ETH zürich

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizera

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
Agroscope

Optical lean phenotyping methods in the context of wheat variety testing

Simon Treier

Diss. ETH No. 31161
2025

Diss. ETH No. 31161

Optical lean phenotyping methods in the context of wheat variety testing

A thesis submitted to attain the degree of

Doctor of Sciences
(Dr. sc. ETH Zürich)

presented by

Simon Philip Treier
born on 11.12.1985

accepted on the recommendation of

Prof. Dr. Achim Walter
Dr. Juan M. Herrera
Dr. Lukas Roth
Prof. Dr. Scott Chapman

2025

Post address:
Agroscope Changins
Simon Treier - DC
Route de Duillier 60
1260 Nyon
Switzerland
www.agroscope.admin.ch
Phone: +41 58 480 84 71

"We demand rigidly defined areas of doubt and uncertainty!".

Douglas Adams, The Hitchhiker's Guide to the Galaxy

Contents

Summary

The worldwide food demand is expected to increase by 35 % to 56 % between 2010 and 2050 due to the growing world population. At the same time, about a third of earth's land surface is already used for agriculture. To increase agricultural production, without expanding the land under cultivation or increasing agrochemical inputs used, sustainable ways of intensification are necessary.

One approach of sustainable intensification is to breed high-performance genotypes that are well adapted to specific environments, and breeding was instrumental in the sharp increase in yields since the Green Revolution. Nevertheless, yields of important field crops, including wheat (*Triticum aestivum* L.), have stagnated since the late 1990s. This has significant implications for future food security, as wheat is one of the most important staple crops. Calories from wheat supplied up to 21 %of the energy consumed by humans.

Genetic gain of wheat was shown to not have declined, and under optimal conditions, increased grain yields can still be realized. But genetic gains were in part counteracted by climate change, which comes with a higher frequency of adverse growing conditions such as drought and heat during sensitive growing stages and leads to a climate that is generally less favorable for agriculture, especially in temperate and hot climates.

Thus, further efforts should focus on adapting genotypes and management practices to local conditions and climates, as interactions between genotypes and environments (G×E) but also management (G×E×M) are responsible for large variability in grain yields. Developing and identifying optimal genotypes for specific environments is paramount to closing the gap between the attainable and the realized yield.

Plant breeding allows for the development of improved varieties. To translate genetic progress in breeding into higher yields, the most suitable genotypes must be used in specific environments. Thus, breeding must be paralleled with a thorough characterization of variety performance in respective environments by conducting multi-environment trials (MET) for variety testing. Results of variety testing are published in annual lists of recommended varieties to allow farmers and other stakeholders to choose varieties that meet the market goals in their specific environments and soils.

Typical traits monitored in variety testing are grain yield at 15 % water content, lodging resistance, early maturity, early heading, sprouting, overwintering, plant height, thousand kernel weight, hectoliter weight, resistance to various diseases such as powdery mildew, rusts, different *Septoria* species and *Fusarium* head blight. On harvested grains, the baking quality, the sedimentation index (Zeleny test), and the protein content are evaluated. These traits are typically still assessed by field observations or laboratory analysis, which is labor intensive and costly, especially since variety testing usually uses METs and thus traits must be assessed on multiple sites. In breeding, high-throughput field phenotyping (HTFP) methods were proposed and developed to make the assessment of plant traits more efficient and also to assess novel traits.

Variety testing could also profit from new HTFP methods, but many of these methods have been tested experimentally under relatively controlled conditions and still have a relatively low technology readiness level (TRL). They have thus not yet been established in the daily practice or variety testing. To close the gap between basic research and methods that are actually

applied by variety testing organizations to finally benefit farmers' production, translational research is necessary.

This thesis focuses on "lean phenotyping" as one aspect of translational research in the context of variety testing. Many of the proposed HTFP workflows are just too expensive in terms of initial investments, operational costs, and labor to be applied in variety testing, especially as within MET, multiple sites must be measured. Switching to cheaper equipment or simplifying workflows is often not easily possible, as the quality of the measured traits is too poor for beneficial integration into variety testing.

Lean phenotyping in this context is understood as translational research to design workflows in such a way that results of sufficient quality can be generated even with more affordable sensors or that the costs of high-quality methods can be reduced. Another aspect concerns the use of new technologies or sensors to measure new traits, provided that they add value for variety testing. The ultimate goal in lean phenotyping is to develop methods with an acceptable balance of costs and benefits.

This thesis was carried out within the Agroscope "Production Technology & Cropping Systems Group", which is responsible for the official variety testing of wheat and other cereals in Switzerland. The thesis is committed to translational research on phenotyping methods for variety testing to further develop them toward lean phenotyping. It aims to evaluate and increase the TRL of existing phenotyping methods under realistic variety testing conditions. Therefore, three optical lean phenotyping methods, drone-based thermal cameras, PhenoCams, and chlorophyll fluorescence sensors were developed, adapted, and examined.

Airborne thermography is a promising method for measuring canopy temperature (CT) to examine the relative fitness of a plant in an environment, especially in the context of heat and drought. With the development of drone-based thermal cameras, airborne thermography became easily accessible and affordable. However, the high variability of CT data from such uncooled thermal cameras makes interpretation very challenging and hindered the broad adoption of this new technology. Therefore, a multi-view approach was adapted for drone-based thermal cameras. Without changing the equipment used, but only with a novel and more comprehensive statistical analysis pipeline, the temporal and spatial variability of CT could be estimated and corrected, allowing for more genotype-specific and consistent measurements. This increased the interpretability of CT, thereby rendering thermal imaging more applicable and therefore more interesting as a phenotyping method in wheat variety testing.

However, CT is an ephemeral trait and influenced by many factors in the short term. The thermal sensor and CT itself are very sensitive to confounding environmental influences. In addition, viewing geometry related effects add uncertainty to CT estimates. These effects mask experimental sources of variance, such as different genotypes and treatments, and while CT is mostly considered a proxy measure of stomatal conductance, the trait also features phenotypic correlations with other traits such as plant height or fractional canopy cover. To gain a thorough understanding of CT as a trait, the sources of variance of drone-based uncooled thermography were thoroughly examined based on 99 flights. Using the thermal multi-view approach developed in the previous step, more than 96.5 % of the initial variance could be explained on average by experimental and confounding sources of variance combined. The insights gained support the planning, conducting, and interpretation of drone-based CT screenings in variety testing.

While drone-based CT represents a new trait that could be useful in the context of evaluating varieties and their resistance to drought and heat, this thesis also developed methods to screen established traits more efficiently and objectively. It is crucial to know the timing of phenological stages and the senescence behavior of genotypes to select for locally adapted varieties and to plan crop management accordingly. Knowing the timing of phenological stages also allows for a more meaningful interpretation of G×E interactions. Capturing these traits

with frequent field visits is very time-consuming. A semimobile PhenoCam setup was used to track phenology and senescence from ear emergence to full maturity. An economic analysis revealed that PhenoCams are economically interesting for observing distant experimental sites. Thus, PhenoCams offer a cost-effective replacement of visual ratings of phenology and senescence, especially in the context of MET.

As for evaluating the timing of phenological stages and senescence, the rating of plant disease infestations under field conditions is time-consuming and prone to subjectivity. Chlorophyll fluorescence (CF) was proposed as a tool to track and rate *Fusarium* head blight infestations and this chapter explored the potential and limitations of CF methods under field conditions. A hand-held CF sensor was used to track *Fusarium* infestations first in a greenhouse trial and the method was then transferred to a field trial and tested for two seasons, together with a CF imaging approach. The tested methods worked well in high-level infestations, but it is hypothesized that they would fail at low-level infestations due to a too low number of measurements and the throughput of the method would need to be increased drastically, *e.g.* by automatization with field robots.

This work provides methodologies and insight for three optical lean phenotyping methods in the context of wheat variety testing. For drone-based thermography, a novel statistical approach was developed to handle the large variability of such data and the approach was applied to examine the manifold sources of variance in CT estimates based on thermal images. PhenoCams were applied to observe phenology and senescence and finally the potential of chlorophyll fluorescence to track the disease progression of *Fusarium* in field conditions was examined.

Zusammenfassung

Die weltweite Nahrungsmittelnachfrage wird, bedingt durch das Bevölkerungswachstum zwischen 2010 und 2050, voraussichtlich um 35 % bis 56 % steigen. Gleichzeitig wird bereits etwa ein Drittel der Erdoberfläche für die Landwirtschaft genutzt. Um die landwirtschaftliche Produktion zu steigern, ohne die Anbaufläche auszudehnen oder den Einsatz von Agrochemikalien zu erhöhen, sind nachhaltige Intensivierungsmethoden notwendig.

Ein Ansatz der nachhaltigen Intensivierung ist die Züchtung von Hochleistungsgenotypen, die an spezifische Umweltbedingungen gut angepasst sind. Die Pflanzenzüchtung spielte eine entscheidende Rolle beim starken Ertragsanstieg seit der Grünen Revolution. Dennoch stagnieren die Erträge wichtiger Feldfrüchte, darunter Weizen (*Triticum aestivum* L.), seit den späten 1990er Jahren. Dies hat erhebliche Auswirkungen auf die künftige Ernährungssicherheit, da Weizen eines der wichtigsten Grundnahrungsmittel ist. Kalorien aus Weizen liefern bis zu 21 % der von Menschen konsumierten Energie.

Es wurde gezeigt, dass der genetische Fortschritt bei Weizen nicht zurückgegangen ist und unter optimalen Bedingungen weiterhin Ertragssteigerungen erzielt werden können. Allerdings wurden die genetischen Gewinne teilweise durch den Klimawandel ausgeglichen, der häufigere ungünstige Wachstumsbedingungen wie Dürre und Hitze in sensiblen Wachstumsphasen mit sich bringt und zu einem allgemein weniger günstigen Klima für die Landwirtschaft führt, insbesondere in gemässigten und heissen Klimazonen.

Daher sollten weitere Anstrengungen darauf abzielen, Genotypen und Anbaumassnahmen an lokale Bedingungen und Klimazonen anzupassen, da Wechselwirkungen zwischen Genotypen und Umweltbedingungen (Genotype×Environment: G×E) sowie dem Management (Genotype×Environment×Management: G×E×M) für eine grosse Variabilität der Kornerträge verantwortlich sind. Die Entwicklung und Identifizierung optimaler Genotypen für spezifische Umweltbedingungen ist von entscheidender Bedeutung, um die Lücke zwischen dem erreichbaren und dem tatsächlich realisierten Ertrag zu schliessen.

Die Pflanzenzüchtung ermöglicht die Entwicklung verbesserter Sorten. Um den genetischen Fortschritt in der Züchtung in höhere Erträge umzusetzen, müssen die am besten geeigneten Genotypen in spezifischen Umweltbedingungen eingesetzt werden. Daher muss die Züchtung mit einer umfassenden Charakterisierung der Leistung von Sorten in den jeweiligen Umweltbedingungen einhergehen, indem Versuche unter verschiedenen Umweltbedingungen (multi-environment trials: MET) zur Sortenprüfung durchgeführt werden. Die Ergebnisse der Sortenprüfung werden in jährlichen Listen empfohlener Sorten veröffentlicht, um Landwirten und anderen Marktakteuren die Auswahl von Sorten zu ermöglichen, die die Marktanforderungen in ihren spezifischen Umweltbedingungen und Böden erfüllen.

Typische Merkmale, die in der Sortenprüfung beobachtet werden, sind Kornertrag bei 15 % Wassergehalt, Lagerfestigkeit, Frühreife, Frühzeitigkeit des Ährenschiebens, Neigung zum Auswuchs, Überwinterung, Pflanzenhöhe, Tausendkorngewicht, Hektolitergewicht sowie die Resistenz gegen verschiedene Krankheiten wie Mehltau, Rostarten, verschiedene *Septoria*-Arten und Ährenfusariose. An geerntetem Getreide werden die Backqualität, der Sedimentationswert (Zeleny-Test) und der Proteingehalt bewertet. Diese Merkmale werden typischerweise noch durch Feldbeobachtungen oder Laboranalysen erfasst, was arbeitsintensiv und kostspielig ist, insbesondere da die Sortenprüfung in METs durchgeführt wird und somit Merkmale

an mehreren Standorten bewertet werden müssen. In der Züchtung wurden Methoden des Hochdurchsatz-Phänotypisierens (high-throughput field phenotyping: HTFP) vorgeschlagen und entwickelt, um die Bewertung von Pflanzenmerkmalen effizienter zu gestalten und neue Merkmale zu erfassen.

Auch die Sortenprüfung könnte von neuen HTFP-Methoden profitieren, doch viele dieser Methoden wurden bisher nur experimentell unter relativ kontrollierten Bedingungen getestet und haben noch einen relativ niedrigen technologischen Reifegrad (technology readiness level: TRL). Sie wurden daher noch nicht in die tägliche Praxis der Sortenprüfung übernommen. Um die Lücke zwischen der Grundlagenforschung und den in der Praxis angewandten Methoden der Sortenprüfung zu schliessen, ist translationale Forschung erforderlich.

Diese Dissertation konzentriert sich auf das Konzept des "Lean Phenotyping" als ein Aspekt der translationalen Forschung im Kontext der Sortenprüfung. Viele der vorgeschlagenen HTFP-Workflows sind für die Sortenprüfung aufgrund hoher Anfangsinvestitionen, Betriebskosten und des hohen Arbeitsaufwands zu teuer, insbesondere da innerhalb von METs mehrere Standorte gemessen werden müssen. Der Gebrauch von günstigeren Instrumenten oder die Vereinfachung von Arbeitsabläufen ist oft nicht ohne weiteres möglich, da die Qualität der gemessenen Merkmale zu gering ist, um eine vorteilhafte Integration in die Sortenprüfung zu ermöglichen.

Lean Phenotyping wird in diesem Kontext als translationale Forschung verstanden, um Arbeitsabläufe so zu gestalten, dass Ergebnisse von ausreichender Qualität auch mit erschwinglicheren Sensoren erzielt oder die Kosten hochwertiger Methoden reduziert werden können. Ein weiterer Aspekt betrifft die Nutzung neuer Technologien oder Sensoren zur Erfassung neuer Merkmale, sofern sie einen Mehrwert für die Sortenprüfung bieten. Das ultimative Ziel des Lean Phenotyping ist die Entwicklung von Methoden mit einem akzeptablen Kosten-Nutzen-Verhältnis.

Diese Dissertation wurde innerhalb der Agroscope Gruppe "Anbautechnik und Sorten Ackerbau " durchgeführt, die für die offizielle Sortenprüfung von Weizen und anderen Getreidearten in der Schweiz verantwortlich ist. Die Dissertation widmet sich der translationalen Forschung zu Phänotypisierungsmethoden für die Sortenprüfung, um diese in Richtung Lean Phenotyping weiterzuentwickeln. Ziel ist es, den TRL bestehender Phänotypisierungsmethoden unter realistischen Sortenprüfungsbedingungen zu bewerten und zu erhöhen. Daher wurden drei optische Lean Phenotyping Methoden entwickelt, angepasst und untersucht: drohnenbasierte Wärmebildkameras, PhenoCams und Chlorophyllfluoreszenz Sensoren.

Die luftgestützte Thermografie ist eine vielversprechende Methode zur Messung der Bestandestemperatur (canopy temperature: CT), um die relative Fitness einer Pflanze in bestimmten Umweltbedingungen zu bewerten, insbesondere im Kontext von Hitze und Trockenheit. Mit der Entwicklung drohnenbasierter Wärmebildkameras wurde die luftgestützte Thermografie leicht zugänglich und erschwinglich. Allerdings erschwert die hohe Variabilität der CT-Daten solcher ungekühlten Wärmebildkameras die Interpretation erheblich und verhinderte eine breite Anwendung dieser neuen Technologie. Daher wurde ein Multi-View-Ansatz für drohnenbasierte Wärmebildkameras adaptiert. Ohne Änderung der verwendeten Ausrüstung, sondern allein durch eine neuartige und umfassendere statistische Analysepipeline, konnte die zeitliche und räumliche Variabilität der CT geschätzt und korrigiert werden. Dies ermöglichte eine spezifischere und konsistentere Messung der Genotypen, wodurch die Interpretierbarkeit der CT verbessert wurde und thermografische Bildgebung als Phänotypisierungsmethode für die Weizensortenprüfung interessanter wurde.

Allerdings ist CT ein flüchtiges Merkmal, das kurzfristig von vielen Faktoren beeinflusst wird. Der Wärmesensor und die CT selbst sind sehr empfindlich gegenüber störenden Umwelteinflüssen. Darüber hinaus erzeugen Effekte im Zusammenhang mit der Betrachtungsgeometrie

Unsicherheiten in den CT-Schätzungen. Diese Effekte überdecken experimentelle Varianzquellen wie unterschiedliche Genotypen und Behandlungen. Obwohl CT hauptsächlich als Proxy-Mass für die stomatäre Leitfähigkeit betrachtet wird, zeigt dieses Merkmal auch phänotypische Korrelationen mit anderen Merkmalen wie Pflanzenhöhe oder Bodenbedeckung. Um ein umfassendes Verständnis von CT als Merkmal zu gewinnen, wurden die Varianzquellen der drohnenbasierten ungekühlten Thermografie anhand von 99 Flügen eingehend untersucht. Mithilfe des im vorherigen Schritt entwickelten thermischen Multi-View-Ansatzes konnten durchschnittlich mehr als 96.5 % der ursprünglichen Varianz durch experimentelle und störende Varianzquellen erklärt werden. Die gewonnenen Erkenntnisse unterstützen die Planung, Durchführung und Interpretation von drohnenbasierten CT-Screenings in der Sortenprüfung.

Während die drohnenbasierte CT ein neues Merkmal darstellt, das für die Bewertung von Sorten und ihrer Widerstandsfähigkeit gegenüber Trockenheit und Hitze nützlich sein könnte, wurden in dieser Dissertation auch Methoden zur effizienteren und objektiveren Erfassung etablierter Merkmale entwickelt. Es ist entscheidend, den Zeitpunkt der phänologischen Stadien und das Seneszenzverhalten von Genotypen zu kennen, um lokal angepasste Sorten auszuwählen und das Kulturmanagement entsprechend zu planen. Das Wissen über den Zeitpunkt der phänologischen Stadien ermöglicht auch eine aussagekräftigere Interpretation von G×E-Interaktionen. Die Erfassung dieser Merkmale durch häufige Feldbesuche ist jedoch sehr zeitaufwendig. Ein semimobiles PhenoCam-Setup wurde verwendet, um die Phänologie und Seneszenz vom Ährenschieben bis zur vollständigen Reife zu verfolgen. Eine wirtschaftliche Analyse ergab, dass PhenoCams für die Beobachtung entfernter Versuchsfelder wirtschaftlich interessant sind. Somit bieten PhenoCams eine kostengünstige Alternative zu visuellen Bewertungen der Phänologie und Seneszenz, insbesondere im Kontext von MET.

Ebenso wie die Bewertung des Zeitpunkts der phänologischen Stadien und der Seneszenz ist die Bewertung von Pflanzenkrankheitsbefällen unter Feldbedingungen zeitaufwendig und subjektiv. Chlorophyllfluoreszenz (CF) wurde als Werkzeug vorgeschlagen, um Ährenfusariosen zu beobachten und zu bonitieren. In diesem Kapitel wurden das Potenzial und die Grenzen von CF-Methoden unter Feldbedingungen untersucht. Ein tragbarer CF-Sensor wurde zunächst in einem Gewächshausversuch zur Erfassung von *Fusarium*-Befall eingesetzt; anschliessend wurde die Methode in einen Feldversuch übertragen und über zwei Vegetationsperioden getestet, zusammen mit einem CF-Bildgebungsansatz. Die getesteten Methoden funktionierten gut bei starkem Befall, es wird jedoch vermutet, dass sie aufgrund einer zu geringen Anzahl von Messungen, bei niedrigen Befallsstärken nicht funktionieren würden. Der Durchsatz der Methode müsste drastisch erhöht werden, z.B. durch Automatisierung mit Feldrobotern.

Diese Arbeit liefert Methoden und Erkenntnisse zu drei optischen Lean Phenotyping Methoden im Kontext der Weizensortenprüfung. Für die drohnenbasierte Thermografie wurde ein neuartiger statistischer Ansatz entwickelt, um die hohe Variabilität solcher Daten zu bewältigen, und dieser wurde angewendet, um die vielfältigen Varianzquellen in CT-Schätzungen auf Basis thermografischer Bilder zu untersuchen. PhenoCams wurden eingesetzt, um Phänologie und Seneszenz zu beobachten, und schliesslich wurde das Potenzial der Chlorophyllfluoreszenz zur Verfolgung der Krankheitsprogression von *Fusarium* unter Feldbedingungen untersucht.

1 General introduction

1.1 The need for improved and adapted crop genotypes

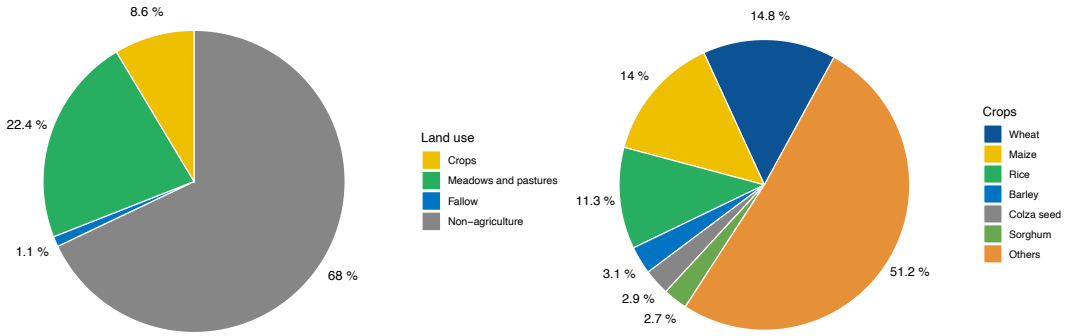
With the world population projected to reach 9.7 billion people by 2050, food demand is expected to increase by 35 % to 56 % between 2010 and 2050 (Van Dijk et al., 2021) and humanity is challenged to respond to this growth in demand with an increased food production. Expanding land under cultivation or intensifying agriculture using more inputs, such as agrochemicals and fertilizers, increases yield, but both strategies have significant adverse effects such as environmental pollution, loss of biodiversity, and negative impacts on rural communities (Kamau et al., 2023). This highlights the need for a sustainable intensification, to produce "more food with less environmental impact" (Godfray and Garnett, 2014).

One approach of sustainable intensification is to breed high-performance genotypes that are well adapted to specific environments. Breeding has been instrumental in the sharp increase in yields since the Green Revolution (T. Fischer et al., 2014; Crespo-Herrera et al., 2017) by developing new genotypes that are more efficient in the conversion of inputs into yield and more adapted to more intensive agricultural practices (M. P. Reynolds, Borrell, et al., 2019). Nevertheless, yields of important field crops, including wheat (*Triticum aestivum* L.), have stagnated in important growing regions across the world since the late 1990s (Ray et al., 2012; Schauberger et al., 2018).

Yet, for wheat, genetic gain was shown to not have declined (*e.g.* Brisson et al., 2010; Gerard et al., 2024). Thus, under optimal conditions, with plants growing to their full potential, increased grain yields can still be realized. However, genetic gains were counteracted in part by climate change. With a higher frequency of adverse growing conditions such as drought and heat during sensitive growth stages, climate change leads to climates that are generally less favorable for agriculture. These trends apply to many regions in temperate climates (Brisson et al., 2010), but especially to regions already in high food insecurity (M. P. Reynolds, Borrell, et al., 2019).

This has significant implications for future food security, as wheat is one of the most important sources of calories for humanity. In 2022, 32 % of earth's surface were used for agriculture and 1487.9×10^{-6} ha or 8.6 % for crop production (FAOSTAT, 2025c). Of the crop area, 220.4×10^{-6} ha or 14.8 % were used for wheat production(FAOSTAT, 2025a). Calories produced with wheat corresponded to 21 % of the energy demand of humanity, assuming direct human consumption and an average daily per capita energy consumption of 2'353 kcal (Berners-Lee et al., 2018; FAOSTAT, 2025b).

Thus, further efforts should focus on adapting genotypes and management practices to local conditions and climates, as interactions between genotypes and environments (G×E) but also management (G×E×M) are responsible for large variability in grain yields (Herrera et al., 2020). Developing and identifying optimal genotypes for respective environments is paramount to close the yield gap, *i.e.* the difference between the attainable and the realized yield (Schils et al., 2018; Gerber et al., 2024). To that end, more resources must be allocated in breeding as one measure to overcome yield stagnation, by developing genotypes resilient to biotic and abiotic stresses (Hickey et al., 2019; Tester and Langridge, 2010).



(a) (b)

FIGURE 1.1: About 32 % of earths land surface are used for agriculture and 8.6 % for growing crops (a). This cropping area is dominated by three main crops. Wheat is grown on about 14.8 %, maize on 14 % and Rice on 11.3 % (b), while the remaining crop cover far less important shares of the cropping area.

1.2 The importance of variety testing

Breeding goals are specific for different regions. For some mega-environments, resistance to heat and drought during grain filling are the main breeding goals, for other regions, resistance to specific diseases or to winter-kill is the main focus of breeding efforts, and very often it is a location-specific combination of multiple goals (M. P. Reynolds, Pask, et al., 2012). Within mega-environments, finer adjustments of the breeding goals leads to the development of genotypes that are highly adapted to very local conditions (Bustos-Korts, Boer, Layton, et al., 2022; C. J. Yang et al., 2024).

Although plant breeding allows the development of improved varieties, to translate genetic progress in breeding into higher yields and close the on-farm yield gap (Cooper et al., 2021), the most suitable genotypes must be used in respective target environments. Thus, breeding must be paralleled with a thorough characterization of genotype performance in respective target environments, which is usually conducted with multi-environment trials (MET), where a set of genotypes is sown in multiple environments and compared to each other (Bustos-Korts, Boer, Layton, et al., 2022). This MET variety testing is usually done by national or regional organizations, which publish annual lists of recommended varieties (Levy et al., 2017; Niedbała et al., 2022; Fang et al., 2024; C. J. Yang et al., 2024).

Only genotypes with improved traits that show trait consistency in diverse environments are eligible for variety registration (H. E. Ahrends et al., 2018; Voss-Fels et al., 2019). Testing for "Value for Cultivation and Use" (VCU) is a mandatory step of the crop variety testing procedure in the European Union (EU) and Switzerland. New varieties are registered and added to the list of recommended varieties if they show superior yield, quality, or agronomic properties. VCU testing guarantees that farmers have access to seeds of good quality and with well-defined properties (Becker, 2011).

To protect the interests of breeders, genotypes must show distinctness, uniformity, and stability (DUS), to be acceptable as a new variety. In DUS trials, genotypes are tested if they are different compared to existing varieties and if they are uniform and stable over time at the population level.

While countries outside EU and Switzerland may have other procedures than VCU and DUS, many crop variety evaluation programs (CVEP) share the concept that new varieties need to outperform older varieties in specific aspects (Fang et al., 2024) and must be distinguishable from them to be accepted for listsf recommended varieties (Cooke and Reeves, 2003).

By testing genotypes in respective climates in local soils with locally representative crop management strategies, variety testing allows agronomists and farmers to decide for varieties that fit the market goals (C. J. Yang et al., 2024), climatic conditions (D. Reynolds et al., 2019; Niedbała et al., 2022) and management strategies of individual farms (Voss-Fels et al., 2019; Rotili et al., 2020; Cooper et al., 2021), leading to higher financial incomes (Levy et al., 2017; Niedbała et al., 2022).

Typical traits monitored in variety testing are grain yield at 15 % water content, lodging resistance, early maturity, early heading, sprouting, overwintering, plant height, thousand kernel weight, hectoliter weight, resistance to various diseases such as powdery mildew, rusts, different *Septoria* species and *Fusarium* head blight. On harvested grains, the baking quality, sedimentation index (Zeleny test), and protein content are evaluated (WBF, 2021). These traits are typically still assessed by field observations or laboratory analysis (Cooke and Reeves, 2003), which is labor intensive and costly, especially since variety testing usually uses METs (Eichi et al., 2020) and thus traits must be assessed on multiple sites.

1.3 High-throughput field phenotyping (HTFP)

The field of genomics has evolved rapidly since the 1990s. However, to fully benefit from the potential of genomics, "plant traits such as growth, development, tolerance, resistance, architecture, physiology, ... yield and ... individual quantitative parameters ..."(L. Li et al., 2014) must be properly assessed by phenotyping (Araus and Cairns, 2014; Cobb et al., 2013) . Therefore, high-throughput plant phenotyping (HTPP) was proposed (Fahlgren et al., 2015; Crain et al., 2018) and developed (Walter et al., 2015; Sun et al., 2022) for combined use with genomics in breeding (S. Michel et al., 2023).

For variety testing, the literature is sparse, but some challenges are very similar to those for breeding, and variety testing is also expected to benefit from new HTPP methods.

Although HTPP includes phenotyping in controlled and field conditions, variety testing always includes field phenotyping, which is the most challenging phenotyping settings due to multiple confounding sources of variance, such as spatial and temporal heterogeneity of traits due to *e.g.*, field gradients, changing weather during measurements or disruptive weather events. The confounding of multiple sources of variance makes data acquisition but also interpretation very challenging (*e.g.* Araus, Kefauver, et al., 2018; Aasen, Kirchgessner, et al., 2020; M. P. Reynolds, S. C. Chapman, et al., 2020). Nonetheless, high-throughput field phenotyping (HTFP) is a very active research field (*e.g.* Ludovisi et al., 2017; Jimenez-Berni, Deery, et al., 2018; Perich et al., 2020; Roth, Rodríguez-Álvarez, et al., 2021), mostly in the context of breeding, yet many HTFP approaches are also promising for variety testing.

1.4 HTFP and the need for translational research

However, many of HTFP methods have not yet been established in the daily practice of breeding or variety testing. For variety testing, digitization has led to improved tools for data organization and analysis (F. Yang et al., 2023; Fang et al., 2024), and for the integration of genomic data (Carvalho et al., 2024; Bruschi et al., 2024), as computing power and statistical software became readily available. Yet, on the side of trait assessment, the impact of new digital phenotyping methods remained very limited. Many phenotyping approaches have been developed to obtain information on crop state, morphology, and performance (*e.g.* Adamsen et al., 1999; Hunt, Doraiswamy, et al., 2013; Hasan et al., 2019; T. Jensen et al., 2007; Gracia-Romero et al., 2017; Jimenez-Berni, Deery, et al., 2018; Yue et al., 2019; H. Wang et al., 2020), for digitalized phenotyping, but few considered the specific needs of variety

testing. For example, Hu et al., 2024 combined coarse satellite data from wheat variety trials at small spatial scales with wheat growth modeling and radiative transfer modeling to retrieve aboveground biomass and assess within-field variability in order to evaluate the quality of trials in a variety testing MET network.

On the one hand, the gap between the development of basic methods and the applicability in variety testing is owed to the fact that variety testing organizations have limited resources (Cullis, A. Smith, et al., 2000), which must be allocated for the assessment of the most relevant traits. HTFP methods often involve considerable initial investment and must first be learned (M. P. Reynolds, Borrell, et al., 2019). On the other hand, the developed approaches have often only been tested experimentally under relatively controlled conditions, in the laboratory or on a research station, but still have a relatively low technology readiness level (TRL, Table 1.1) in the range of $\sim$ 3 and 5 (*cf.* Table 1.1). Some phenotyping methods and concepts have possibly been demonstrated under field conditions (TRL $\sim$ 6 - 7), but prototypes have not yet made it into application in an "operational environment" (European Commission, 2014) or even day-to-day practice (TRL $>$ 7).

The application of HTPF in practice also requires increased interdisciplinarity, as specific technical knowledge (*e.g.* the piloting of drones or the operation of specialized cameras or sensors) is necessary for the measurements, then, the data generated require more comprehensive data management, and finally, the meaningful interpretation of the data requires both technical and agronomic knowledge (Kholová et al., 2021).

Table 1.1: Technology readiness levels (TRL) as proposed by the European Commission; Mankins, 1995; European Commission, 2014).

TRL	Desrition
1	basic principles observed
2	technology concept formulated
3	experimental proof of concept
4	technology validated in lab
5	technology validated in relevant environment (industrially relevant environment in the case of key enabling technologies)
6	technology demonstrated in relevant environment (industrially relevant environment in the case of key enabling technologies)
7	system prototype demonstration in operational environment
8	system complete and qualified
9	actual system proven in operational environment (competitive manufacturing in the case of key enabling technologies; or in space)

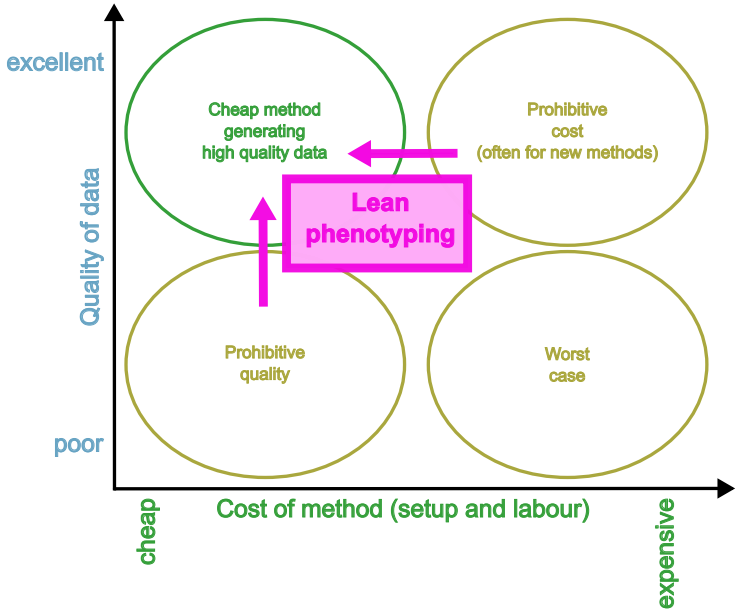
To close the gap between basic research and methods that are actually applied by variety testing organizations to finally benefit farmers' production, "translational research" is necessary (M. P. Reynolds, Borrell, et al., 2019). "This kind of research is often seen as more complicated and time-consuming than basic research and less sexy than working at the 'cutting edge' where research is typically divorced from agricultural realities in order to achieve faster and cleaner results" (CIMMYT, 2019).

1.5 Lean phenotyping for variety testing

This thesis focuses on "lean phenotyping" as one aspect of translational research in the context of variety testing. Many of the proposed HTFP workflows are just too expensive in terms of initial investments, operational costs, and labor to be applied in variety testing, especially as within MET, multiple sites must be measured (Eichi et al., 2020). Switching to cheaper equipment or simplifying workflows is often not easily possible, as the quality of the measured traits is too poor for beneficial integration into variety testing.

Lean phenotyping in this context is understood as translational research to design workflows in such a way that results of sufficient quality can be generated even with less expensive sensors or that the costs of high-quality methods can be reduced (Fig .1.2). Another aspect concerns the use of new technologies or sensors to measure new traits, provided that they add

value for variety testing. The ultimate goal in lean phenotyping is to develop methods with an acceptable balance of costs and benefits (López-Lozano and Baruth, 2019) that are applicable in a day-to-day variety testing routine.



(a)

FIGURE 1.2: Cost-Quality considerations in lean phenotyping. The ultimate goals in lean phenotyping is to develop methods with an acceptable balance of costs and benefits through translational research.

1.6 What is a good trait in crop performance assessment?

The primary target of breeding and variety testing is the selection of genotypes with superior primary traits such as yield and quality. Both are the result of complex G×E×M interactions (Cooper et al., 2021). These primary traits are attained in specific environments (Laidig et al., 2008). For example, a variety can attain high yields due to late maturity combined with efficient remobilization of carbohydrates (E. A. Chapman et al., 2021), due to increased water use efficiency (Rebetzke et al., 2013), or because its leaves were less affected by foliar disease (Zetzsche et al., 2020). A high-yielding genotype in one environment might perform poorly in another environment, *e.g.* due to water stress and primary traits thus suffer from low genotype specificity, expressed as heritability, over different locations and years (Araus, Slafer, et al., 2008; Bustos-Korts, Boer, Malosetti, et al., 2019).

For a more complete understanding and a more accurate prediction of primary traits, secondary traits, also called component traits (Bustos-Korts, Boer, Malosetti, et al., 2019), can be assessed. A secondary trait is adding value to prediction and decision making when it's heritability is higher between locations and years, than heritability of primary traits. So, there should be variability of the trait between genotypes, and this variability should be less affected by environmental conditions. Secondary traits should also show a genetic correlation with a primary trait and the assessment of the secondary trait should be rapid, reliable, and affordable (Araus, Slafer, et al., 2008; Bustos-Korts, Boer, Malosetti, et al., 2019; M. P. Reynolds, S. C. Chapman, et al., 2020).

Most of the aforementioned variety testing traits (lodging resistance, early maturity, early heading, sprouting, overwintering, plant height, thousand kernel weight, hectoliter weight, resistance to various diseases such as powdery mildew, rusts, different *Septoria* species and

Fusarium) can be considered secondary traits. One objective of lean phenotyping for variety testing is to measure them more efficiently.

1.7 Introducing new traits

Well established and new digital sensors also come with new opportunities to assess new secondary traits related to plant growth, in addition to traditional variety testing traits. Those may also include traits related to more environmentally friendly crop production, notably increased resource use efficiency (RUE) but also more resilient production systems in the face of a changing climate (Shiferaw et al., 2013; FAO, 2017). Such new traits would support official genotype testing to contribute to RUE and resilience of cropping systems by improving variety testing procedures and making available information about the performance of crop varieties associated with RUE and resilience. Based on such new traits, farmers and other stakeholders such as processors and mills, could be able to make more informed variety choices considering local climates as well as biotic and abiotic growing conditions. Thus, as a second objective of lean phenotyping for variety testing, such new traits are explored and developed.

1.8 Aims and structure of this thesis

The need for translational research in variety testing has been recognized by the European Union (Invite, 2025; InnoVar, 2025) and the Australian government (GRDC, 2025), and corresponding research projects have been initiated, underlining the importance of the topic. This thesis was conducted within the the "Production Technology & Cropping Systems Group" of Agroscpoe (`www.agroscope.admin.ch`), which is in charge for the official variety testing of wheat and other cereals in Switzerland. The thesis is committed to translational research on phenotyping methods for variety testing to further develop them towards lean phenotyping. It aims to evaluate and increase the TRL of existing phenotyping methods under realistic variety testing conditions. Therefore, three optical lean phenotyping methods, drone-based thermal cameras, PhenoCams, and chlorophyll fluorescence sensors were developed, adapted, and examined in the following chapters:

Chapter 2 - Improving drone-based uncalibrated estimates of wheat canopy temperature in plot experiments by accounting for confounding factors in a multi-view analysis:

With the development of drone-based thermal cameras, airborne thermography, previously only possible with helicopters or other manned platforms, became easily accessible and affordable. However, the large variability of the thermal imaging data from uncooled thermal cameras, mainly due to thermal drift, made the quantitative interpretation very challenging. In this chapter, a multi-view method was introduced to conduct an image-wise sequential analysis of thermal images instead of an orthomosaic-based analysis. Knowing the trigger timing of individual images, thermal drift over the sequence of individual flights could be estimated, as well as effects related to viewing geometry, and thus be corrected for in a mixed model approach. With the multi-view method, the consistency and genotype specificity of canopy temperature (CT) measurements was significantly improved compared to approaches relying on orthomosaics in a two-year field variety testing trial with winter wheat. Thermal imaging became more reliable without changing the equipment used but through a novel more comprehensive statistical analysis of the data, rendering thermal imaging more applicable, and thus more interesting as a phenotyping method in wheat variety testing.

Chapter 3 - Multi-view can explain large fractions of variance and their sources in drone-based thermography on wheat plots:

Understanding a phenotyping method is crucial for a correct interpretation of the results to avoid erroneous conclusions on physiological relations between the measured trait and the status of the plant. CT, especially when based on uncooled thermal cameras, is a challenging trait to measure and interpret, not only due to thermal drift. While CT is mostly considered as a proxy measurement of stomatal conductance, the trait also features phenotypic correlations with other traits like plant height or fractional canopy cover and is affected by other short-term sources of variance in field conditions. Based on the thermal multi-view method presented in Chapter 2, this chapter identified and analyzed manifold sources of variance in CT measurements from 99 flights with a drone-based thermal camera flown over two years on two different trials. The experimental sources of variance (genotypes and treatments) were disentangled from the confounding sources of variance, and together they explained large fractions of the initial variance of CT. Ignoring confounding sources led to erroneous conclusions about the phenotypic correlations of CT with other traits. Based on an extensive and diverse dataset, this chapter allows a comprehensive understanding of CT as a trait, which supports the planning, conducting, and interpretation of drone-based CT screenings in variety testing.

Chapter 4 - Comparison of PhenoCams and drones for lean phenotyping of phenology and senescence of wheat genotypes in variety testing:

It is crucial to know the timing of phenological stages and the senescence behavior of genotypes to select for locally adapted varieties and to plan crop management accordingly. Knowing the timing of phenological stages also allows for a more meaningful interpretation of G×E interactions, *e.g.* to distinguish variety adaptation to stresses from stress avoidance. Capturing these traits with frequent field visits is very time-consuming. In contrast to Chapter 1 and 2, where the potential of a relatively new trait was tested, this Chapter developed an alternative and cost-effective approach with full field applicability to capture traits which are well established but generally assessed by visual ratings. A semimobile PhenoCam setup was used to track phenology and senescence from ear emergence to full maturity. The method was compared with visual reference methods, with which it was strongly correlated. An economic analysis revealed that PhenoCams are an interesting option to observe distant experimental sites. Thus, PhenoCams offer a cost-effective replacement of visual ratings of phenology and senescence, especially in the context of MET.

Chapter 5 - Evaluating the potential of chlorophyll fluorescence to detect and rate *Fusarium* head blight on field experiments for winter wheat variety testing:

As for evaluating the timing of phenological stages and senescence, the rating of plant disease infestations under field conditions is time-consuming, and prone to subjectivity due to rater bias. Chlorophyll fluorescence (CF) was proposed as a tool to track and rate *Fusarium* head blight infestations and this chapter explored the potential and limitations of CF methods under field conditions. A hand-held CF point sensor was used to track *Fusarium* infestations first in a greenhouse trial and the method was then transferred to a field trial and tested for two seasons, together with a CF imaging approach. The tested methods worked well in high-level infestations, but due to the low number of measurements that can be taken with a certain time, it is hypothesized that they would fail in low-level infestations and the throughput of the method would need to be increased drastically, *e.g.* by automatization with field robots.

Chapter 6 - General discussion and conclusion:

Finally, the contribution of the different approaches and chapters to lean phenotyping is discussed together with possible future pathways for the individual approaches, but also for their integration in future variety testing setups.

2 Improving drone-based uncalibrated estimates of wheat canopy temperature in plot experiments by accounting for confounding factors in a multi-view analysis

Simon Treier[1,2], Juan M. Herrera[1], Andreas Hund[2], Norbert Kirchgessner[2], Helge Aasen[3], Achim Walter[2], Lukas Roth[2]

1 – Production Technology & Cropping Systems Group, Agroscope, Route de Duiller 60, 1260 Nyon, Switzerland

2 – ETH Zurich, Institute of Agricultural Sciences, Universitätstrasse 2, 8092 Zurich, Switzerland

3 – Earth Observation of Agroecosystems Team, Agroecology and Environment Division, Agroscope, Reckenholzstrasse 191, 8046 Zürich, Switzerland

This chapter was published in ISPRS Journal of Photogrammetry and Remote Sensing Volume 218, doi: 10.1016/j.isprsjprs.2024.09.015,

Abstract

Canopy temperature (CT) is an integrative trait, indicative of the relative fitness of a plant genotype to the environment. Lower CT is associated with higher yield, biomass and generally a higher performing genotype. In view of changing climatic conditions, measuring CT is becoming increasingly important in breeding and variety testing. Ideally, CTs should be measured as simultaneously as possible in all genotypes to avoid any bias resulting from changes in environmental conditions. The use of thermal cameras mounted on drones allows to measure large experiments in a short time. Uncooled thermal cameras are sufficiently lightweight to be mounted on drones. However, such cameras are prone to thermal drift, where the measured temperature changes with the conditions the sensor is exposed to. Thermal drift and changing environmental conditions impede precise and consistent thermal measurements with uncooled cameras. Furthermore, the viewing geometry of images affects the ratio between pixels showing soil or plants. Particularly for row crops such as wheat, changing viewing geometries will increase CT uncertainties. Restricting the range of viewing geometries can potentially reduce these effects. In this study, sequences of repeated thermal images were analyzed in a multi-view approach which allowed to extract information on trigger timing and

viewing geometry for individual measurements. We propose a mixed model approach that can account for temporal drift and viewing geometry by including temporal and geometric covariates. This approach allowed to improve consistency and genotype specificity of CT measurements compared to approaches relying on orthomosaics in a two-year field variety testing trial with winter wheat. The correlations between independent measurements taken within 20 min reached 0.99, and heritabilities 0.95. Selecting measurements with oblique viewing geometries for analysis can reduce the influence of soil background. The proposed workflow provides a lean phenotyping method to collect high-quality CT measurements in terms of ranking consistency and heritability with an affordable thermal camera by incorporating available additional information from drone-based mapping flights in a post-processing step.

2.1 Introduction

Canopy temperature (CT) of wheat (*Triticum aestivum* L.) is an integrative trait "being associated with yield in a range of conditions" (M. P. Reynolds, Pask, et al., 2012). "It is indicative of the relative fitness of a genotype to the environment" (M. P. Reynolds, Pask, et al., 2012). Lower CT is associated with higher yield, biomass and generally a higher performing genotype. CT is tightly linked to stomatal conductance (*e.g.* Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019) and different traits might lead to low CT, *e.g.* a root system that increases water supply to the plant, high intrinsic radiation-use efficiency, photo-protective mechanisms that increase radiation-use efficiency and green area throughout the growth cycle or a late senescence and consequently a larger green area during later stages (Perich et al., 2020; M. P. Reynolds, Pask, et al., 2012). Therefore, CT can be used as an indirect selection criterion for yield (*e.g.* Das, S. C. Chapman, et al., 2021).

Thermal measurements have been proposed for breeding programs at least since the 1980s (Blum et al., 1982; Lepekhov, 2022), but standard procedures with handheld thermometers have their shortcomings, especially because distortions by rapidly changing environmental conditions should be avoided (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; Pask et al., 2012). Main sources of short-term variability in environmental conditions include wind, sunlight, clouds, and air temperature (M. P. Reynolds, Pask, et al., 2012). Thus, genotypes should be measured within a short period, *e.g.* within 30 min (Z. Wang et al., 2023), but this number is highly dependent on the rate of change in environmental conditions. Thermal infrared (TIR) cameras mounted on unmanned aerial vehicles are therefore an interesting option to measure many experimental units in a relatively short time and thus reduce the short-term variability of measurements.

CT is linked to vapor pressure deficit and consequently air temperature (Idso et al., 1981). A higher air temperature leads to higher CT differences which increases ratio of genotypic variability of CT to residual variability of CT. So, thermal surveys pose challenges when applied in temperate climates where hot and dry conditions with a high vapor pressure deficit (VPD), are less frequent and therefore CT differences between genotypes less distinct (Messina and Modica, 2020).

To get accurate CT measurements, calibrated TIR cameras must be used. Cooled TIR cameras are accurate but heavy and cannot be mounted on a lighttwiweight drone (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016). Uncooled calibrated TIR cameras must be calibrated with reference temperature targets (Aragon et al., 2020; Kelly et al., 2019; Nugent et al., 2013), it takes specific system knowledge to operate them (Perich et al., 2020), but they still have limited accuracy (Kelly et al., 2019; Perich et al., 2020) and might need recalibration after having been operating for some months (Aragon et al., 2020). However, there are uncalibrated TIR cameras that can be operated with standard drones and standard software. Such sensors are not well suited to measure absolute CT accurately, but they hold

the potential to measure relative CT consistently (Kelly et al., 2019). Measuring such relative differences might be sufficient in cases where genotype differences are to be identified (H. G. Jones, Serraj, et al., 2009), *e.g.*, in breeding and variety testing. Yet, the relative differences must be consistent for measurements taken within a short interval, *e.g.* 30 min.

Uncooled TIR cameras are prone to thermal drift problems (Kelly et al., 2019; Mesas-Carrascosa et al., 2018; Z. Wang et al., 2023; Yuan and Hua, 2022) where the TIR measurement changes are influenced by the temperature of the sensor. This introduces another source of variance of CT which is not related to the state of the canopy itself or the canopy environment. Additional confounding effects include vignetting, *i.e.* distortions caused by the lens optics where image edges appear darker (or cooler for thermography) than the central regions (Kelly et al., 2019; Yuan and Hua, 2022). The summation of all effects makes it challenging to derive accurate temperature data with both uncalibrated or calibrated uncooled TIR cameras (Kelly et al., 2019; Malbéteau et al., 2021). Research is tackling this issue by different approaches.

Nugent et al. (2013) highlight the importance to include the sensor temperature in the analysis of TIR images, and Ribeiro-Gomes et al. (2017) and Kelly et al. (2019) demonstrate how this inclusion can be achieved in field environments. However, sensor temperature is not always available and Yuan and Hua (2022) proposed a simplified correction for non-uniformity and vignetting based on a single image taken after a flight. Mesas-Carrascosa et al. (2018) and Z. Wang et al. (2023) used drift correction methodology based on features that appear on multiple overlapping images to create corrected orthomosaics. Malbéteau et al. (2021) corrected for temporal trends by normalizing data of single flight lines to previous flight lines of the same flight on orthomosaics. As wind is one of the most important environmental drivers of sensor temperature, Kelly et al. (2019) and Yuan and Hua (2022) examined the relation between wind and sensor temperature while Malbéteau et al. (2021) showed how different wind conditions result in different CT estimates.

Perich et al. (2020) used uncorrected orthomosaics to extract plot-based values. They then proposed including spatial correction with the R-package SpATS (Rodríguez-Álvarez et al., 2018) to account for spatial and temporal trends simultaneously in a subsequent step. However, they observed that temporal effects of rapidly changing environmental conditions remain a challenge. While parts of temporal effects are absorbed in the spatial correction process and confound the spatial trend and its interpretability, others remain uncorrected and bias the TIR signal. To overcome these limitations, temporal effects need to be mitigated when creating the orthomosaics, as done by Malbéteau et al. (2021), Mesas-Carrascosa et al. (2018) or Z. Wang et al. (2023) prior to orthomosaic analysis. While promising correction approaches exist for orthomosaics, they are often based on assumptions such as the similarity of surface temperature within a specific land cover type (Z. Wang et al., 2023). Such assumptions are not valid in wheat variety testing as CT variances are examined within the same land cover type. In addition, any artifacts of an erroneous correction are propagated to the analysis in orthomosaics but the information on the correction applied is not available with the final CT estimate.

To the best of our knowledge, airborne TIR imaging in agriculture was either based on single images (*e.g.* Deery, Rebetzke, Jimenez-Berni, James, et al., 2016) or orthomosaics, *i.e.* large composite images of a series of images with large overlap (*e.g.* Das, J. Christopher, Apan, Choudhury, et al., 2021; Francesconi et al., 2021; Malbéteau et al., 2021; Messina and Modica, 2020; Perich et al., 2020; Z. Wang et al., 2023). The advantages and disadvantages of these methods are discussed in Perich et al. (2020). In short, single images are limited in resolution, and therefore only a limited land surface can be captured at once. When creating orthomosaics, the information of multiple images has to be blended into a single large orthomosaic and the different blending methods may lead to different results (Aasen and Bolten, 2018; Malbéteau et al., 2021; Perich et al., 2020). Furthermore, information is lost in the aggregation process,

as spots that appear on multiple images with specific viewing geometries are blended into a single pixel on the orthomosaic.

An alternative is to skip the orthomosaic processing step and work with original image sequences, a novel method for thermal imaging proposed in this study. To avoid the loss of information in the orthomosaic blending process, Roth, Aasen, et al. (2018) developed a method to analyze RGB drone images without the need to merge individual images into an orthomosaic. Single images can be examined with respect to trigger timing and the geometric relations between the experimental unit, sun stand, and drone position. Transferring such an approach to thermal imaging will provide the means to analyze sources of variance in CT for field experiments. With multi-view imaging, temporal and geometric trends are not disregarded at the creation of orthomosaics but are used to improve the statistical analysis of CT data. The information on the correction applied is available with the final CT estimate and can be consulted when results are inconsistent. All together, multi-view imaging enables to handle confounding factors that affect the interpretation of CT. Such an informed analysis is crucial in variety testing and breeding as temporal, spatial, and geometric trends of CT might mask effects of genotypes or treatments otherwise. By estimating the different sources of variance, they can be corrected for, revealing the actual effects of the experiment that are of interest.

Mixed models are a widely used statistical tool to separate and estimate different sources of variance in agronomic trials (*e.g.* Gilmour et al., 1997; H. P. Piepho and E. R. Williams, 2010; Hans-Peter Piepho et al., 2012). Estimating continuous covariate effects such as spatial or temporal trends is often done with auto-regressions and/or smoothing splines (*e.g.* Cullis, A. B. Smith, et al., 2006; Rodríguez-Álvarez et al., 2018; Velazco et al., 2017). It is hypothesized that post-processing multi-view images with mixed models will improve CT measurements on wheat in plot experiments. The step of correcting an orthomosaic in pre- or post-processing can be skipped. Instead, the correction can be integrated in the analysis of the experiment directly, using common tools to analyze designed experiments, namely, mixed models.

In addition to including covariates in the estimation of CT, knowing the viewing geometry for each measurement allows for the selection of measurements with preferable viewing geometries. Das, S. C. Chapman, et al. (2021) and Pask et al. (2012) described the impact of soil on the measurement of apparent CT. It is hypothesized that by selecting for oblique (*i.e.* less vertical) viewing angles and measurements perpendicular to the sowing row direction, the fraction of plants visible in TIR images can be increased, and the influence of soil on measurements can be reduced.

This study sought to improve the measurement of genotype related CT variance in the context of wheat variety testing by a drone-based thermography lean-phenotyping approach. TIR images from an affordable uncooled and uncalibrated off-the-shelf TIR camera were georeferenced and information on trigger timing and on geometric relations between the sun, the region of interest (ROI) and the drone was exploited in a multi-view approach. It was tested if the integration of such temporal and geometric covariates in mixed models allows to account for the different sources of variance of CT measurements and thereby to correct for unwanted sources of variance. We hypothesized that this correction enables an improved quality of thermal measurements in terms of consistency and heritability with relatively simple equipment and without the need for in-field reference procedures.

2.2 Methods

2.2.1 Field experiments and data acquisition

TIR measurements were conducted on wheat variety testing experiments of winter wheat for two consecutive years (2020–2021 and 2021–2022) on fields of the agricultural research station of Agroscope, at Changins, Switzerland [46°23′55.4″N 6°14′20.4″E, 425 m.a.s.l., the World Geodetic System (WGS) 84]. The soil of the experimental site is a shallow Calcaric Cambisol (Baxter, 2007; Cárcer et al., 2019).

Air temperature, rainfall, radiation, wind speed, wind direction, relative humidity and VPD were obtained from a weather station of Meteoswiss which was located about 800 m from the experimental site at Changins [46°24′3.7″N 6°13′39.6″E, 458 m.a.s.l., WGS 84].

The two years showed very contrasting weather conditions (Fig. S1.2). While 2021 was a relatively cool year with almost 700 mm of precipitation from the beginning of the year to harvest, there was just 280 mm precipitation for the same period in 2022. The average temperature between beginning of May and harvest was 2.9 °C warmer in 2022 than 2021. Therefore, wheat developed faster in 2022 and heading and harvest occurred earlier.

The measurement periods were between onset of heading and early senescence. The trial comprised 30 modern registered European winter wheat varieties and is further referred to as the EuVar trial. The same varieties were sown over the two years. Three treatment regimes were applied to these genotypes in both years. In the "maximal" treatment, one growth regulator and one fungicide treatment were applied. In the "medium" treatment, there was just the growth regulator application and not the fungicide application. In the "minimal" treatment, neither a growth regulator nor a fungicide were applied. Fertilization and herbicides were applied according to the Proof of Ecological Performance (PEP) certification guidelines (Swiss Federal Council, 2013), which represent a minimal standard for best practice conventional agriculture in Switzerland. Each variety-treatment combination was repeated three times in plots of 1.05 m x 8 m each. Each plot contained eight sowing rows of the same wheat genotype with a spacing of 15 cm between them. The genotypes were randomly distributed within blocks of 3 by 10 plots and these blocks randomly nested within three treatment replicates. Each treatment replicate contained three blocks and every block was treated with one of the three treatments. The 270 plots of the experiment span over 27 rows (which followed tractor track direction) and 10 columns (Fig. S1.1).

The two experiment-year combinations are further referred to as EuVar21 and EuVar22 according to year of harvest. Table S1.1 gives an overview on the different treatments and the most important field interventions and Table S1.2 displays details of the chemical products used.

Flights were conducted between onset of flowering and early senescence at two and four dates in 2021 and 2022 respectively. On specific dates, multiple flights were conducted at different time slots. To account for short term variability, within each time slot at least two, mostly three flights were conducted with the same settings. A group of flights that were conducted at one time slot and date is further called a flight campaign. In total, 39 flights were performed (for more details, see Supplementary Materials section S1.5).

A description of the equipment and the settings used and of the flight planning can be found in Supplementary Materials section S1.6. Heading of drone and TIR camera remained relatively stable throughout the flight and did not change with flight path direction changes. The resulting flight duration was between 7 and 9 min depending on wind conditions and the total area recorded. The experiments were neighbored by border plots and other experiments. To fully profit from the advantages of the methodology proposed in this study, flights covered not just the experiments but all wheat plots in the respective field surroundings, *i.e.* border plots and other experiments on the same field. This allowed to reduce border effects by taking

advantage of temporal and spatial corrections, as will be described later on. Supplementary Materials section S1.7 summarizes the pre-flight procedure. In short, the camera was turned on 15 min before each flight in 2021 and 30 min in 2022 to allow the temperature signal to stabilize. The TIR images were saved as radiometric JPEG format.

For post-processing in the Structure-from-Motion-based photogrammetry software Agisoft Metashape (Agisoft LCC, St.Peterburg, Russia) and to allow time series analysis, thermal ground control points (GCPs) were distributed in the field in an evenly spaced shifted grid pattern (for more details, see Supplementary Materials section S1.8).

For the multi-view approach, digital elevation models (DEM) were needed on which the images could be projected. TIR images often do not provide enough spatial detail to generate DEMs with sufficient quality (*e.g.* Malbéteau et al., 2021). TIR based DEMs may appear flat with no distinct plot pattern. Therefore, flights were also conducted with a Micasense RedEdge-MX Dual multispectral sensor, which allows for more spatial detail. Although this sensor produces multispectral data with 10 bands, only the RGB bands were used for this study, and the data is further referred to as RGB data.

2.2.2 TIR data processing overview

The multi-view approach allowed to include covariates such as trigger timing and viewing geometry parameters of single measurements in the analysis. To examine if this allowed to better compensate for temporal and spatial trends, different multi-view approaches were compared to the standard orthomosaic approach (Fig. 2.1). First, TIR images were georeferenced. TIR data was then extracted from georeferenced orthomosaics as well as georeferenced single images. For the multi-view approach, trigger timing was extracted along with covariates related to viewing geometry for each plot on each image (green section in Fig. 2.1). TIR data was then treated by different statistical approaches (blue section) and the approaches were compared to each other (violet section).

2.2.3 TIR image pre-processing

Radiometric JPEG format contains an 8-bit gray scale JPEG image as well as a 14-bit array with digital numbers (DN), which represent the magnitude of TIR radiation (Kelly et al., 2019). The DNs in the 14-bit arrays of the radiometric JPEGs were transformed to TIFF files representing temperature in $°C$ x 1000 by using a Python 3.8 script (van Rossum, Guido and Drake, Fred L., 2009) and a modified version of the Flir Image Extractor (`https://github.com/ITVRoC/FlirImageExtractor`), which allowed for batched processing.

The 14-bit TIFF files of the radiometric image as well as the RGB images were aligned in the structure-from-motion-based software Agisoft Metashape Professional (Agisoft LLC, St. Petersburg, Russia) and georeferenced (for details, see Supplementary Materials section S1.9). Plot masks were created for each plot in Qgis 3.16 (QGIS Development Team, 2022), to determine the ROIs from which data was used for analysis. To account for border effects in the field and for inaccuracies of georeferencing and superimposition of different flights, a border buffer of 25 cm was applied to all masks on plot width. On plot length, the buffer was up to 1 m, leaving at least a surface of 2.1 m^2 to be analyzed in each plot. The plot masks were saved to GeoJSON format.

Imaging techniques deliver pixel values in a 2-D space. In order to evaluate experimental units, pixels within ROIs in this 2-D space must be analyzed. Usually, this is done using zonal statistics, *i.e.*, the pixels within ROIs are reduced to single values using statistical aggregation functions. In this work, an empirically determined specific percentile for each year was used.

As selection criteria for percentile determination, generalized heritability (Oakey et al., 2006, Eq. 2.10, Eq. S1.1, Eq. S1.2) of different percentiles was calculated for each flight. The

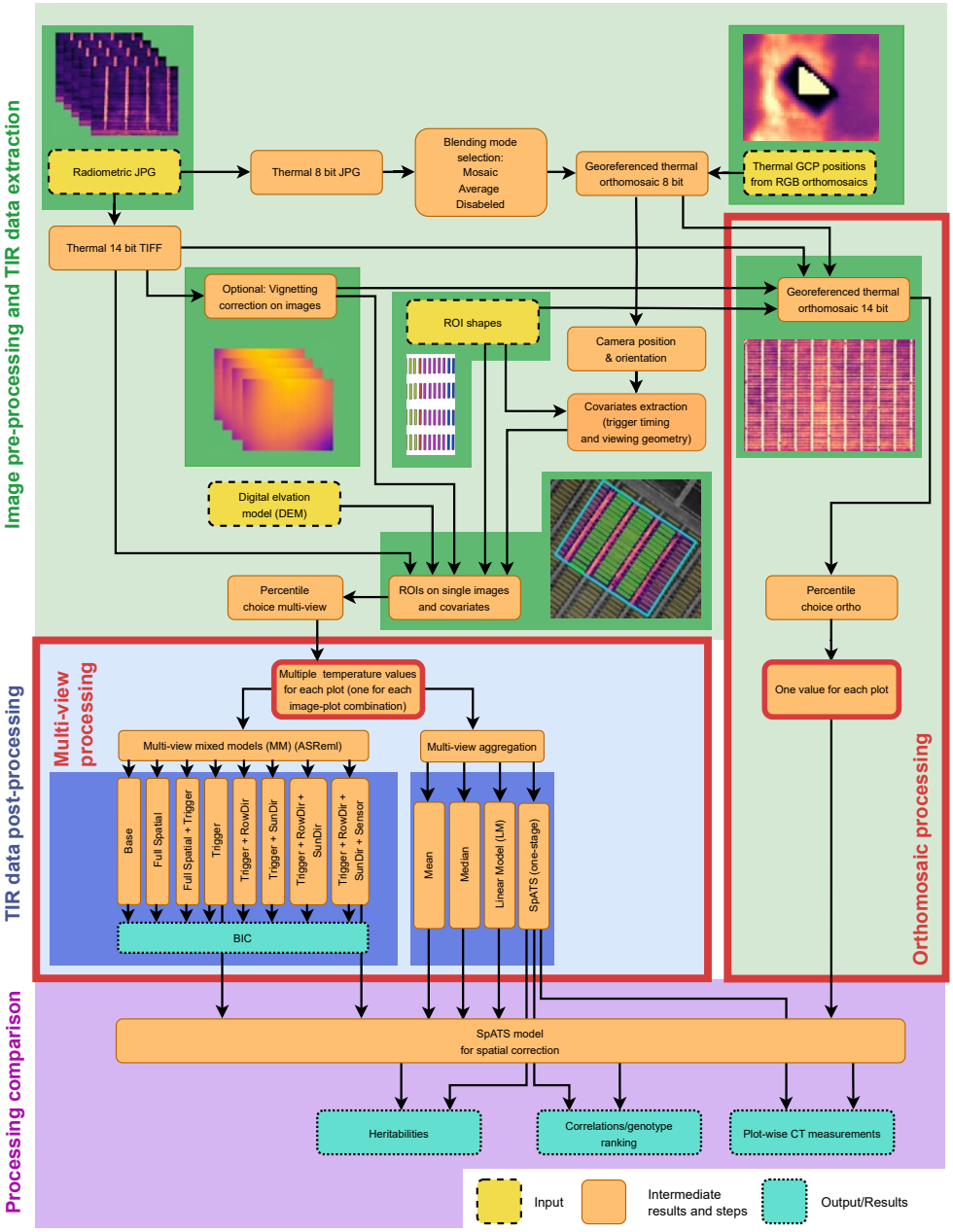


FIGURE 2.1: Overview on the different steps of TIR data processing methods that were compared in this study. Orthomosaics were composed by different blending modes. After image pre-processing (green section), TIR information was analyzed on orthomosaics or with different multi-view approaches (blue section). Plot values were estimated based on multi-view data by using mixed models of different complexity. In addition, multi-view data was aggregated to plot values by relatively simple aggregations methods. The results were compared to each other by means of correlation, genotype ranking consistency and heritability of plot-wise apparent CT (violet).

values within the ROIs were reduced to a single value by using the respective percentile. For each percentile, heritabilities were calculated in SpATS (Rodríguez-Álvarez et al., 2018), which is an easy-to-use tool for spatial analysis commonly used in agricultural research and thermography (Anderegg, Yu, et al., 2020; Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Perich et al., 2020), which also includes a mixed model for experimental design factors. The resulting percentile-heritability relations were plotted for graphical comparison. Two quantitative criteria were used to select the percentiles: Select a percentile in the center of a percentile region where (1) the heritability is close to the maximum, and (2) closely adjacent percentiles have similar heritabilities, *i.e.* the heritability is stable in the respective percentile region. For each year, the optimal percentile was determined. The values within the ROIs were reduced to a single value by using the optimal percentile for all flights within one year. One value per plot was then used as plot-wise CT value in further analysis.

The internal temperature of the sensor is constantly changing, due to the interplay of heating sensor electronics and an ever changing exposure to sun and wind during flights. This is leading to constantly changing non-uniformity effects which mix up with vignetting and distort TIR images (Kelly et al., 2019; Yuan and Hua, 2022).

Yuan and Hua (2022) proposed to use a single image taken shortly after a drone flight with a TIR sensor to correct for these effects. We considered this procedure too complex for day-to-day operations. Instead, it was tested if a simplified, overall vignetting mitigation could improve measurement quality. To that end, a mean overall vignetting effect was estimated by calculating a mean vignetting effect over 413 images in an indoor experiment (procedure described in detail in Supplementary Materials section S1.10). The image corresponding to a mean estimated vignetting effect was subtracted from all the TIR images to get vignetting-corrected images (*e.g.* Figs. S1.3 & S1.4). Subsequent analysis was conducted on images with and without vignetting correction for both, the orthomosaic and multi-view methods.

2.2.4 DEM creation

DEMs were created on the basis of aligned images in Agisoft Metashape and could be derived from thermal data in 2021, but not in 2022. Therefore, DEMs in 2022 were generated from RGB data. Both methods allowed generating DEMs of sufficient positioning precision (positioning RMSE vertical: 2.5 cm, horizontal: 1.5 cm based on Agisoft alignment error estimates for ground control points). For each year, a representative DEM was chosen that was created from images taken after the wheat stem elongation phase and before early senescence, when the canopy height remained stable. The quality of the DEMs was checked by visually inspecting the plausibility of the positioning of the masks projected on single images in multi-view pre-processing. The projected masks needed to be centered within plots and of rectangular shape (*e.g.* Fig. 2.2). In 2021, the DEM was based on the second flight of the thermal campaign flown on 2021-06-12 at 12:30, at a flight height of 40 m. The ground sampling distance (GSD) of the TIR images was 5.15 cm/pix and the spatial resolution of the DEM was 41 cm/pix. With this coarse resolution, inconsistencies such as holes in the DEM could be leveled out. The DEM used in 2022 was based on the data generated on 2022-06-04 with the Micasense sensor at a flight height of 40 m. The GSD of the images was 2.71 cm/pix. The DEM did not exhibit holes and the spatial resolution of the DEM was set to 2.71 cm/pix too.

2.2.5 Orthomosaic pre-processing

TIR orthomosaics were created by the three blending modes available in Agisoft Metashape, as described in the Agisoft Metashape professional edition user manual (Agisoft, 2023):

- Mosaic: A two-step approach where larger features are composed based on multiple images while details are taken from a single image.
- Average: A weighted average for all pixels on the orthomosaics.
- Disabled: Pixels are taken from a single close-to-nadir image.

The blending modes in orthomosaic composition were compared to each other by means of generalized heritability (Oakey et al., 2006) similar to Perich et al. (2020). TIR data was aggregated within ROIs by multiple percentiles and heritabilities were calculated for multiple percentiles on each flight for the three different blending modes. The resulting percentile-heritability relations of the three blending modes were plotted for graphical comparison.

The best performing blending mode was then applied to determine the optimal percentile for data aggregation by zonal statistics. The percentile-heritability relations were analyzed on all flights within one year. The optimal percentile for each year was applied for all flights within this year.

2.2.6 Multi-view pre-processing

The camera positions (longitude, latitude, height) and orientations (pitch, roll, yaw) at the moment of triggering for the single images were estimated in an indirect sensor orientation approach (Benassi et al., 2017) in Agisoft Metashape after aligning images. Using the previously estimated trigger positions, the single images were projected on the DEMs (Fig. 2.2) by ray tracing as described in Roth, Aasen, et al. (2018) and Roth, Camenzind, et al. (2020). This allowed to project geographic coordinates (*e.g.* EPSG:2056 reference system) to image coordinates. As a result, plot masks of ROIs were created for each trigger position (*i.e.* for each image) where at least one plot was entirely inside the field of view (FOV) of the camera. As coordinates were identical for 8-bit JPEG images and 14-bit intensity value arrays, the image-wise masks could directly be applied to the temperature TIFF files. This approach of identifying the ROIs for each plot on every single image is referred to as multi-view. For each plot on each TIFF file, all percentiles were extracted with a Python 3.8 script and saved to a CSV file.

As plot-wise data was extracted for each image, the trigger timing could be determined from image meta data. By knowing the trigger timing of each image and the position of the experiment, the position of the sun could be determined as azimuth and elevation angle in Python using a script by John Clark Craig (`https://levelup.gitconnected.com/python-sun-position-for-solar-energy-and-research-7a4ead801777`, 2021). As Cartesian (*i.e.* orthogonal) coordinates were used and the position of the sun, the position of the plot centers and the position and orientation of the camera at the moment when the image was triggered were known, this allowed to calculate the geometric relations between sun, plot and drone by trigonometry as listed in Table 2.1 and illustrated in Fig. 2.3.

b

2.2.7 TIR data post-processing

After data extraction, TIR data was processed by different methods with the aim of finding a robust, yet simple processing method for TIR multi-view data (blue section of Fig. 2.1). In the following, the different processing steps of the different methods are described. The presentation of single steps follows the structure of Fig. 2.1. TIR data was processed with the standard orthomosaic method which served as a baseline. This method was compared to several multi-view methods, starting with relatively simple multi-view aggregation and going to approaches including statistical models of increasing complexity to estimate plot-wise CT.

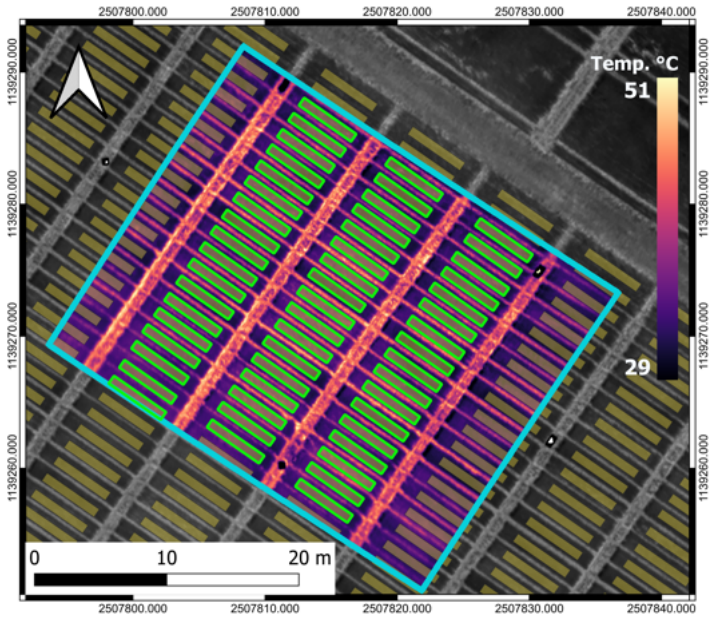
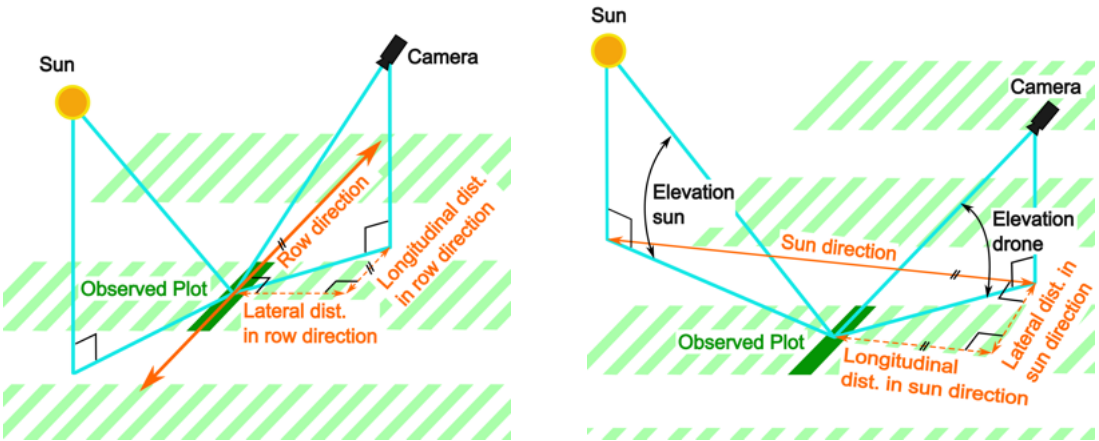


FIGURE 2.2: Example of a TIR image, projected on a DEM. The DEM (gray-scale, in the background) defined the surface on which the TIR image (blue margin) was projected on. Plots were defined for the whole field (shaded in yellow). Plot shapes that fell entirely within the extent of the TIR image (green margins) were projected to image coordinates and all plot-wise TIR percentiles were extracted.



(a) Plot position in row direction

(b) Plot position in sun direction

FIGURE 2.3: By knowing the position of the sun, the position of the plot and the position and orientation of the camera when an image is triggered, different geometric relations can be calculated, such as the position of the plot relative to the drone in row (or sowing) direction (a) or relative to the sun (b). The dimensions of interest and related covariates are shown in orange. Important angles related to drone and sun are named. Small black angle marks and short parallel black lines indicate perpendicularity and parallelism, respectively.

TABLE 2.1: List of covariates calculated from multi-view data and used in the mixed model.

Covariate	Description	Metric
Trigger timing	The time stamp when each TIR image was taken	seconds from start of flight
Lateral dist row dir.	Lateral distance of the plot relative to the drone in sowing row direction	meters from planar position of drone
Lateral dist sun dir.	Lateral distance of the plot relative to the drone in sun direction (*i.e.* orthogonal to principle plane of the sun)	meters from planar position of drone
Longitudinal dist row dir.	Longitudinal distance of the plot relative to the drone in sowing row direction	meters from planar position of drone
Longitudinal dist sun dir.	Longitudinal distance of the plot relative to the drone in sun direction (*i.e.* in the principle plane of the sun)	meters from planar position of drone
Sensor x	X coordinate of the plot center on the sensor plane (image coordinates)	pixel no. in x from bottom-left
Sensor y	Y coordinate of the plot center on the sensor plane (image coordinates)	pixel no. in y from bottom-left

2.2.7.1 Multi-view simple aggregation post-processing

The orthomosaic method only yielded one value per plot to be analyzed in a final statistical analysis. In contrast, the multi-view method provided several values for each plot (originating from different images), which were aggregated by different methods prior to final analysis. As shown in Fig. 2.1 (blue section), the simplest way to aggregate values from different images j to plot values θ_p is by calculating the mean or median of all measurements per plot p, and ignoring the effects of genotype (i), treatment (k) and replication (n),

$$\hat{\theta}_{p\_mean} = \text{mean}(\theta_{jp})\,. \tag{2.1}$$

$$\hat{\theta}_{p\_median} = \text{median}(\theta_{jp})\,. \tag{2.2}$$

A more complex way is to correct for the effect of trigger timing with a simple linear model simultaneously for all images (Eq. 2.3), *e.g.*, in Base-R (R Development Core Team, 2022). Here, the measured temperature θ_{jp} of the p^{th} plot on the j^{th} image is composed of an estimated image effect ν_j, a plot effect ϕ_p and an error e_{jp}, ignoring i, k and n,

$$\theta_{jp} = \nu_j + \phi_p \ + e_{jp}. \tag{2.3}$$

The image effect ν_j estimates an image-specific CT contribution at trigger time j. ϕ_p corresponds to a plot-specific CT contribution of the p^{th} plot. In the model fitting process, the error term is minimized by varying the estimated values of the two effects. The temporal variance is assumed to be attributed to the image effects. The estimated plot effect can then be used for further processing as estimate of relative plot CT without the temporal effect,

$$\hat{\theta}_{p\_LM} = \phi_p. \tag{2.4}$$

For the simple linear model, further denoted LM, all the plots within the fields were analyzed. Different experiments were covered as well as border plots.

2.2.7.2 Multi-view mixed models post-processing

The repeated plot-values originating from the multi-view method allow to model relations to geometric covariates and trigger timing. Including these relations might increase the explained variance of TIR measurements.

To include these covariates, mixed models were applied. With mixed models, a response variable can be modeled by explanatory categorical factors, covariates and an error term representing variance that cannot be explained by the model. The data is clustered according to categorical factors, and regression parameters in mixed models can be cluster specific as well. This enables for example the modeling of genotype- and treatment-specific responses. The factors can either be fixed or random. Within the random factors, effects are cluster-specific. Fixed factors have fixed effects, and regression parameters apply to the whole population at observation (Hartung and H.-P. Piepho, 2007; Wu, 2010).

Mixed models of varying complexity were fitted in ASReml-R (Butler, 2019). The parameters of the mixed model (Eq. 2.5) are explained in Table 2.2. The terms were grouped by types of terms ("Design-Factor, "Spatial-Autoregression", "Spatial-Smoothing-Spline", etc.). Note that not all models included all term types. Table 2.3 describes the different models and which term types were included in each model. The index j is written in parentheses to represent both, models that do consider trigger timing and those that do not ("MM Base", "MM Full spatial").

Modeling started with the baseline "MM Base" model where only experimental design factors (genotype, treatment, replicate, plot position, plot) were included. This model was then increased in complexity by iteratively including a subset of additional factors as well as temporal and geometric covariates (Table 2.1). This led to nested models where simpler models were fully included in more complex models, culminating in the most complex model,

$$
\begin{aligned}
\theta_{i(j)knp} = {} & \theta_i + \tau_k + \phi_p + r_n + (\tau r)_{kn} + && \text{(Design-Factors)} \\
& (\alpha\beta)_{c(p)r(p)} + \alpha_{c(p)} + \beta_{r(p)} + && \text{(Spatial-Autoregression)} \\
& f_{\text{spl}\times\text{spl}}\big(c(p), r(p)\big) + f_{\text{spl}}(c(p)) + f_{\text{spl}}(r(p)) + && \text{(Spatial-Smoothing-Spline)} \\
& f_{\text{spl}}(j) + && \text{(Temporal-Trend)} \\
& f_{\text{spl}\times\text{spl}}(\lambda_{\text{lon,Row,jp}}, \lambda_{\text{lat,Row,jp}}) + && \text{(Row-Direction-Trend)} \\
& f_{\text{spl}\times\text{spl}}(\lambda_{\text{lon,Sun,jp}}), \lambda_{\text{lat,Sun,jp}})) + && \text{(Sun-Direction-Trend)} \\
& f_{\text{spl}\times\text{spl}}(s_{x,jp}, s_{y,jp}) + && \text{(Sensor-Plane-Trend)} \\
& e_{i(j)knp} && \text{(Residuals)}
\end{aligned}
\tag{2.5}
$$

Just like with the LM, the CT was assumed to be influenced by categorical factors. In contrast to the LM, more than two factors were included. These factors and their rationale are described in the following.

In addition to the plot effect, the design factors included genotypes, treatments, replications, and an interaction between treatment and replication, since treatments could react differently within replications. For the spatial part, an effect of the spatial coordinates, described as columns $c(p)$, rows $r(p)$ and their interaction (*i.e.*, a two-dimensional grid) was assumed to impact the CT values. This impact was assumed to be autocorrelated, *i.e.* the spatial effect of the plot at a specific position was assumed to be more closely related to that of its neighbor plot than to a more distant plot. The "Full Spatial" model contained, in addition to autocorrelated effects, a spatial model, assuming the effects of columns and rows to follow independent smoothing splines in both directions, and in addition a two-dimensional smoothing spline in both directions. With the "Full Spatial" model, it was tested whether a model with more degrees of freedom in the spatial dimension provides a better fit.

In addition to design factors, temporal and geometric covariates were added. The temporal trend, defined along the trigger timing in seconds after the start of the respective flight, was modeled as a smoothing spline. Geometric covariates for three geometric dimensions were included as three independent two-dimensional smoothing splines. The first two dimensions, "Row-Direction-Trend" and "Sun-Direction-Trend" (Fig. 2.3), represented the position of the

plot below the drone, described in a Cartesian coordinate system with the drone position defined as the origin of the coordinate system. x and y of the coordinate system were the lateral and longitudinal distances in the respective dimension. The third geometric dimension, "Sensor-Plane-Trend", described the position of the plot center on the image with x and y coordinates, where the origin was bottom left of the image.

The models were fitted for every flight separately, as the impact of covariates was assumed to vary between flights. As for the LM, all plots within the fields were analyzed. To account for this in mixed models, varieties were given unique names within each experiment, so the same variety name did not appear in two different experiments, which reduced the complexity of the models. A simple additive effect for treatments was assumed for the estimation of plot-wise CT as some models with an interaction between treatments and genotypes proved to be too computationally intensive at this stage.

With the Bayesian information criterion (BIC), the quality of the model fit was compared. BIC was preferred over pure likelihood as it penalizes complex models and therefore over-fitting. It was also preferred over the Akaike Information Criterion (AIC) as BIC penalizes complex models with redundant variables stronger than AIC. Lower BIC values indicate preferable models (Schwarz, 1978; Stoica and Selen, 2004).

After fitting the models (Eq. 2.5), plot-wise CT values were estimated in a similar approach as for the other, simpler models (Eq. 2.1, 2.2 & 2.4). Specifically, $\hat{\theta}_{p\_MM}$ were estimated as sum of genotype effects (θ_i), treatment effects (τ_k), plot effects (ϕ_p), and replication effects (r_n),

$$\hat{\theta}_{p\_MM} = \theta_i + \tau_k + \phi_p + r_n \ . \tag{2.6}$$

As with the LM, the term related to the temporal trend $f_{\text{spl}}(j)$ was not included in the prediction. In addition, terms related to spatial effects of columns or rows and geometric trends were discarded. $\hat{\theta}_{p\_MM}$ therefore represents the plot values corrected for temporal or geometric trends and for spatial trends related to columns and rows.

2.2.8 Methods comparison

The final results of the different methods were single plot values per flight. To compare the quality of the different methods, the plot-wise CT values were compared to each other after a spatial correction (Fig. 2.1, violet section).

The plot values of the orthomosaic method, the aggregated plot-wise multi-view values (Eq. 2.1, 2.2), the multi-view values estimated with the LM (Eq. 2.4) and the plot-wise results from the CT estimations with the mixed models (Eq. 2.6) were first fitted with a spatial model (Eq. 2.7) in the R package SpATS (Rodríguez-Álvarez et al., 2018). Because of the low absolute temperature accuracy of uncooled and uncalibrated TIR cameras, the retrieval of accurate absolute CT is very challenging, especially if larger field trials are covered (H. G. Jones, Serraj, et al., 2009; Kelly et al., 2019). Therefore, relative temperature differences were analyzed, as relative temperature differences are commonly used for the grading of plant performance, assuming that CT rankings are reproducible and consistent between measurements under similar conditions (H. G. Jones, Serraj, et al., 2009; Prashar and H. Jones, 2014; Das, J. Christopher, Apan, Roy Choudhury, et al., 2021).

Just CT estimates of plots belonging to EuVar were used as input to the SpATS-models and plots of other experiments and border plots were skipped at this stage.

TABLE 2.2: Terms of the mixed models (Eq.2.5). Note that not all term types are used in all models.

Term type	**Term**	**Description**	**Part**
Design-Factors:	θ_{i}	Genotype effect of the i^{th} genotype (unique for each experiment within field)	Random
	τ_{k}	Treatment effect of the k^{th} treatment (unique for each experiment within field)	Fixed
	ϕ_{p}	Effect of the p^{th} plot	Random
	r_{n}	Effect of the n^{th} replication	Random
	τr_{kn}	Interaction of the k^{th} treatment and the n^{th} replication	Random
Spatial-Autoregression:	$(\alpha\beta)_{c(p)r(p)}$	Two-dimensional spatial autocorrelation model based on row and column position in the field	Random
	$\alpha_{\mathrm{c(p)}}$	One-dimensional autocorrelation model for columns in the field (orthogonal to tractor track direction)	Random
	$\beta_{\mathrm{r(p)}}$	One-dimensional autocorrelation model for rows in the field (in tractor track direction)	Random
Spatial-Smoothing-Spline:	$f_{\mathrm{spl}\times\mathrm{spl}}\big(c(p), r(p)\big)$	Two-dimensional spatial smoothing spline model based on row and column position in the field	Random
	$f_{\mathrm{spl}}(c(p))$	One-dimensional smoothing spline model for columns in the field (orthogonal to tractor track direction)	Random
	$f_{\mathrm{spl}}(r(p))$	One-dimensional smoothing spline model for rows in the field (in tractor track direction)	Random
Temporal-Trend:	$f_{\mathrm{spl}}(j)$	Trigger timing smoothing spline along the j sequential trigger events	Random
Row-Direction-Trend:	$f_{\mathrm{spl}\times\mathrm{spl}}(\lambda_{\mathrm{lon,Row,jp}}, \lambda_{\mathrm{lat,Row,jp}})$	Two-dimensional spatial smoothing spline model based on longitudinal and lateral distance of the plot relative to the drone in row direction	Random
Sun-Direction-Trend:	$f_{\mathrm{spl}\times\mathrm{spl}}(\lambda_{\mathrm{lon,Sun,jp}}), \lambda_{\mathrm{lat,Sun,jp}}))$	Two-dimensional spatial smoothing spline model based on longitudinal and lateral distance of the plot relative to the drone in sun direction	Random
Sensor-Plane-Trend:	$f_{\mathrm{spl}\times\mathrm{spl}}(s_{x,jp}, s_{y,jp})$	Two-dimensional spatial smoothing spline model based on plot center position on the sensor plane of the thermal sensor in x and y (image coordinates)	Random
Residuals:	$e_{\mathrm{i(j)knp}}$	Residual term for the i^{th} genotype, the j^{th} trigger event, the k^{th} treatment, the n^{th} replication and the p^{th} plot	Random

TABLE 2.3: Term type combinations used in plot-wise CT estimation with mixed models (Eq.2.5). Starting with a simple "Base" model, models increase in complexity further down by including different sets of term types. For detailed information on the terms in each term type, see Table 2.2. The prefix "M" has been omitted in mixed model names in the table for simplicity.

Mixed model (MM)	**Term types**	**Description of model**
Base	Design-Factors + Spatial-Autoregression + Residuals	Includes the experimental design (genotypes, treatments, replications) and a simple spatial model.
Full Spatial	Design-Factors + Spatial-Autoregression + Spatial-Smoothing-Spline + Residuals	The "Base" model enhanced by a complex spatial model in the style of Velazco et al. (2017) which includes a random term for each row and column, an auto correlated interaction term and a bivariate smoothing spline between the two.
Full spatial + Trigger	Design-Factors + Spatial-Autoregression + Spatial-Smoothing-Spline + Temporal-Trend + Residuals	"Full spatial" model enhanced by the temporal dimension of trigger timing.
Trigger	Design-Factors + Spatial-Autoregression + Temporal-Trend + Residuals	The "Base" model enhanced by the temporal dimension of trigger timing.
Trigger + RowDir	Design-Factors + Spatial-Autoregression + Temporal-Trend + Row-Direction-Trend + Residuals	Integrates the relative position of the plot in row (*i.e.* sowing) direction in the "Trigger" model.
Trigger + SunDir	Design-Factors + Spatial-Autoregression + Temporal-Trend + Sun-Direction-Trend + Residuals	Integrates the relative position of the plot in sun direction in the "Trigger" model.
Trigger + RowDir + SunDir	Design-Factors + Spatial-Autoregression + Temporal-Trend + Row-Direction-Trend + Sun-Direction-Trend + Residuals	Integrates the "Trigger + RowDir" and the"Trigger + SunDir" models into one model.
Trigger + RowDir + SunDir + Sensor	Design-Factors + Spatial-Autoregression + Temporal-Trend + Row-Direction-Trend + Sun-Direction-Trend + Sensor-Plane-Trend + Residuals	Integrates the spatial dimensions of the sensor plane (image coordinates) in the "Trigger + RowDir + SunDir" model.

$$\begin{aligned}\hat{\theta}_p = \hat{\theta}_{iknp} = \theta_i + \tau_k + \left(\theta_n\tau_n\right)_{ik} + \phi_p + r_n + & \quad \text{(base model)} \\ \tau r_{kn} + & \quad \text{(repl. } \times \text{ treat. (just EuVar))} \\ f\big(c(p), r(p)\big) + \psi_{c(p)} + \psi_{r(p)} + & \quad \text{(spatial model)} \\ e_{iknp} & \quad \text{(error term)}\end{aligned} \tag{2.7}$$

A smooth bi-variate surface which was defined by the positions of the plots within columns and rows ($f(c(p), r(p))$) was included in the model together with a random effect for columns and rows ($\psi_{c(p)} + \psi_{r(p)}$). With SpATS models just covering plots of respective experiments, they included an interaction between the i^{th} genotype and the k^{th} treatment $(\theta_n\tau_n)_{ik}$. The remaining terms were equal to the terms in Eq. 2.5 and can be looked up in Table 2.2. While ASReml-R also provides the functionality to calculate heritabilities and predict single plot values, the inclusion of the full experimental design as in Eq. 2.7 in one stage proved to be too computationally intensive due to the interaction term $(\theta_n\tau_n)_{ik}$. Therefore, the two-stage approach for the mixed models with a subsequent analysis in SpATS was applied, but in contrast to the simpler methods in the comparison, most of the spatial correction was done within the mixed model before SpATS spatial correction. This two-stage approach furthermore allows a full comparability of the mixed model approach with simpler methods since all approaches relied on the SpATS model.

From the SpATS formula, plot-wise values are predicted as genotype effect θ_i, treatment effect τ_k and the error e_{iknp}, where the error represents variance that could not be explained with the SpATS model,

$$\hat{\theta}_{p\_SpATS} = \theta_i + \tau_k + e_{iknp}\ . \tag{2.8}$$

To test the quality of CT estimates, Pearson correlation, genotype rank consistency, and heritability were used as quantitative criteria, as done in other studies (Oakey et al., 2006; H. G. Jones, Serraj, et al., 2009; Rodríguez-Álvarez et al., 2018).

The correlations were calculated between flights within years. To avoid inflated correlations, dominant treatment effects were removed before correlation calculations by subtracting estimated treatment effects from plot-wise CT values. If measured under similar conditions, high correlations between flights taken close to each other are indicative of the consistency of the method, which means that the ranking of CT estimates remains similar between two flights. The correlation between flights within the same campaign is therefore an important criterion of consistency and quality. High correlations between flights taken at distinct times or dates, *i.e.* between different campaigns , are indicative of CT consistency as a measurement over time. The consistency between campaigns might be affected by changes in meteorological conditions, but also phenology, when taken at different dates. Although strong correlations might also be expected between campaigns, they are, therefore, less indicative of the consistency of the used method itself than correlations within campaigns.

Along with the correlations and CT ranking, the genotype ranking consistency between flights within treatments allows for robust conclusions about genotypes' CT. To capture this measure quantitatively, the measurement means per genotype were ranked for each flight within each treatment, and the consistency of the genotype ranking was examined as the standard deviation (sd) of the genotype ranking throughout the flights of one campaign, defined as:

$$\sigma_{gen\_r} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{2.9}$$

where x corresponds to the ranking of a genotype mean of one flight i, $\overline{x}$ to the mean of the genotype rank of that respective genotype across all flights within one campaign and n to the total number of the flights within one campaign. The sd of genotype ranking σ_{gen_r} (Eq. 2.9) provided a tangible metric of ranking consistency, and a lower value indicated greater consistency. σ_{gen_r} was calculated within the three treatments separately. One value was calculated for each genotype within each treatment for all campaigns of selected methods before and after correction in SpATS. The values were visualized in box plots for comparison and pairwise t-tests were applied to examine whether the different methods produced significantly different σ_{gen_r} values.

Heritability served as a measure to determine how well the methods are suitable to detect genotype-specific differences in CT. It is a measure that quantifies how much of the total phenotypic variance (*i.e.* the variance of the observed values, *e.g.* CT) is explained by the genotypes (Oakey et al., 2006; Rodríguez-Álvarez et al., 2018). Standard heritability is the fraction of genotypic variance σ_g^2 and the sum of genotypic variance and error variance σ_e^2 divided by the number of replications r:

$$H_s^2 = \frac{\sigma_g^2}{(\sigma_g^2 + \frac{\sigma_e^2}{r})}. \tag{2.10}$$

The possible value of heritability ranges from 0 to 1. A high heritability means that a trait can be selected for, as the variance between genotypes is considerably larger than within genotypes. A heritability of 0 indicates that the variance is not related to the genotype at all, and therefore a trait with 0 heritability is not interesting in breeding or variety testing.

The heritability provided in SpATS is an extension of Eq. 2.10 which can be used for more complex variance structures, *e.g.*, unbalanced designs, the so-called generalized heritability (Oakey et al., 2006). While Eq. 2.10 showcases the principles of heritability, for the interested reader, the formula framework of generalized heritability is provided in Eq. S1.1 & Eq. S1.2. For more details on generalized heritability, see Oakey et al. (2006) and Rodríguez-Álvarez et al. (2018).

Except for the orthomosaic-based data, weights were included in the fitting process in SpATS where the weights w were equal to the inverted plot-wise standard error (se) estimates ($w = se^{-1}$) of the respective plot-wise CT estimation (Roth, Rodríguez-Álvarez, et al., 2021).

2.2.9 One-stage approach

All methods described so far were two-stage approaches where plot-wise CT values were estimated first with a subsequent spatial correction in SpATS. To offer a pragmatic solution, an additional one-stage approach was tested where the multi-view raw data was directly fitted in SpATS. To that end, the term ν_j was added to Eq. 2.7 for the effect of the j^{th} trigger event (Eq. 2.11).

$$\begin{aligned}
\hat{\theta}_{ijknp} = \theta_i + \tau_k + \left(\theta_n \tau_n\right)_{ik} + \phi_p + r_n + & \quad \text{(base model)} \\
\tau r_{kn} + & \quad \text{(repl. } \times \text{ treat.)} \\
f\big(c(p), r(p)\big) + \psi_{c(p)} + \psi_{r(p)} + & \quad \text{(spatial model)} \\
\nu_j + & \quad \text{(trigger timing)} \\
e_{ijknp} & \quad \text{(error term)}
\end{aligned} \tag{2.11}$$

As with the other approaches, plot-wise CT values were then predicted with Eq. 2.8 for comparison.

2.2.10 Data quality improvements by data selection

Using a multi-view approach allows to select measurements according to values of geometric covariates as well as to modify the number of the measurements included in the analysis.

When changing from a nadir oriented view to a more oblique view, the avoidance of the most nadir oriented measurements leads to a reduction of apparent soil cover and therefore soil signal in the more oblique measurements (Aasen and Bolten, 2018; Pask et al., 2012; Perich et al., 2020). Whether and how the selection of a specific viewing-geometry impacts the CT estimates was tested by excluding most nadir oriented data in a data-treatment experiment. Pearson correlations and heritability were used to estimate how the nadir exclusion influences the quality of the results with regard to consistency and genotype specificity. Most nadir oriented measurements were excluded for every flight in swaths in direction of sowing. Swath width of exclusion was 0 m (*i.e.* no exclusion), 2 m, 4 m and 6 m from the line parallel to sowing direction directly below the drone. This led to swath widths of 0 m, 4 m, 8 m and 12 m. The measurements for every flight were then fitted with the "MM Trigger" model (Table 2.3) and SpATS according to Eq. 2.7 in a two-stage approach. The fitted plot-wise values were correlated to all other flights of the same swath width of nadir exclusion and heritabilities were calculated for comparison.

Excluding measurements reduces the number of observations available for the analysis. To examine the effect of a reduction of the number of observation included in analysis, the number of observations for each plot in each flight was varied from 1 to 9 observations per single plot in a data-treatment experiment. The observations were chosen randomly and the procedure was repeated five times for each number of observation. Values were fitted with the pragmatic one-stage approach (Eq. 2.11) in SpATS as "MM Trigger" produces very erratic estimates of temporal trends when number of observations is low. The fitted plot-wise values were correlated to all other flights of the same number of observations and heritabilities calculated. Correlation values and heritabilities were grouped over all flights for each number of observations for comparison.

2.3 Results

2.3.1 TIR data processing and processing comparison

2.3.1.1 Example of selected correction steps

Fig. 2.4 provides an overview on how some of the methods and the spatial correction in SpATS affected the CT estimates. Three methods were chosen for a comparison. The "Ortho" method provides a base-line for comparison, "Agg.-Mean" is a multi-view approach without correction before SpATS, and "MM Trigger" is a multi-view approach using trigger timing as a covariate for corrections of thermal drift in a mixed model. As case example, relative CT values were visualized for the first flight of the campaign on 2022-05-18 at 16.00.

The field maps of $\hat{\theta}_{p\_Ortho}$ and $\hat{\theta}_{p\_mean}$ before spatial correction in SpATS contain strong trends. While these trends at first sight appear to be spatial, they are in reality composed of both spatial and temporal trends. The CT estimates span wide ranges within genotypes. After correcting for temporal and also most dominant spatial trends, CT estimates based on $\hat{\theta}_{p\_MM}$ do not show strong trends anymore and the within-genotype variance decreased. These "$\hat{\theta}_p$ before SpATS" values were the input values for the spatial correction in SpATS. The estimates after the spatial correction $\hat{\theta}_{p\_SpATS}$ are show on the right side of Fig. 2.4. No strong trends could be detected anymore for any of the three methods after spatial correction and within-genotype variance decreased for all three. The within-genotype variance of 'MM Trigger" is already lower before final spatial correction in SpATS than for the "Ortho" method after

spatial correction. The ranking of the genotypes is very similar between the three methods after SpATS, but not before, where $\hat{\theta}_{p\_\boldsymbol{Ortho}}$ and $\hat{\theta}_{p\_\boldsymbol{mean}}$ show similar general trends of ranking between the two methods but not compared to $\hat{\theta}_{p\_\boldsymbol{MM}}$.

2.3.1.2 Percentile choice for data aggregation and blending mode choice in orthomosaic composition

To find the best suited percentile for the aggregation, percentile-heritability relations of all flights were visualized for both, orthomosaic (Fig. S1.5b & Fig. S1.6b) and multi-view method (Fig. S1.7).

The median (*i.e.* the 50$^{\text{th}}$ percentile) fulfilled the two criteria of high heritability and stability of heritability over closely adjacent percentiles in both years. Differences in heritabilities of different percentiles between the orthomosaic (Fig. S1.5b & Fig. S1.6b) and multi-view (Fig. S1.7) methods were small. Hence, the 50$^{\text{th}}$ percentile was chosen for both methods for later method comparison. The orthomosaic blending mode "Mosaic" led to the highest and most stable heritabilities and was therefore chosen for further analysis.

2.3.1.3 Covariates related to trigger timing and viewing geometry

The "Base" model (design factors only) and the "Base + Full Spatial" model failed to fit in the mixed model stage for 12 out of 39 flights and 9 out of 39 flights, respectively. Just by including trigger timing in mixed models, models converged for all flights. When comparing BICs of models, the "Base" model (design factors only) always showed a higher BIC and therefore higher lack of fit than more complex models that include covariates (Fig. 2.5). Increasing complexity of the spatial model in the "Base + Full Spatial" model did not improve the models while adding trigger timing significantly improved the performance in all cases. The inclusion of "Sun-Direction-Trend" improved most models significantly. "Row-Direction-Trend" slightly improved the models while considering the position of the plot on the sensor plane (*i.e.* image coordinates of the plot center, denoted "Sensor-Plane-Trend") did not lead to any improvement.

2.3.1.4 Example of thermal drift

A strong drift of TIR measurements along trigger timing, *i.e.* a strong temporal trend was observed for all measurements. Patterns were similar for all flights (*e.g.* Fig. 2.6). Analyzing the estimated temperature drift with time ($f_{\text{spl}}(j)$ in Eq. 2.5) with the "MM Trigger" mixed model in relation to relative movements along the main flight direction (Fig. 2.6) revealed a strong link between main direction of flight and direction of TIR drift. A change of temporal trend coincided very often with a change of motion direction. Temperature frequently changed more than 10 °C within one flight line. The direction of this relation was not persistent and the temporal trend sometimes increased or decreased for the same direction of motion within a flight campaign or even within a single flight.

2.3.1.5 Consistency of plot-wise CT estimates and genotype CT ranking

As a metric of consistency, correlations of plot-wise values $\hat{\theta}_{p\_\boldsymbol{SpATS}}$ between flights within years were calculated, as well as the sd of genotype rankings within campaigns.

Plot-wise CT estimation with the best performing yet most complex mixed model "MM Trigger + RowDir + SunDir + Sensor" was applied to all plots within the field with subsequent spatial correction in SpATS. The correlations between $\hat{\theta}_{p\_\boldsymbol{SpATS}}$ of different flights ranged from moderate to very strong, with generally stronger correlations for flights that were taken within a shorter period (closer to the diagonal of the correlation table) and weaker for flights that

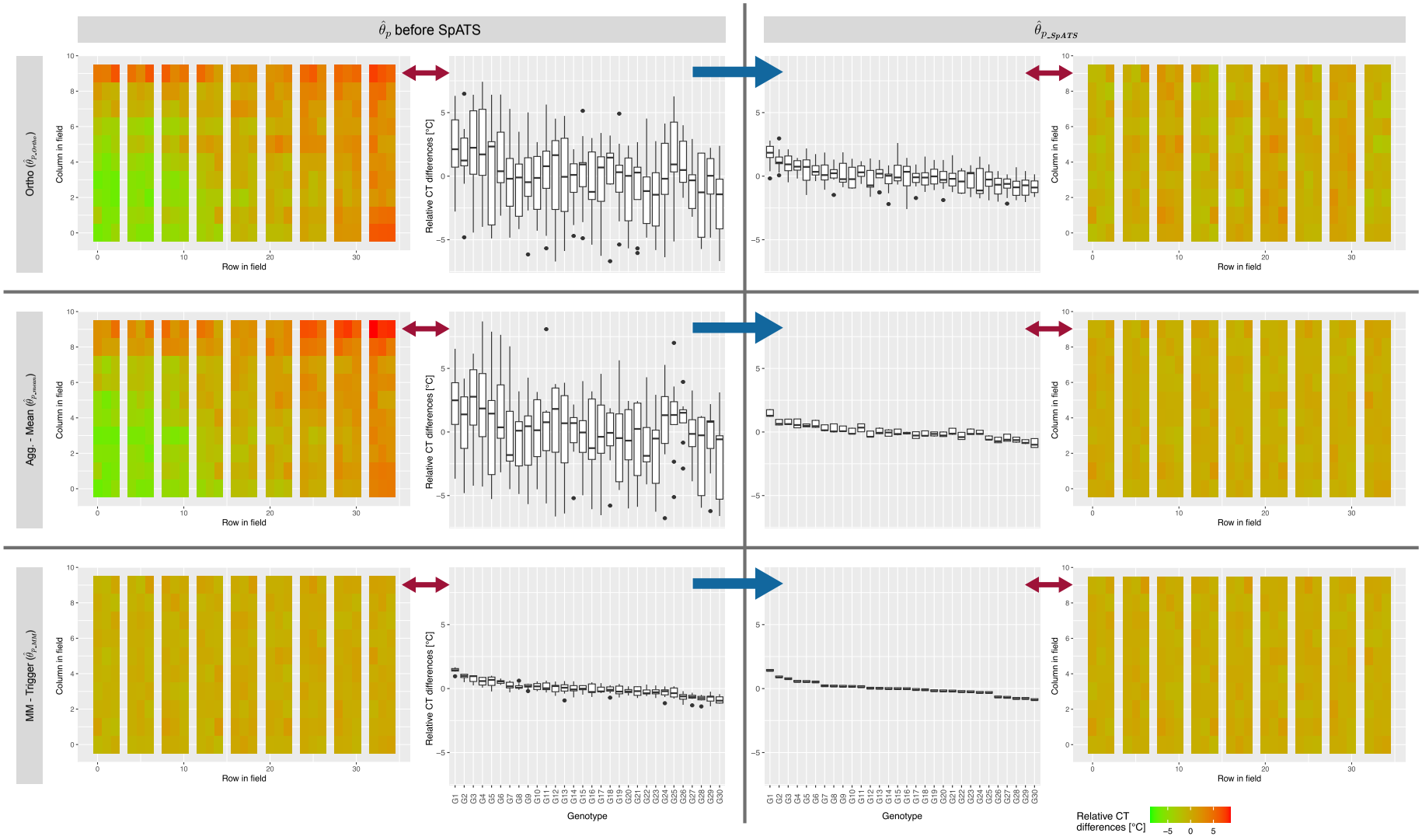


FIGURE 2.4: Data visualization of CT differences from mean CT within the first flight of the campaign on 2022-05-18 at 16.00. CT was estimated with the orthomosaic method ($\hat{\theta}_{p\_\boldsymbol{Ortho}}$), the "Agg.-Mean" aggregation method ($\hat{\theta}_{p\_\boldsymbol{mean}}$) and the "MM Trigger" mixed model ($\hat{\theta}_{p\_\boldsymbol{MM}}$). Values are shown before spatial correction ($\hat{\theta}_p$ before SpATS) and after correction in SpATS ($\hat{\theta}_{p\_\boldsymbol{SpATS}}$). The field maps show the measured CT differences of the plots at their specific location in the field. The box plots show the values ordered by the 30 genotypes (G1...G30). Genotypes were arranged in decreasing order of values according to data of the "MM Trigger" method after spatial correction in SpATS. The treatments had very little effect on CT values, and the values of all three treatments were summarized in the same box plot. Therefore, each box plot is based on 9 data points.

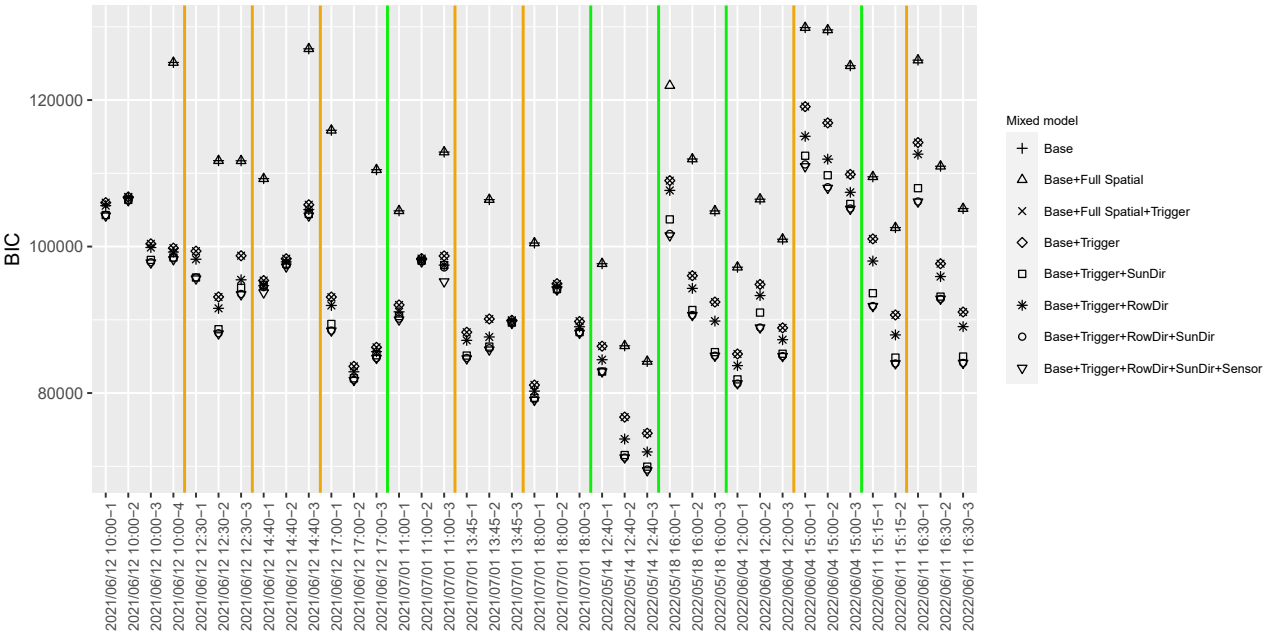


FIGURE 2.5: The Bayesian information criterion (BIC) for all flights. With BIC, the quality of the model fit was compared. Lower BIC values indicate preferable models. Green lines separate different measurement days, orange lines different flight campaigns within the same day.

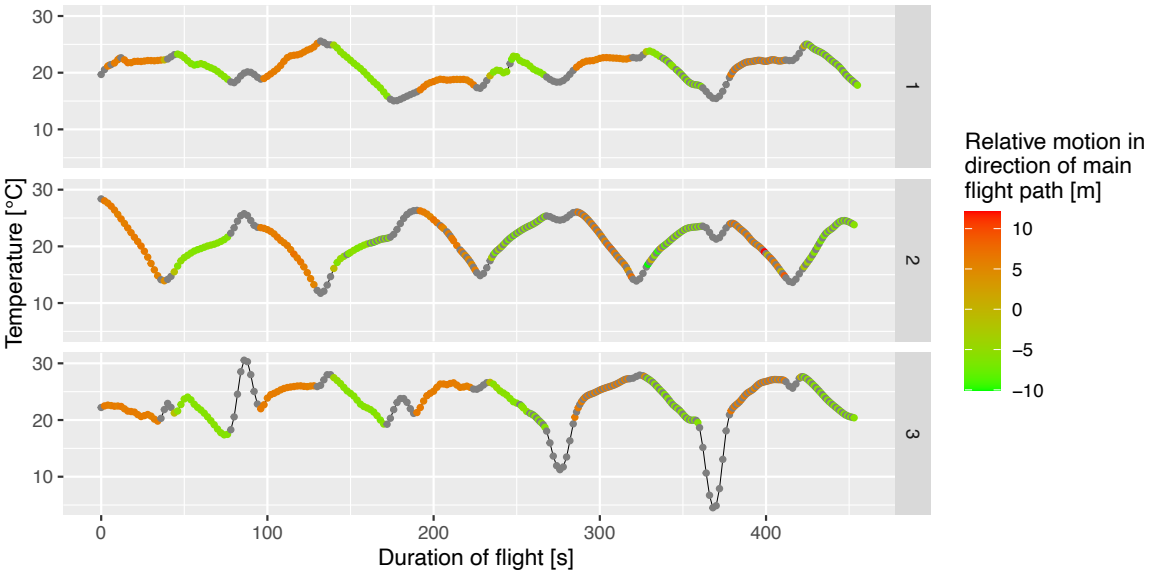


FIGURE 2.6: Estimated thermal drift of TIR measurements throughout the duration of flights for the three flights of the 13.45 campaign on 2021-07-01 on EuVar21. Rows 1 to 3 represent the three different flights of the same campaign. Flight plan and sensor orientation were identical for the three flights which were all conducted within 30 min. The colors indicate the motion in direction of the main flight path. Red indicates flights in one direction and green in the opposite direction of the flight path grid. For gray points, temporal drift was modeled but there was no corresponding measurement of motion along the main flight path.

were taken at times further apart. These patterns were consistent over both years (Figs. 2.7 & S1.8).

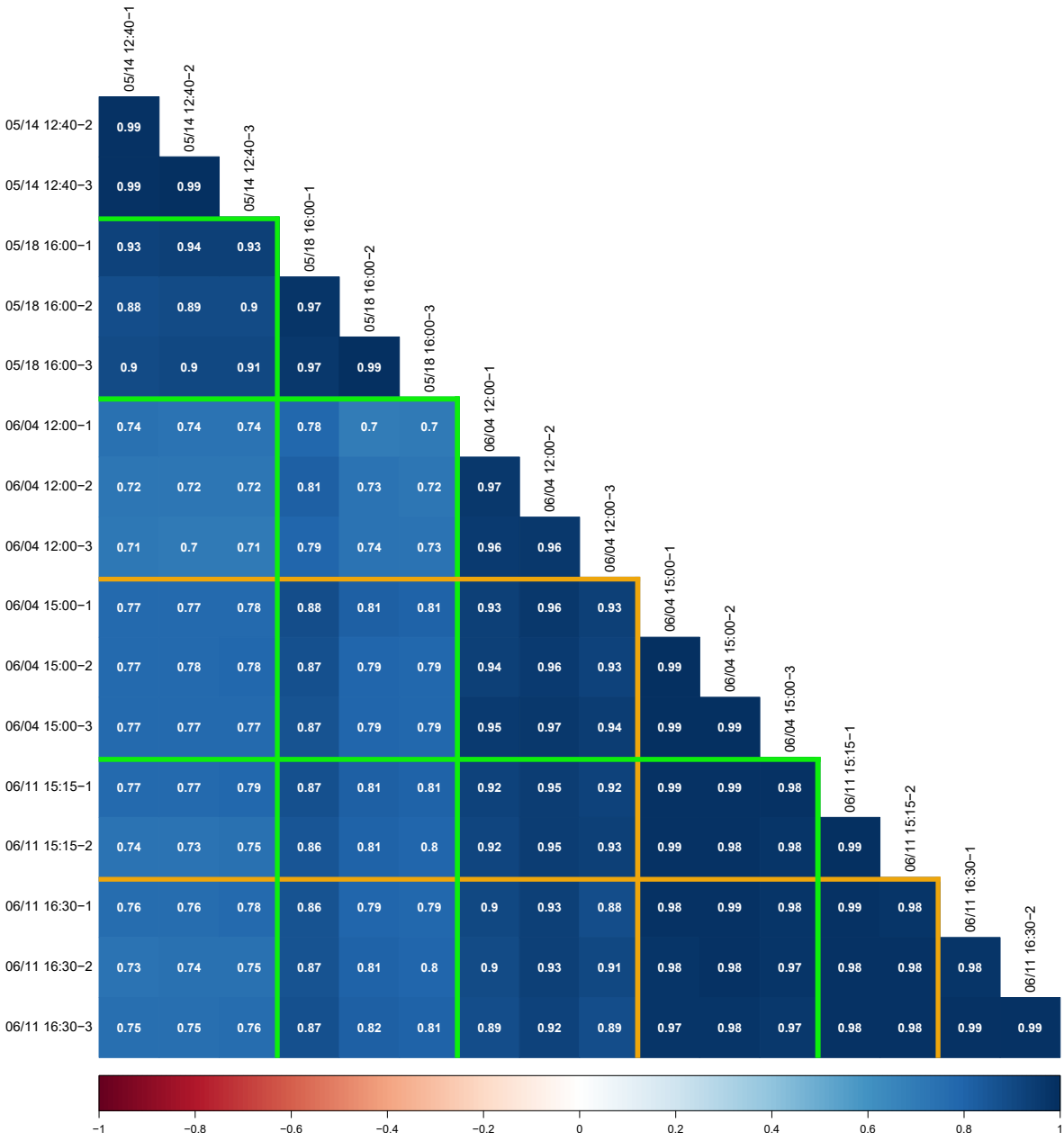


FIGURE 2.7: Pearson's correlations of EuVar22 plot values $\hat{\theta}_{p\_SpATS}$ after correction for spatial and temporal covariates ("MM Trigger + RowDir + SunDir + Sensor" and subsequent fitting with SpATS) and removing dominant treatment effects. Green lines separate different measurement days, orange lines different flight campaigns within the same day. All correlations are significant at $P < 0.001$.

Mean plot-wise correlations of CT measurements over all dates were calculated for different CT pre-processing and post-processing methods (similar to Fig. 2.7) and correlations aggregated in box plots for comparison (Fig. 2.8a). As flights were conducted in a specific phenological window (onset of heading - early senescence), correlations were strong not just within campaigns but also between campaigns. Consequently, all correlations of one season were summarized in the same box plot. Correlations were weakest for the orthomosaic method. Mean correlations of the orthomosaics method were 0.62 and 0.78 in 2021 and 2022, respectively. Correlations were stronger for the median multi-view aggregation method (0.63/0.81 for 2021/2022, respectively), the "LM" (0.68/0.80) and the mean multi-view aggregation (0.71/0.85). Correlations were strongest for the one-stage SpATS model (0.76/0.86), the mixed models "MM Trigger" (0.76/0.87) and "MM Trigger + RowDir + SunDir + Sensor" (0.76/0.88) . Correlations of plot-wise CT measurements were similarly strong for both years within campaigns (Figs. 2.7, S1.8). Vignetting correction almost did not change the values, and the values mentioned are those without vignetting correction.

The sd of genotype ranking within campaigns σ_{gen_r} (Eq. 2.9) was calculated for all processing methods (Fig. 2.8b). The values of the method "SpATS (one-stage)" before spatial correction correspond to unadjusted mean values as for the method "Agg. - Mean". σ_{gen_r} was lowest after mixed model pre-processing and spatial correction in SpATS in both years but was similarly low for the "SpATS (one-stage)" approach. Spatial correction had a large effect for models without mixed model pre-processing. σ_{gen_r} was very similar for the "Ortho" and "Agg.-Mean" method before and after SpATS in both years. The genotype ranking within campaigns was therefore most consistent for the approaches with mixed model pre-processing, but similarly consistent for the "SpATS (one-stage)" approach. Mean and median values of σ_{gen_r} for all methods are shown in Table S1.4.

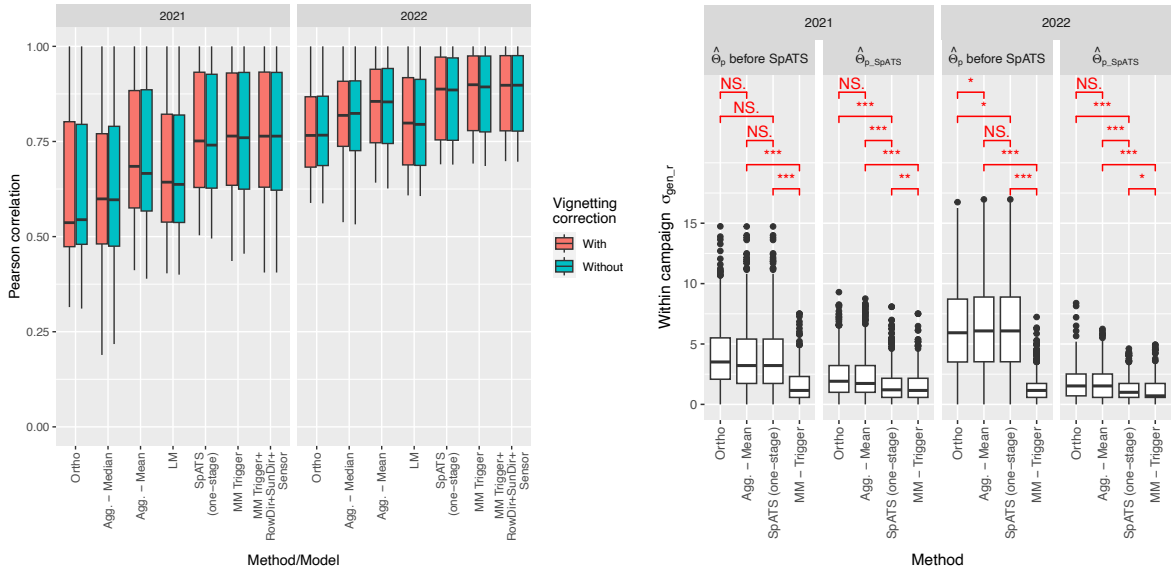
2.3.1.6 Genotypic specificity of apparent CT

Heritabilities were generally high to very high (Fig. 2.9). The aggregation methods "Mean" and "Median" provided the lowest heritability estimates with the highest variability between flights of the same campaign, followed by the "Ortho" method. The CT estimation methods "LM" and "SpATS (one-stage)" mostly showed slightly higher and less variable heritabilities than the "Ortho" method. Plot-wise CT estimation with the "MM Trigger" method and "MM Trigger + RowDir + SunDir + Sensor" consistently showed the highest and least variable heritabilities. Often the "Trigger" model showed slightly higher heritabilities than the more complex method. The difference between heritabilities of data without and with vignetting correction was minimal with the average absolute difference between the two being 0.005 over all methods tested. No clear trend could be observed for the sequence of the individual flights within a campaign.

2.3.2 Analysis on quantity and quality of observations included in multi-view models

2.3.2.1 Selection of non-nadir measurements

Excluding measurements that were closest to the line in nadir direction below the drone and parallel to row direction increased heritability consistently for both years (Fig. 2.10a). The correlation between the flights within one swath width of nadir-view exclusion got weaker in general with increasing swath width (Fig. 2.10b).



(a) (b)

FIGURE 2.8: Consistency of plot-wise CT estimates and genotype CT ranking. (a) Pearson's correlations of plot-wise CT measurements $\hat{\theta}_{p\_SpATS}$ within EuVar. Correlations were calculated for each flight within both years, but not across years. CT was estimated with the orthomosaic method, two different aggregation methods ("Agg.-Median" & "Agg.-Mean"), the "LM", one-stage SpATS and two mixed model methods ("MM Trigger", "MM Trigger + RowDir + SunDir + Sensor"). Correlations were calculated for data with and without vignetting correction after spatial correction in SpATS. (b) The sd of genotype ranking σ_{gen_r} (Eq. 2.9) within campaigns was arranged for four different processing methods and two years before and after spatial correction in SpATS. Each box plot is based on 90 σ_{gen_r} values from the 30 genotypes sown within three treatments each year for all campaigns within one year (7 campaigns in 2021 and 6 campaigns in 2022). Red marks indicate the significance of the differences between the groups based on a pairwise t-test. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

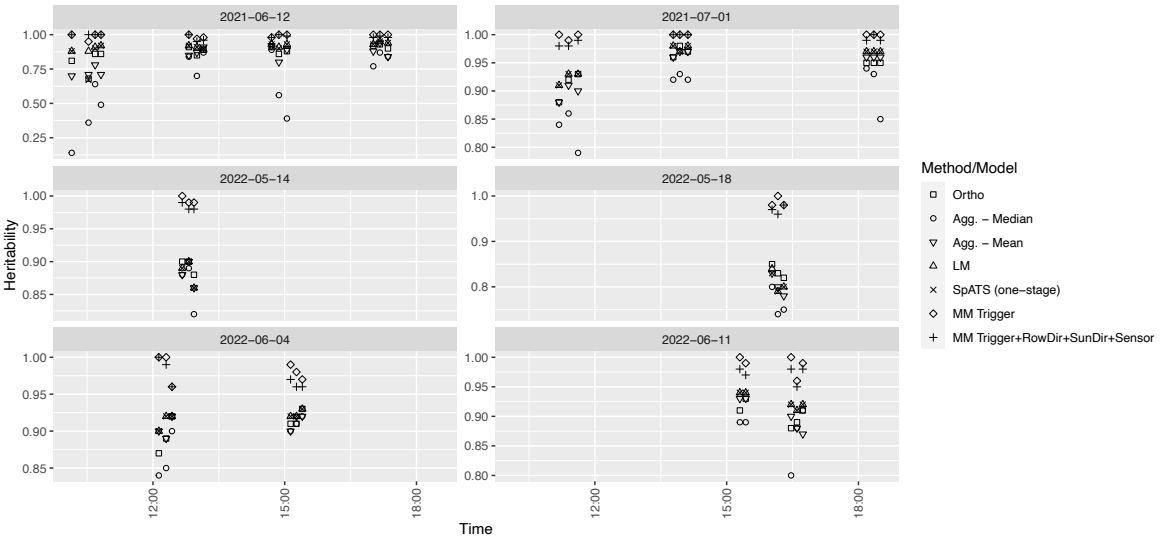
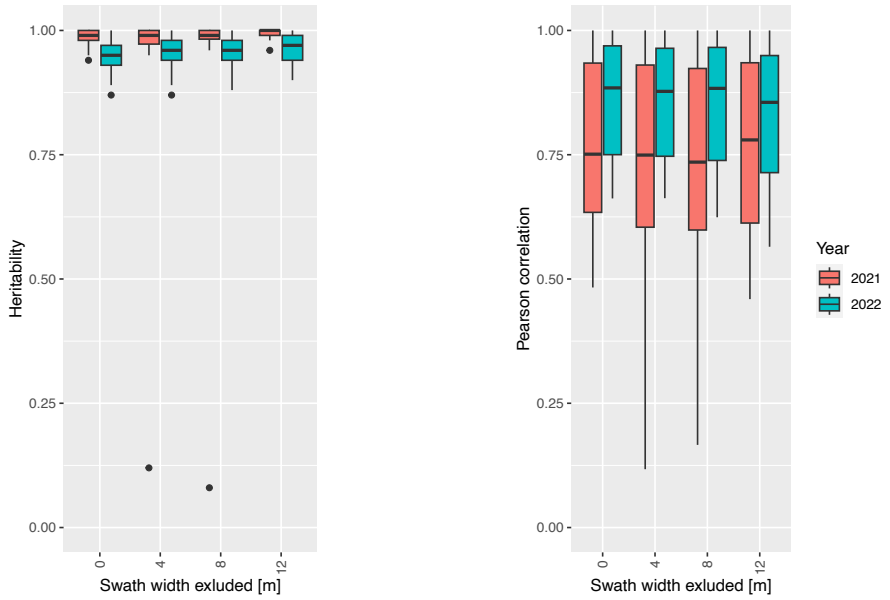


FIGURE 2.9: Heritability for all flights on EuVar grouped by date and time. The shapes indicate the different methods and models used in plot-wise CT estimation. Each group of two to four flights within the different time slots represents a campaign. Note that the scale of heritability is varying between the plots to allow to represent very different value ranges between different dates.



(a) (b)

FIGURE 2.10: Heritabilities (a) and the correlations (b) for all flights, for which data was excluded in nadir oriented swaths of different widths.

2.3.2.2 Number of observations included in models

Heritabilities were calculated for 1 to 9 randomly chosen observations for each plot. The procedure was repeated five times for each flight and values were fitted with the SpATS one-stage approach (Eq. 2.11). When comparing the resulting correlations and heritabilities in box plots, they consistently increased with increased number of observations for both years (Fig. 2.11a) but seem to asymptotically approach a maximum. Also the correlation between the flights increased with the number of observations, indicating that measurements became more consistent (Fig. 2.11b).

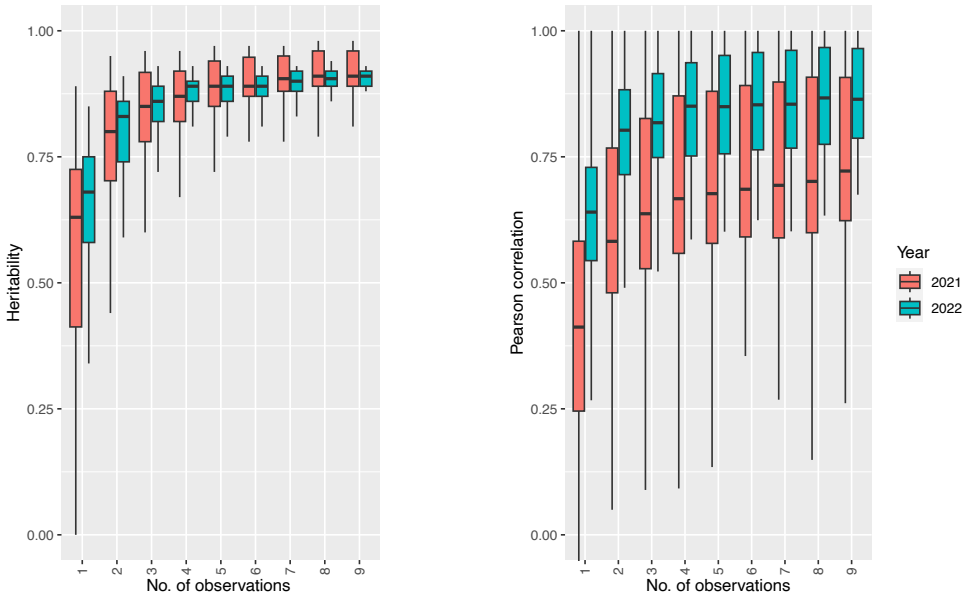
2.3.3 Weather conditions during flights

Flights were conducted in conditions suitable for flying (low wind, dry canopy, no rain). Within these conditions, no obvious dependence of heritability on environmental parameters such as temperature, solar radiation, wind speed, wind direction, relative humidity or VPD could be found (Fig. S1.9 and S1.10).

2.4 Discussion

2.4.1 The performance of multi-view methods

The results demonstrated the large influence of temporal, spatial, and geometrical trends on CT measurements (*e.g.*, Fig. 2.4), and how different methods lead to different CT estimates. After a final spatial correction with SpATS, strong trends had largely disappeared for all three methods, but within-genotype variance still differed significantly between the three methods. "Ortho" processing showed the largest within-genotype variance. A larger variance within genotypes reduced the heritability, as it decreased the ratio of genotypic variance, *i.e.* the variance caused by different genotypes, divided by the sum of genotypic and unexplained error variance (Eq. 2.10). From the CT values arranged by genotypes, it therefore became evident



(a) (b)

FIGURE 2.11: Heritabilities (a) and the correlations (b) for all flights and specific numbers of observations per flight.

that heritability increased for "Agg.-Median" method and was highest for "MM Trigger". The latter contained just a very low within-genotype variance after a final spatial correction. A decreased within-genotype variance also allows for a more consistent genotype ranking.

The multi-view method improved the genotype ranking consistency of CT within campaigns, and highly genotype specific CT measurements could be derived in the very contrasting conditions of the wet and cool year 2021 and the hot and dry year 2022. Using simple aggregation functions such as the mean and median to aggregate multiple values per plot showed generally lower heritabilities than using orthomosaics. The results indicate that a weighted spatial aggregation as done in the orthomosaic generation is superior to simple aggregation methods, but inferior to multi-view methods including mixed models or the "SpATS (one-stage)" method.

Both the orthomosaic method and mean and median aggregation do not compensate for temporal effects. Consequently, the subsequent processing of plot values (*e.g.*, in SpATS) is assumed to correct for both spatial and temporal trends simultaneously in such situations. Usually, drones fly perpendicularly or parallel to row directions in experiments. While the sequence of images is lost when aggregating using the mean or median, nadir oriented parts of images are getting the highest weight in the orthomosaic blending mode "Mosaic", which will partly preserve the triggering sequence. Consequently, a spatial correction of plot values can correct partially for spatial and temporal trends for blended orthomosaics, but not for aggregated values when using the mean or median.

Working with multi-view data allows to reduce temporal trends in plot-wise CT estimation. Including trigger timing in CT estimation was improving model fits and correlations the most but the fits could not be improved by a more complex spatial model. This shows that models are correcting for temporal effects and not for spatial effects that are mixed up with temporal effects. The separation of spatial and temporal trends is possible because even with a flight path that is parallel to row or column direction, each plot is recorded at multiple drone passes with opposing flight directions. The conditions on the sensor are not always the same when flying over the same plot. This becomes evident when examining the temporal pattern of the

thermal drift *e.g.* in Fig. 2.6. At about 135 s after flight start, temperature is estimated to be at a local maximum for the flights 1 and 3 and a local minimum for flight 2, but all three flights were conducted within 30 min. Such large differences can be explained by thermal drift (*e.g.* Kelly et al., 2019) but not with large CT changes in the field under relatively stable conditions. This separation of trends might be the main reason why all methods that included temporal trends showed strong correlations between plot-wise CT estimates. While the most complex plot-wise CT estimation with mixed models led to the highest correlations, the relatively simple CT estimation with the one-stage SpATS model led to good results as well while being far less complicated and computationally intensive than the mixed models computed with ASReml-R. The simple model, considering trigger timing and including a simple spatial model, might be sufficient for many cases.

Nevertheless, high heritabilities and correlations were achieved with all methods and even with the orthomosaic method, the estimated heritabilities were often higher than what was reported in comparable experiments (*e.g.* Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; Perich et al., 2020). The very high heritabilities in this study might in part be due to the properties of the experiments such as the chosen genotypes which originated from all over Europe. This led to a diverse set of genotypes which showed a more heterogeneous behavior than variety trials with genotypes adapted to conditions in Switzerland. In addition, the treatments had relatively little effect on the performance of the varieties which increased the number of effective replicates to nine and in turn led to a more robust estimation of genotypic variances.

2.4.2 Continuous thermal drift and influence of wind

Our data suggest that thermal drift is continuing throughout the flights, regardless of the previous stabilization regimen. This indicates that a thermal equilibrium in the sensor is not reached during the flights (*e.g.* Yuan and Hua, 2022).

In accordance with Kelly et al. (2019), we assume changing wind conditions on the sensor to be the main source of temperature drift. While flights were conducted in conditions with relatively low wind speeds, the wind conditions on the sensor kept changing constantly throughout the flights, in particular at the turning points of the flight path. Kelly et al. (2019) had shown that a wind speed difference of as low as $2\,\mathrm{m\,s^{-1}}$ is sufficient to trigger large thermal drift. A change in flight direction came with a change of wind direction and speed the sensor was exposed to and changes in thermal drift often coincided with changes of main flight direction. Although this is in line with the findings of previous studies (Kelly et al., 2019; Malbéteau et al., 2021; Yuan and Hua, 2022), we demonstrated the relation between flight path-related changes of wind conditions on the sensor and CT readings in-flight and continuously for the first time to the best of our knowledge.

Kelly et al. (2019) and Yuan and Hua (2022) reported the sensor to need several minutes to reach an equilibrium after changing wind conditions. This is much longer than the interval between changes in wind conditions caused by changes in the flight path, which is typically below 1 min, and the sensor does not have the time to reach an equilibrium during a flight. It has been suggested to mount shields to protect the sensor from exposure to wind (*e.g.* Kelly et al., 2019). Yet, this would also increase payload and reduce the agility of the gimbal. In addition, the potential of such a shielding to reduce sensor drift might be limited, as wind is only one of several potential drivers of sensor temperature.

Drift is most pronounced after turning on the TIR sensors and therefore, stabilization procedures are suggested in literature. In this work, temperature stabilization period was 15 min in 2021 and was increased to 30 min in 2022. This was longer than the 10 min recommended for handheld thermometers in Pask et al. (2012) and in the range of the 30 min recommended in

Jimenez-Berni, P. J. Zarco-Tejada, et al. (2009). Kelly et al. (2019) and Yuan and Hua (2022) showed that under laboratory conditions, the largest drifts of TIR cameras often occur during the first 30 min. In 2022, heritabilities were generally lower than in 2021, when stabilization period was shorter. We therefore conclude that other parameters are more relevant for the quality of drone-based TIR imaging than increasing the temperature stabilization period on the ground beyond 15 min.

Within campaigns, there was no clear trend that first flights showed a lower heritability than later flights of the same campaign. The suggestion of Kelly et al. (2019) to hover the drone for 15 min prior to measurements to stabilize it with in-flight conditions did not prove to be helpful for the multi-view approach in our study. Continuous thermal drift throughout the flight cannot be considered in any pre-flight stabilization procedure alone. Also, in-flight stabilization procedures, where the drone is hovered over the field prior to measurements, just help to mitigate effects from rather constant propeller slipstream but not from changing direction of flight and wind.

2.4.3 Analysis on quality and quantity of observations included in multi-view models

Multi-view allows to select data according to viewing geometry. When measuring CT of wheat crops with a handheld sensor, it is recommended to measure at an oblique angle to reduce the influence of the soil (Pask et al., 2012). By excluding most nadir-oriented measurements in aerial thermography, the average fraction of plant pixels per measurement can be increased. Measurement values are therefore more related to actual CT and less to canopy cover and soil temperature. This is most likely also leading to more accurate (though not necessarily higher) correlations between flights, as different traits such as stomatal conductance and canopy cover are unmixed to a certain degree. Heritability and correlation between flights within the same year also depend on the number of observations included.

The results showed that more observations per plot make the measurements more genotype specific and consistent. This effect is not unique to multi-view imaging: also in orthomosaic blending, information of multiple images is aggregated into one orthomosaic. Unlike with orthomosaics, with multi-view, we can determine the influence of the number of observations by excluding random observations. Consequently, the added value of repeated measurements per plot could be estimated, which would allow to estimate the minimum number of measurements to be included and to plan flights accordingly (flight height and overlap). A trade-off between maximizing number of observations and optimizing quality by data selection must be found for individual CT measurement campaigns. When nadir-oriented measurements are excluded, the number of measurements per plot might become so low that it deteriorates the CT estimates, which was demonstrated with weakening correlations when swath width of nadir-view exclusion was increased.

2.4.4 Sensitivity of the approach

The estimated plot-wise CT estimates within single flights span over a range of 2.96 °C on average over all flights. Within this range, the measurements were shown to be highly consistent within the same campaign over the 270 plots of EuVar by means of correlation and within-campaign genotype ranking consistency. In addition, significant differences could be found between the 30 genotypes. The high genotype specificity of the CT values was confirmed by the high heritabilities. This indicates a high sensitivity of our approach for relative CT differences. This sensitivity is clearly below a relative sensitivity of 1 °C which is stated in Mesas-Carrascosa et al. (2018) to be required for TIR measurements in agriculture. However,

the sensitivity is restricted to relative differences in CT. For absolute values, as required in many applications for crop physiology, additional in-field calibration would be needed.

Kelly et al. (2019) showed that also uncooled and uncalibrated TIR cameras show a relatively constant relation (*i.e.* slope) between DNs (the original raw values of the thermal camera) and temperature of reference objects. The authors found that mainly the offset is changing between flights. This supports our findings that the multi-view approach allows to represent relative temperature differences well, but the estimation of absolute values is prone to large errors. The narrow ranges of genotypic differences found indicate that the accuracy of uncooled yet calibrated TIR cameras of ±5 °C (Kelly et al., 2019; Perich et al., 2020) is not sufficient without a post-processing correction step.

2.4.5 Correlation and within-campaign genotype ranking consistency as measure of the methods consistency

In this work, correlations between CT of different flights and within-campaign genotype ranking were considered as indicators of consistency of the different approaches. Another option would be to correlate flight data with ground measurements. Nevertheless, ground reference measurements are subject to drift as well, and consequently should be taken in the same period as the TIR measurements. Ground reference measurement should also have the same response time and response pattern to changing environmental conditions as CT (H. G. Jones, Serraj, et al., 2009).

While measuring variability of CT *e.g.* between treatments and genotypes just once is a bad indicator of systematic and consistent CT differences (H. G. Jones, Serraj, et al., 2009), very strong and significant ($P < 0.001$) correlations between repeated measurements of apparent CT in independently processed flights were reached in this work. Together with the within-campaign genotype ranking, this demonstrates a high consistency and high reliability of the multi-view method.

2.4.6 Covariates in mixed models

The covariates included in the mixed models were chosen to represent the main trends assumed to be influencing the apparent temperature. Trigger timing was included to correct for trends related to sensor drift (Kelly et al., 2019). Lateral and longitudinal distance of the plot from the drone in sowing row direction were assumed to be related to changing apparent canopy cover (Aasen and Bolten, 2018; Pask et al., 2012; Perich et al., 2020). Anisotropy of wheat canopies, *i.e.* the directional dependence of the reflectance of TIR radiation on the crop surface, was assumed to be correlated with the lateral and longitudinal distance of the plot from the drone in sun direction (H. G. Jones, Serraj, et al., 2009; Nicodemus, 1977; Perich et al., 2020). The x and y image coordinates of the "Sensor-Plane-Trend" were intended to describe sensor related trends such as vignetting.

2.4.7 Including all plots to avoid border effects

Temporal drift was estimated based on all plot-wise measurements available, *i.e.* also border plots and plots of other experiments were included. When a drone is flying over an experiment in swaths, at the beginning and at the end of the swath, the temporal density of data points may decrease. For these regions, the estimation of the temporal drift is unbalanced and can take on extreme values (see extremely warm/cold gray points in Fig. 2.6). When including all plots in the plot-wise CT estimation models, the estimates of apparent CT within the experiment of interest are less impaired by the effect of reduced density, as the plots at the beginning and end of swaths are border plots or belong to other experiments that are not

in the focus of the study. In addition, the inclusion of other experiments and border plots increases the data available for more robust estimation of trends. Rodríguez-Álvarez et al., 2018 for example included 31 trials on one field for estimation of spatial trends before analying experiments separately.

2.4.8 Image pre-processing and TIR data extraction

Vignetting correction affected neither the correlations between measurements nor the heritabilities of the single flights significantly. Nevertheless, it was important to include it in the analysis as its spatial patterns potentially might mix up with the covariates related to viewing geometry. Decreasing the variance that might stem from vignetting previous to modeling decreased the risk of overestimation of geometry related effects in mixed models which might be concurrent with vignetting.

The choice of the 50^{th} percentile for plot-wise data aggregation allowed for highly consistent and heritable CT measurements. Together with nadir-view exclusion, a smart selection of a fitting percentile contributes to mitigating a bias by the background in mixed pixels. More reasoning on vignetting and zonal data aggregation by specific percentiles is provided in sections S1.17 and S1.18, respectively.

2.4.9 Benefits of additional data available in multi-view

For existing approaches of drone-based CT measurement, analysis is usually conducted on orthomosaics (*e.g.* Francesconi et al., 2021; Malbéteau et al., 2021; Perich et al., 2020). The presented image-wise multi-view approach allows for more detailed information on temporal trends, measurement geometries and uncertainty estimates. Such information is lost to a large extent when conducting analysis on orthomosaics. Mesas-Carrascosa et al. (2018) and Z. Wang et al. (2023) also used information of multiple images for an estimate of temporal drift. Mesas-Carrascosa et al. (2018) retrieved features from overlapping parts of images from multiple drone passes of the same flight while Z. Wang et al. (2023) just used features from consecutive images. They both used the differences between the features that appear on multiple images to correct the orthomosaic for temporal drift. In contrast, we extracted CT of the specific plots on single images directly. This automatic process enables an efficient information retrieval directly from overlapping images, which in turn increases the efficiency of trend estimation. In addition, multiple covariates can be calculated for each measurement, increasing the available information for a subsequent analysis. Aasen and Bolten (2018) estimated the position of pixels relative to the sun on single images by using a fixed orientation of the camera during the flight for hyperspectral information. The multi-view approach allows to calculate such geometric relations independently of the orientation of the camera. The interplay of wind conditions and flight direction on CT estimates was examined in Malbéteau et al. (2021). By visualizing temperature drift in relation to flight direction with a high temporal resolution, their findings could be complemented with continuous in-flight drift dynamic estimates. Deery, Rebetzke, Jimenez-Berni, James, et al. (2016), Deery, Rebetzke, Jimenez-Berni, Bovill, et al. (2019) and Perich et al. (2020) used correlations between measurements at different times and heritabilities as quality criteria of the experiment, H. G. Jones, Serraj, et al. (2009) used consistency of genotype ranking, while Malbéteau et al. (2021) used pixel-based standard deviation of the input-data to check quality. Based on previous studies, here correlations, genotype rankings, and heritabilities were also used as quality criteria, but the inverted standard error of the measurements per plot was included for weighting in heritability calculations as an uncertainty estimate. While the swath based approach of Malbéteau et al. (2021) corrects the input-data before analysis, with the multi-view approach, the different trends and effects are estimated in a statistical model. Estimated trends and standard errors are available for an in-depth

analysis together with multiple covariates, but the input-data remains unchanged, providing a comprehensive and detailed overview on the quality of the data. While such comparisons over different experiments have to be done with due caution, correlations and heritabilities in this study were as high or higher than what was reached in Deery, Rebetzke, Jimenez-Berni, Bovill, et al. (2019) and Perich et al. (2020) with calibrated TIR cameras. With an uncooled and uncalibrated TIR camera, correlations and heritabilities higher than 0.95 were reached by exploiting covariates available through the multi-view approach which allowed to correct for thermal drift and viewing geometry related effects. Multi-view as a lean phenotyping approach has therefore the potential to significantly improve CT measurements in the context of variety evaluation without the need for more expensive equipment or elaborate in-field reference procedures.

2.4.10 Cheat sheet for drone based multi-view thermography

Finally, based on the findings of this study and complemented from literature (Kelly et al., 2019; Yuan and Hua, 2022), the most important findings on an optimal procedure to measure CT in a multi-view approach are summarized in Fig. 2.12. The cheat sheet follows the logic of the work flow and is divided into the stages before the flight, flight, data extraction and analysis. If these recommendations are followed, the most important findings of this study can be incorporated into drone-based CT measurements.

2.4.11 Outlook

Further research might include streamlining the processing for simple implementation and using the method in combination with a stationary local sensor with a high absolute measurement accuracy for in-field normalization to derive accurate absolute CT values on the whole fields. Temporal drift information might be included in an orthomosaic blending procedure where each image gets an offset estimate by multi-view, allowing for more accurate and consistent orthomosaics. Alternatively, several geometric covariates could be calculated and their contribution to total variance examined in a multi-view approach. If information on wind direction and speed on the field are available at a high temporal and spatial resolution, CT and thermal drift could be related to the influence of changing wind conditions and gusts. Finally, while the method was developed for wheat phenotyping in variety testing experiments, it might be suitable for other field crops and even for observations beyond agriculture. Once the thermal images are aligned and georeferenced, the method is semiautomatic. As simple requirement, georeferenced polygons of the targeted ROIs must be available, and those ROIs must be small enough to appear entirely in multiple images of a flight. The back-projection of ROIs to images does not need any manual intervention, wherefore the whole process could be automatized. With the data retrieved, mixed models and linear models could be fitted for non-designed experiments (*e.g.*, land surface monitoring) as well as designed experiments (*e.g.*, breeding experiments) alike. Larger areas could be covered by flying at higher altitudes. Those adaptations would pave the way to apply the presented method not just for breeding and variety testing, but also, *e.g.*, to detect stressed patches in fields to improve irrigation efficiency, or variable-rate fertilization applications in precision agriculture (Romano et al., 2011; Messina and Modica, 2020; Chandel et al., 2022).

2.5 Conclusion

In this study, a multi-view approach for consistently measuring relative CT of wheat with a drone-based uncooled and uncalibrated TIR camera without any in-situ field references was

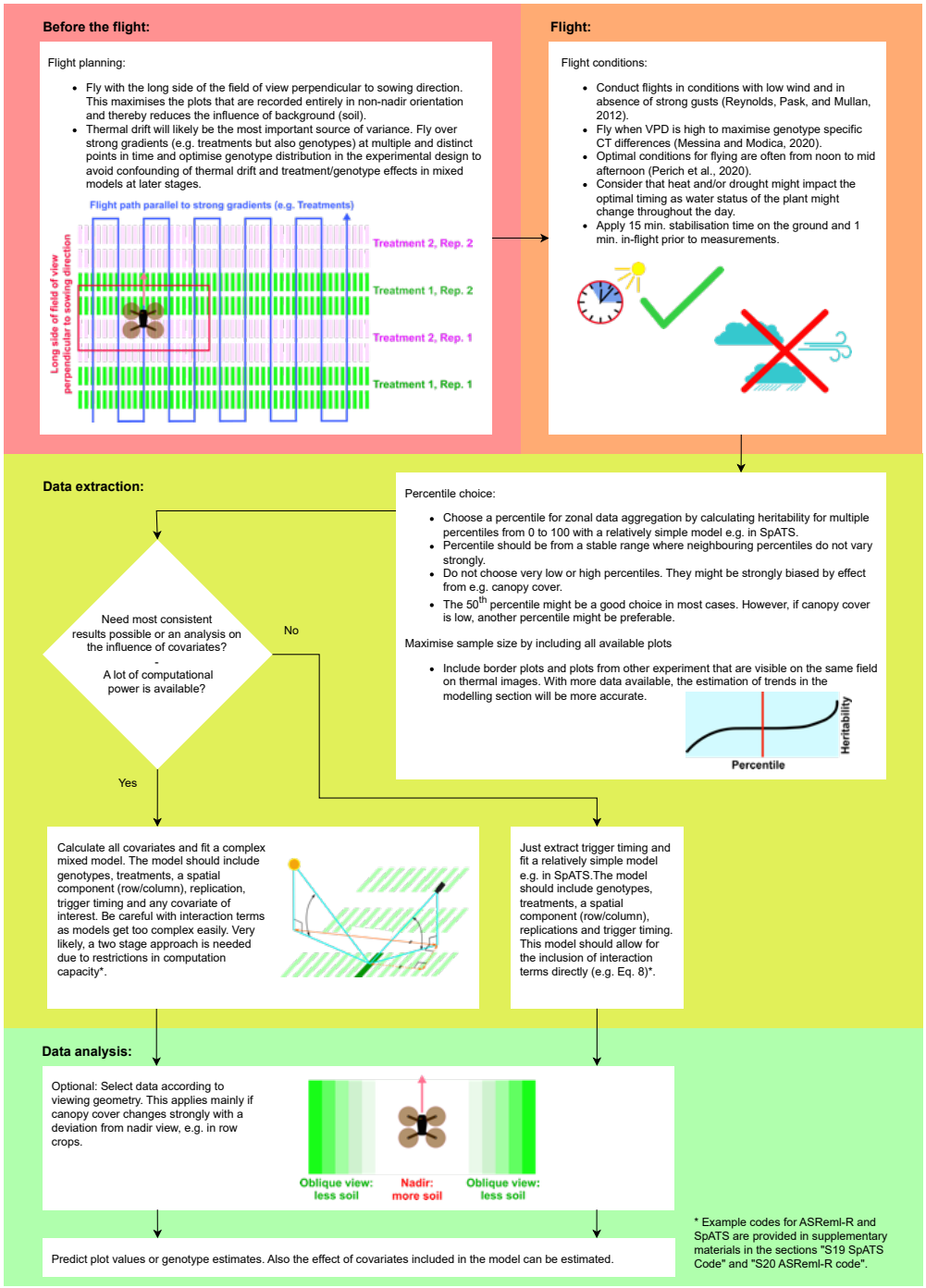


FIGURE 2.12: Cheat sheet, giving an overview on most relevant considerations when measuring CT with a drone based multi-view approach.

presented. The quality of the measurements was assessed by means of correlations between measurements taken at different times, genotype ranking consistency between flights and heritability. Contrary to standard orthomosaic approaches, multi-view allows to calculate and include several covariates in the analysis which improved the CT estimates in terms of correlation, ranking consistency and heritability. The trigger timing, describing thermal drift during a flight, was by far the most beneficial covariate to be included. Integrating other covariates related to viewing geometry with respect to position of the plot and the sun relative to the drone showed additional potential to improve CT estimates. The proposed approach enables the disentanglement of spatial drift from temporal drift.

The ability for detrending CT data, together with the option to select measurements according to viewing geometry paves the way for using drone-based thermography with relatively simple equipment as a lean-phenotyping method without complex calibration procedures. Yet, the method is limited to relative temperature differences and does not correct for errors in absolute CT values.

To facilitate the implementation of multi-view thermography, a computationally inexpensive and easy to apply model is provided based on the R-package SpATS. A cheat sheet outlines the complete procedure to facilitate its implementation.

In future research, the method might be used in combination with a ground-based sensor with a high absolute measurement accuracy for in-field normalization to derive accurate absolute CT values. In addition, in situations where an orthomosaic is required, temporal drift information might be used as image-specific offset information in orthomosaic processing to create more consistent orthomosaics.

Data availability

A working example of the procedure suggested in this publication including source code and example data is provided on github (`https://github.com/TreAgron/ThermalMultiviewExample.git`).

Authors' contribution

Simon Treier: conceptualization, planning and execution of the experiment, data collection, methodology, software, formal analysis, visualization, writing – original draft. Lukas Roth: conceptualization, supervision, methodology, review & editing. Juan M. Herrera: project administration, funding acquisition, conceptualization, planning of the experiment, supervision, methodology, acquisition, writing – review & editing. Norbert Kirchgessner: conceptualization, review & editing. Achim Walter, Helge Aasen, Andreas Hund: writing – review & editing.

Funding

This study was in part supported by the two H2020 projects InnoVar and Invite.

Acknowledgments

We thank Johanna Antretter, Fernanda Arelmann Steinbrecher, Ulysse Schaller, Matthias Schmid and Julien Vaudroz for rating of phenology; Nicolas Vuille-dit-Bille for the support in collecting and processing drone data; Nicolas Widmer and his team as well as Yann Imhoff for field management; Margot Visse-Mansiaux for support in setting up the experiments.

3 Analysis of variance and its sources in UAV-based multi-view thermal imaging of wheat plots

Simon Treier[1,2], Lukas Roth[2], Andreas Hund[2], Helge Aasen[3], Lilia Levy Häner[1], Nicolas Vuille-dit-Bille[1], Achim Walter[2], Juan M. Herrera[1]

1 – Production Technology & Cropping Systems Group, Agroscope, Route de Duiller 60, 1260 Nyon, Switzerland

2 – ETH Zürich, Institute of Agricultural Sciences, Universitätstrasse 2, 8092 Zürich, Switzerland

3 – Earth Observation of Agroecosystems Team, Agroecology and Environment Division, Agroscope, Reckenholzstrasse 191, 8046 Zürich, Switzerland

This chapter was resubmitted to Plant Phenomics after revision.

Abstract

Canopy temperature (CT) estimates from drone-based uncooled thermal cameras are prone to confounding effects, which affects the interpretability of CT estimates. Experimental sources of variance, such as genotypes and experimental treatments blend with confounding sources of variance such as thermal drift, spatial field trends, and effects related to viewing geometry. Nevertheless, CT is gaining popularity to characterize crop performance and crop water use, and as a proxy measurement of stomatal conductance and transpiration. Drone-based thermography was therefore proposed to measure CT in agricultural experiments. For a meaningful interpretation of CT, confounding sources of variance must be considered. In this study, the multi-view approach was applied to examine the variance components of CT on 99 flights with a drone-based thermal camera. Flights were conducted on two variety testing field trials of winter wheat over two years with contrasting meteorological conditions in the temperate climate of Switzerland. It was demonstrated how experimental sources of variance can be disentangled from confounding sources of variance and on average more than 96.5 % of the initial variance could be explained with experimental and confounding sources combined. Not considering confounding sources led to erroneous conclusions about phenotypic correlations of CT with traits such as yield, plant height, fractional canopy cover, and multispectral indices. Based on extensive and diverse data, this study provides comprehensive insights into the manifold sources of variance in CT measurements, which supports the planning and interpretation of drone-based CT screenings in variety testing, breeding, and research.

3.1 Introduction

Canopy temperature (CT) of wheat (*Triticum aestivum* L.) is a proxy-measurement of stomatal conductance (*e.g.* Anderegg, Aasen, et al., 2021; Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Rebetzke et al., 2013; M. P. Reynolds, Pask, et al., 2012) and transpiration (Jiang and Islam, 1999) that is negatively correlated with yield in well-watered conditions (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; R. A. Fischer et al., 1998; Pask et al., 2012; Rebetzke et al., 2013; Roche, 2015), *i.e.* a lower CT is generally associated with higher yield. CT is more sensitive to changes in the water status of plants than other optical measurements such as the Normalized Difference Vegetation Index (NDVI), and shows a faster response time to physiological changes in the plant (Baluja et al., 2012; Damm et al., 2022; Messina and Modica, 2020; P. Zarco-Tejada, González-Dugo, L. Williams, et al., 2013). This makes CT especially interesting for measuring plant performance in dry and/or hot conditions. Therefore, it was proposed to be used in cereal breeding (*e.g.* Anderegg, Aasen, et al., 2021; Brennan et al., 2007; Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Perich et al., 2020; Rebetzke et al., 2013; M. P. Reynolds, Pask, et al., 2012; Romano et al., 2011), in research and precision agriculture (*e.g.* Chandel et al., 2022; Maes and Steppe, 2012; P. Zarco-Tejada, González-Dugo, and Berni, 2012), *e.g.*, to detect water stress. Thermal infrared (TIR) cameras mounted on drones allow the efficient measurement of many experimental units (Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Perich et al., 2020). However, various sources of variance can adversely affect TIR measurements and increase uncertainties when estimating CT. Spatiotemporal and geometric patterns superimpose with the effects of specific genotypes or treatments (*e.g.* Kelly et al., 2019). Therefore, the measurement and interpretation of CT data is not trivial (Perich et al., 2020). Elaborated measurement procedures and statistical methods are needed to disentangle the sources of variance that influence CT measurements.

The most important sources of variance and their main drivers/causes are summarized in Table 3.1. First, CT is sensitive to short-term changes in environmental conditions. Solar radiation, air temperature, relative humidity of the air, vapor pressure deficit (VPD), and cloud cover are all interlinked and affect CT measurements directly by changing the heat balance of the canopy, for example, by fluctuating radiation or indirectly by impacting stomatal conductance (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; Pask et al., 2012; Perich et al., 2020; Rebetzke et al., 2013). Such environmental effects might mask more subtle plant responses (Damm et al., 2022). To reduce distortions by the environment, it is recommended to fly in stable conditions, *i.e.* when there are no clouds or haze and wind speeds are low with no gusts. However, also under stable conditions, solar radiation and VPD are constantly changing, and the conditions may differ at the beginning and the end of the flight, particularly for long-duration flights (Z. Wang et al., 2023).

Due to a limited payload of drones, uncooled thermal cameras are commonly used in field phenotyping. They often depend on Vanadium Oxide (VOx) microbolometers, which are arranged in focal plane arrays (FPA) (*e.g.* Das, J. Christopher, Apan, Roy Choudhury, et al., 2021; Kelly et al., 2019; Messina and Modica, 2020; Perich et al., 2020; Treier et al., 2024; Yuan and Hua, 2022). Such cameras are prone to thermal drift, where the measured temperature varies as a result of short-term temperature fluctuations the FPA of the sensor and the camera optics are exposed to (Messina and Modica, 2020; Nugent et al., 2013). This holds true for both radiometrically calibrated and uncalibrated cameras. Thermal drift is known to be a significant confounding source of variance in CT measurements, and the literature proposes different approaches to correct for it in data pre-processing (*e.g.* Kelly et al., 2019; Mesas-Carrascosa et al., 2018; Z. Wang et al., 2023; Yuan and Hua, 2022) and analysis (Treier et al., 2024). Kelly et al. (2019) and Yuan and Hua (2022) examined the importance of wind

TABLE 3.1: Overview on most important sources of variance of drone-based thermal measurements.

Variance source	**Variance driver/cause**	**Temporal behaviour**	**Primary type of correction**	**Reference**
Solar radiation	Weather	Dynamic (short term)	Temporal	M. P. Reynolds, Pask, et al. (2012)
VPD*				Idso et al. (1981)
Wind				M. P. Reynolds, Pask, et al. (2012)
Thermal drift	Sensor temperature	Dynamic (short term)	Temporal	Nugent et al. (2013)
Non-uniformity effects			-	Nugent et al. (2013)
Field heterogeneity	soil water content, water logging, soil compaction etc.	Stable throughout single flights	Spatial	Perich et al. (2020)
Treatment effects	Field management	Stable throughout single flights	Treatment	M. P. Reynolds, Pask, et al. (2012)
Plant height	Genotype/Field management	Stable throughout single flights	Genotype/ Treatment	Prashar and H. Jones (2014)
Soil cover				Aasen, Honkavaara, et al. (2018)
Stomatal conductance				M. P. Reynolds, Pask, et al. (2012)
Phenology				Prashar and H. Jones (2014)
Stay green				Anderegg, Aasen, et al. (2021)
Rooting depth (water availability)				M. P. Reynolds, Pask, et al. (2012)
Vignetting	Sensor/Optics	Rather stable	Geometric	Aasen, Honkavaara, et al. (2018)
BRDF**	Viewing geometry	Stable throughout single flights	Geometric	Schaepman-Strub et al. (2006)
Apparent soil cover				Pask et al. (2012)
Atmospheric effects				Jimenez-Berni, P. J. Zarco-Tejada, et al. (2009)

*Vapour pressure deficit (VPD), **Bidirectional reflectance distribution function (BRDF)

conditions on the sensor as an important driver of sensor temperature and TIR readings. Kelly et al. (2019), Malbéteau et al. (2021) and Treier et al. (2024) demonstrated how TIR readings change with relative motion along the main flight direction of the drone as a result of changing wind conditions the sensor is exposed to.

Thermal drift is not homogeneous throughout the FPA and leads to non-uniformity effects (*e.g.* Nugent et al., 2013). Other non-uniformity effects are caused by dark signal noise and vignetting (Aasen, Honkavaara, et al., 2018). The latter describes the alteration of the signal in dependence of the path of radiation through the lens optics, leading to distortions where the edges of the image appear darker (or cooler for thermography) than the central regions (Aasen, Honkavaara, et al., 2018; Kelly et al., 2019; Yuan and Hua, 2022).

The viewing geometry also alters the TIR readings. The signal is subject to surface anisotropy, that is, the signal is altered depending on the direction from which it is emitted/reflected from the surface (Aasen and Bolten, 2018; Aasen, Honkavaara, et al., 2018; Perich et al., 2020), which can be described with a bidirectional reflectance distribution function (BRDF) (Nicodemus, 1977; Schaepman-Strub et al., 2006). Additionally, viewing geometry alters the fraction of soil visible between rows in row crops. At a more nadir-oriented view, the fractional canopy cover (FCC) is at a minimum and increases with more oblique viewing geometry, mainly perpendicular to the sowing rows (Roth, Aasen, et al., 2018). It is therefore recommended to measure at oblique angles (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; Pask et al., 2012; Rebetzke et al., 2013). However, with drone-based cameras, this is not always possible, and excluding nadir-oriented measurements comes with trade-offs. Just including measurements from oblique angles is more canopy specific and less related to FCC, but it also decreases the maximum number of measurements that can be taken per plot

when less oblique measurements are excluded, which is deteriorating the consistency of the measurements (Treier et al., 2024).

While the sources of variance of the TIR measurements considered so far included instantaneous environmental conditions, the sensor, and the viewing geometry, the experiment at observation itself constitutes an important source of variance. In the case of wheat variety testing trials, different genotypes are arranged in the field in blocks of multiple randomly arranged replications which allow to disentangle effects of field heterogeneity from genotype effects. Field heterogeneity might be caused by differences in soil water content, soil depth, soil fertility, water logging, soil compaction, root disease, and other factors (*e.g.* Araus, Kefauver, et al., 2018; Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Rebetzke et al., 2013). For some studies, different field management practices, *e.g.*, different irrigation or fertilization regimens, are applied to the genotypes. Genotypes, treatments, and field heterogeneity lead to distinct phenotypes, and phenotype-specific CT differences might be explained by different traits and not stomatal conductance alone, although phenotypic traits often are interlinked with each other. Quantitative trait loci have been shown to be often pleiotropic or co-located for CT and yield, above-ground biomass, plant height, and other traits (*e.g.* Mason and Singh, 2014; Rebetzke et al., 2013, citet in Deery, Rebetzke, Jimenez-Berni, James, et al., 2016). CT is strongly affected by above-ground biomass, morphological parameters such as plant height, FCC, leaf area index (LAI), rooting behavior, late senescence behavior, and consequently a larger green area during later stages, and even by the spatial orientation of leafs and spikes (Anderegg, Aasen, et al., 2021; Anderegg, Kirchgessner, et al., 2024; Oberholzer et al., 2017; Prashar and H. Jones, 2014; Perich et al., 2020; Rebetzke et al., 2013; M. P. Reynolds, Pask, et al., 2012). All of these sources are not independent. FCC for example might be caused by the genotype but also the field management or field heterogeneity, and an increased FCC might reduce the impact of the viewing geometry as also at nadir view, little soil is visible when FCC is saturated.

To observe the effects of genotypes and treatments on CT, uncertainties of CT estimates must be mitigated by estimating and correcting confounding sources of variance. For example Rebetzke et al. (2013) applied a mixed model and included the time of CT sampling as a fixed linear effect. Treier et al. (2024) proposed a multi-view approach in which CT estimates were derived from sequences of thermal images. Unlike approaches where CT estimates rely on orthomosaics, multi-view allowed for multiple CT estimates per plot and flight and to estimate covariates related to trigger timing and viewing geometry for each single measurement. The authors showed how the inclusion of trigger timing as a random effect in linear mixed models was allowing to increase consistency and genotype-specificity of the CT estimates. The aim of the study at hand was to empirically demonstrate how the multi-view approach can be used to disentangle multiple sources of variance and to separate undesired sources of variance from desired sources in a first step. A second aim was to show why these corrections matter with respect to the interpretability of the data. To that end, the multi-view method was applied in two wheat variety testing trials with contrasting field management regimens in two consecutive years of contrasting meteorological conditions. Complementary measurements were conducted to test hypotheses on wind conditions on the sensor, canopy cover, LAI, above-ground biomass, and plant height as important drivers of the thermal signal.

3.2 Methods

3.2.1 Field experiments and data acquisition

TIR measurements were conducted in two winter wheat variety testing experiments for two consecutive years (2020-2021 and 2021-2022) in the fields of the Agroscope agricultural research

station, Changins, Switzerland [46°23′55.4″N 6°14′20.4″E, 425 m.a.s.l., the World Geodetic System (WGS) 84]. The soil of the experimental site is a shallow Calcaric Cambisol (Baxter, 2007; Cárcer et al., 2019).

One trial comprised 30 modern registered European winter wheat varieties and is further referred to as the EuVar trial. The same varieties were seeded for the two years in three different treatment regimes. In the "maximal" regimen, one growth regulator and one fungicide treatment were applied. In the "medium" regimen, there was only the growth regulator application and not the fungicide application. In the "minimal" regimen, neither a growth regulator nor a fungicide was applied (see Tables S2.1 and S2.2 for more details). Fertilizers and herbicides were applied in three splits and at equal rates to all treatments according to the Proof of Ecological Performance (PEP) certification guidelines (Swiss Federal Council, 2013), which represent a minimal standard for best-practice for conventional agriculture in Switzerland. Each variety-treatment combination was repeated on three plots. Within single plots, eight sowing rows of the same wheat genotype were sown with a spacing of 15 cm between them resulting in an observable canopy of about 1.25 m x 6.7 m each. Within blocks of 3 by 10 plots, the genotypes were randomly distributed, and these blocks were randomly nested within three treatment replicates. Each replicate contained three blocks, and each block was treated with one of the three treatments. The 270 plots of the experiment span over 27 rows (which followed the tractor track direction) and 10 columns (Fig. S2.1). This experiment, the TIR data acquisition and multi-view processing were first described in Treier et al. (2024), where the same authors demonstrated the robustness of the multi-view approach and the method was shown to outperform commonly used orthomosaic-based approaches. The Methods are partially described here and in the Supplementary Materials for clarity, but for more information, it is referred to the study mentioned.

The second trial, further denoted SwiVar, comprised modern winter wheat genotypes and mixtures of two genotypes. The genotypes included registered varieties and candidate lines for inscription in the Swiss list of recommended wheat varieties. In the first year, there were 34 pure genotypes and two genotype mixtures. In the second year, there were 35 pure genotypes and one mixture. 31 genotypes and one mixture stayed the same between the two years. This performance trial included two different nitrogen treatment regimens. In one regimen, nitrogen fertilization was carried out according to common local agricultural practice following the PEP guidelines. In the second fertilizer regimen, no nitrogen fertilizer was applied. Herbicides were applied in both treatments according to the PEP guidelines. Each genotype was repeated in each treatment three times, resulting in 216 plots with the same row spacing as in EuVar and a canopy of about 1.25 m x 4.3 m each. Within the treatments, the plots were arranged in a randomized complete block design and the treatments were grouped into two separate blocks of 6 x 18 plots due to restrictions in available space and for simplifying nitrogen management (Fig. S2.2). In 2021, a sowing error occurred in three plots of one replication of SwiVar, which were seeded with the variety of the border plots and for these three genotypes, there were just two replications in the fertilized regimen (Fig. S2.2). The three plots were included in the analysis as genotype "border". SwiVar22 received an irrigation of 30 mm on 2022-05-23 due to lack of rain (Fig. S2.5).

The different experiment-year combinations are further referred to as EuVar21, EuVar22, SwiVar21 and SwiVar22 according to year of harvest. Tables S2.1 and S2.2 give an overview on the different treatments, fertilizer applications and the most important field interventions while Table S2.3 displays details on the chemical products used.

Air temperature, rainfall, radiation, wind speed, wind direction, relative humidity and VPD were obtained by a weather station of Meteoswiss which was located about 800 m from the experimental site at Changins [46°24′3.7″N 6°13′39.6″E, 458 m.a.s.l., WGS 84].

2021 was a relatively cool year with almost 700 mm precipitation between the beginning of

the year and harvest, while there was just 280 mm of precipitation for the same period in 2022. The temperature was on average 2.9 °C warmer from May to harvest for 2022 compared to 2021, and wheat developed faster in 2022 with the heading occurring 6 days earlier (Fig. S2.5). Harvest was 20 days earlier for EuVar22 compared to EuVar21. SwiVar22 was harvested 13 days before SwiVar21.

Flights were carried out between the onset of flowering and mid-senescence. In 2021, flights were conducted on two dates in each trial. In 2022, flights were conducted on four dates on EuVar22 and on six dates on SwiVar22 respectively. On specific dates, multiple flights were conducted at different time slots. To account for short-term variability, within each time slot at least two, mostly three flights were conducted with the same settings. A group of flights that were conducted at one time slot and date is further called a flight campaign. In total, 39 flights were performed on EuVar and 60 on SwiVar (for more details, see Supplementary Materials sections S2.4 & S2.5). Drone flights generally took place under close to optimal conditions with relatively low wind, although conditions in 2022 were more optimal than in 2021, when high and semitransparent cloud layers led to fluctuating light intensities for some flights in 2021 (Figs. S2.6 & S2.7).

The drone flew over the plots at a height of approximately 40 m, which allowed for a ground sampling distance (GSD) of about 5.2 cm/pixel. With a plot width of 1.5 m, this GSD resulted in more than 20 rows of pixels within the plots after excluding the border areas of the plots while still allowing for relatively short flights. The heading of drone and TIR camera was set to remain stable throughout the flight and did not change with changes in flight path direction. The resulting flight duration was between 6 and 9 min depending on the wind conditions and the total area recorded. The settings used resulted in an image pattern in which each spot in the trial was recorded on at least nine images from different perspectives. The camera was pointing toward the ground orthogonally (*i.e.* in nadir orientation). An uncalibrated DJI Zenmuse XT TIR sensor (SZ DJI Technology Co. Ltd., China) was used and a detailed description of the equipment and the settings used and of flight planning can be found in Supplementary Materials section S2.6. The experiments were neighbored by border plots and other experiments. To increase the number of measurements available for trend estimation, the flights covered not just the experiments but all wheat plots in the respective field surroundings, that is, border plots and other experiments on the same field. This helped reduce border effects by improving temporal and spatial corrections, as described in Treier et al. (2024). Supplementary Materials section S2.10 summarizes the pre-flight procedure. In short, the camera was turned on at least 15 min before each flight to allow the temperature signal to stabilize. The TIR images were saved as radiometric JPEG format. Following the protocol of Treier et al. (2024), no radiometric calibration was applied for later processing and only the internal calibration provided by the manufacturer was used.

For post-processing in the Structure-from-Motion-based photogrammetry software Agisoft Metashape (Agisoft LCC, St.Peterburg, Russia) and to allow time series analysis, thermal ground control points (GCPs) were distributed in the field in an evenly spaced shifted grid pattern (for more details, see Supplementary Materials section S2.11).

For the multi-view approach, digital elevation models (DEM) were needed on which the images could be projected. DEMs were based on both, TIR images and RGB images. For more details on the creation of DEMs, refer to Supplementary Materials section S2.7.

3.2.2 TIR image pre-processing

From radiometric JPEG format, 14-bit TIFF files were derived representing temperature in $°C$ x 1000 by using a Python 3.8 script (van Rossum, Guido and Drake, Fred L., 2009), a modified version of the Flir Image Extractor (`https://github.com/ITVRoC/FlirImageExtractor`).

The 14-bit TIFF files of the radiometric images as well as the RGB images were aligned in the structure-from-motion-based software Agisoft Metashape Professional (Agisoft LLC, St. Petersburg, Russia) and georeferenced (for details, see Supplementary Materials section S2.12). Plot masks were created for each plot in Qgis 3.16 (QGIS Development Team, 2022), to determine the regions of interest (ROIs) from which the data was used for analysis. A buffer of at least 25 cm was applied on plot width and length to account for inaccuracies in georeferencing.

The image information was reduced to a single value for each plot in each image by using the optimal percentile of all pixel values within each plot in each image. The procedure for finding an optimal percentile was described in Treier et al. (2024). In short, for each percentile, heritabilities were calculated from a mixed model with the R package SpATS (Rodríguez-Álvarez et al., 2018). The resulting percentile-heritability relations were plotted for graphical comparison and optimal percentile selection. The same percentile was used for the aggregation of all flights on one experiment within one year (for more details, see Supplementary Materials section S2.8).

3.2.3 Multi-view pre-processing

The single images were projected on the RGB DEMs by ray tracing as described in Roth, Aasen, et al. (2018), Roth, Camenzind, et al. (2020) and Treier et al. (2024). This allowed the projection of geographic coordinates (*e.g.* EPSG:2056 reference system) to image coordinates. As a result, plot masks of ROIs were created for each trigger position (*i.e.* for each image), where at least one plot was entirely inside the field of view (FOV) of the camera. For each plot on each TIFF file, all percentiles were extracted with a Python 3.8 script.

As plot-wise data was extracted for each image, the trigger timing could be determined from image meta data. The trigger timing of each image and the position of the experiment was known while the position of the sun was determined for each measurement as azimuth and elevation angle in Python using a script by John Clark Craig (`https://levelup.gitconnected.com/python-sun-position-for-solar-energy-and-research-7a4ead801777`, 2021). As Cartesian (*i.e.* orthogonal) coordinates were used and the position of the sun, the position of the plot centers and the position and orientation of the camera at the moment when the image was triggered were known, this allowed to calculate the geometric relations between sun, plot and drone by trigonometry as listed in Table S2.5 and illustrated in Fig. 3.1 (for more details, see Supplementary Materials section S2.9).

3.2.4 TIR data post-processing

After data extraction, the contribution of the different sources of CT variance to the total CT variance was estimated and CT was corrected for confounding sources of variance. Although the sources of variance might differ, they might be corrected by the same type of correction (Table 3.1). For example, while variance sources related to weather are ideally avoided by flying without wind and clouds, they still might affect the measurements in a temporal pattern. Such temporal variation mixes with the thermal drift, and is thus corrected by the same type of correction (Z. Wang et al., 2023). Correction for the different types of correction was achieved in a two-step approach (Fig. 3.2), as the computational burden of a one-stage approach was too heavy for multi-view data (Treier et al., 2024) and stage-wise approaches are proposed for the analysis of complex agricultural trials (Hans-Peter Piepho et al., 2012). In a first stage, the TIR measurements were corrected for non-geometric sources of variance. The residuals of the first stage were then analyzed to reveal the importance of geometric effects in a partial least squares regression (PLSR) analysis in a second stage. A plot-wise mean was calculated as a reference baseline. In the following, the two-stage approach is described in detail.

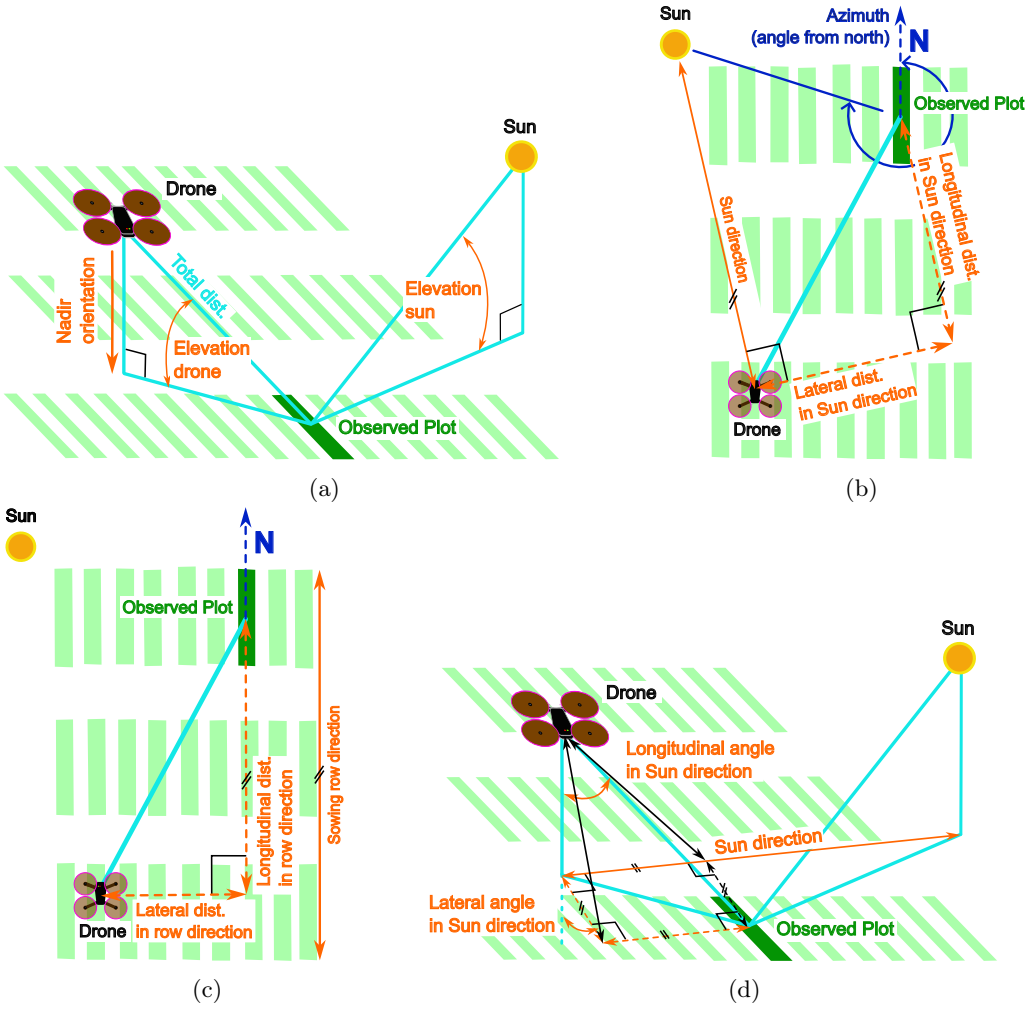


FIGURE 3.1: By knowing the position of the sun, the position of the plot and the position and orientation of the camera when an image is triggered (a), different geometric relations can be calculated. The position directly below the drone is in nadir orientation. The vertical angle at which drone and sun are seen from the observed plot are the elevations of drone and sun respectively. The azimuth of the sun is the clockwise horizontal angle at which the sun is seen from the observed plot from north (b). The position of the plot can be described as planar distance between drone and plot in direction of the sun (b) or in sowing row direction (c). Another option to describe the positon of the plot relative to the drone is by viewing angles as is shown for angles relative to sun direction (d), but not shown for the sowing row direction. Elements in the principal optical planes in drone or sun direction are in bright blue, cardinal direction in dark blue. The dimensions of interest and related covariates are in orange. Small black angle marks and short parallel black lines indicate perpendicularity and parallelism respectively.

3.2.4.1 Mixed model

The multi-view method provided several CT estimates for each plot (originating from different images). For each measurement, covariates related to trigger timing and viewing geometry were available which were used to analyze sources of variance and to correct the TIR measurements.

A mixed model (Eq. 3.1) was fitted in ASReml-R (Butler, 2019) to correct for temporal and spatial trends and experimental design factors (experiments, genotypes, treatments, replications). ASReml-R was chosen over other mixed model software due to its capability to model complex variance structures, which was important for the best possible consideration of nested structures (*e.g.* border plots) and temporal trends in this study. The mixed model

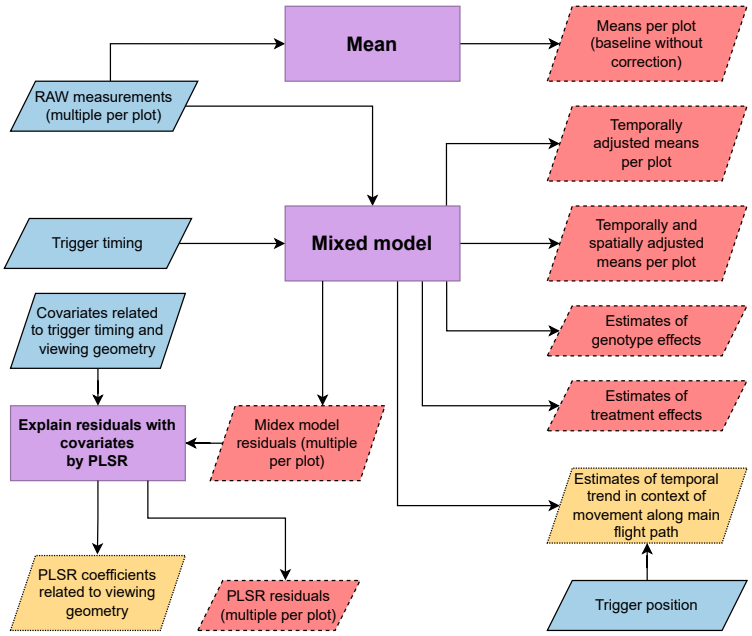


FIGURE 3.2: Flow-chart depicting the process of step-wise TIR measurement correction. TIR measurements and covariates (blue/solid-border parallelograms) were processed in different steps (purple rectangles) to derive estimates of plot-wise canopy temperature and residuals (red/dashed-border parallelograms) as well as trends related to trigger timing and viewing geometry (yellow/dotted-border parallelogram).

used was introduced and tested for robustness in Treier et al. (2024), where the single terms are explained in detail and mentioned here for clarity. Plot-based repeated CT measurements θ_{ijknp} for the i^{th} genotype, j^{th} trigger event, k^{th} treatment, n^{th} replication, and p^{th} plot were decomposed in factors related to genotypes (θ_{i}), treatments (τ_{k}), replications (r_{n}) and plots (ϕ_p) within a field. A temporal trend was modeled as a smooth spline $f_{\text{spl}}(\lambda j)$ along the sequential trigger events λ_j, where a trigger event j corresponds to a specific thermal image. A spatial model comprised two one-dimensional autocorrelation parts in row direction $f_{\text{AR(1)}}(r(p))$ (following tractor tracks) and column direction $f_{\text{AR(1)}}(c(p))$, where $f_{\text{AR(1)}}$ is a first order autoregression function of respective rows and columns at positions of plots in row direction $r(p)$ and column direction $c(p)$. In addition, a two-dimensional spatial autocorrelation $f_{\text{AR(1)}\times\text{AR(1)}}(c(p), r(p))$ was included in the spatial model. e_{ijknp} are measurement specific residuals. Genotypes and treatments were given unique IDs for each experiment covered in one flight, so the same genotype or treatment ID did not appear in the experiment of interest (*i.e.*, either EuVar or SwiVar) and also surrounding experiments or border plots at the same time, reducing the complexity of data structure to be handled by the models. The k experiment-specific treatments therefore also implicitly describe the different experiments. An interaction between the i^{th} genotype and the k^{th} treatment $(\theta\tau)_{ik}$ was applied only to the experiments of interest. For parts of other surrounding experiments and border plots, a simple additive effect was assumed for genotype and treatment for simplicity and to reduce computational capacity needed. ASReml-R allows to specify model terms for subsets of data with the "at()" statement, and the data could be processed differently for plots belonging to different experiments and border plots with the same ASReml-R model. The interaction between the k^{th} treatment and the n^{th} replication $(\tau r)_{kn}$ was just applied to EuVar, as the treatments were nested within the replications in EuVar, but not in SwiVar.

$$\begin{aligned}\theta_{ijknp} = {} & \theta_i + \tau_k + \phi_p + r_n + (\theta\tau)_{ik} + (\tau r)_{kn} + && \text{(Design-Factors)}\\ & f_{\mathrm{AR(1)\times AR(1)}}(c(p), r(p)) + f_{\mathrm{AR(1)}}(c(p)) + f_{\mathrm{AR(1)}}(r(p)) + && \text{(Spatial-Autoregression)}\\ & f_{\mathrm{spl}}(\lambda_j) + && \text{(Temporal-Trend)}\\ & e_{ijknp} && \text{(Residuals)}\end{aligned} \tag{3.1}$$

3.2.4.2 Estimate the temporal trend

Mixed models decompose variance into variance components and the different components can subsequently be included in models to predict the effects of individual variables. The temporal trend was estimated as the effect of the j^{th} trigger event/image along the duration of a single flight modeled with a smooth spline $f_{\mathrm{spl}}(\lambda_j)$ in Eq. 3.1.

3.2.4.3 Plot-wise CT estimates

After fitting the models by Eq. 3.1, single plot-wise CT values ($\hat{\theta}_p$) were estimated with different prediction models to estimate the effect and importance of different variables within the mixed model.

To have a baseline for comparison, the mean plot temperature $\hat{\theta}_p^{mean}$ was calculated on the measurements of the individual images j available for one plot p without applying the mixed model or considering any covariates,

$$\hat{\theta}_p^{mean} = \mathrm{mean}(\theta_{jp})\,. \tag{3.2}$$

A first mixed model-based prediction model included all variance components of the mixed models except for the temporal trend λj (Eq. 3.3). It estimated the individual plot-wise CT values as the sum of genotype effects (θ_i), treatment effects (τ_k), plot effects (ϕ_p), row r_p, column c_p and replication effects (r_n) at the position of plot p,

$$\hat{\theta}_p^{t\_c} = \hat{\theta}_{ikpr_{(p)}c_{(p)}n} = \theta_i + \tau_k + \phi_p + r_p + c_p + r_n\,. \tag{3.3}$$

By discarding the temporal trend in the prediction, the plot-wise estimates were plot-wise means $\hat{\theta}_p^{t\_c}$ adjusted along the temporal dimension and therefore temporally corrected (t_c). In the next step, the spatial trends of row r_p and column c_p were discarded in prediction,

$$\hat{\theta}_p^{ts\_c} = \hat{\theta}_{ikpn} = \theta_i + \tau_k + \phi_p + r_n\,. \tag{3.4}$$

The plot-wise estimates $\hat{\theta}_p^{ts\_c}$ of Eq. 3.4 were temporally and spatially corrected (ts_c). To consider possibly strong treatment effects, for each flight, the mean treatment temperatures were calculated and subtracted from $\hat{\theta}_p^{ts\_c}$,

$$\hat{\theta}_p^{t\_defl} = \hat{\theta}_{ipn} = \hat{\theta}_{ikpn} - mean(\tau_k)\,. \tag{3.5}$$

The plot-wise estimates $\hat{\theta}_p^{t\_defl}$ represent the sum of a genotpye, a genotype-treatment interaction, a plot, and a replication effect after subtracting a mean treatment effect $mean(\tau_k)$, leaving out all other effects of Eq. 3.1. They are temporally and spatially corrected, and

treatment effects were deflated (t_defl), meaning that only a possible genotype-treatment interaction is left in the estimate, but not the main treatment effect.

Predictive models (Eq. 3.3, Eq. 3.4 & Eq. 3.5) just comprised plots belonging to EuVar or SwiVar. As uncooled and uncalibrated TIR cameras provide a low absolute temperature accuracy, just relative temperature differences between the plots were analyzed from this stage onward (H. G. Jones, Serraj, et al., 2009; Kelly et al., 2019).

For a comparison of the effects of the single variables, the variance of the plot-wise estimates derived from the different prediction methods (Eq. 3.2 - Eq. 3.5) was calculated for all flights.

3.2.4.4 Multispectral measurements

The trials were also monitored with an airborne Micasense RedEdge-MX Dual multispectral camera (MicaSense Inc., Seattle, Washington, USA) throughout the growing season. With multispectral data, vegetation indices (VI) were calculated to obtain approximative estimates of LAI and biomass. The images were aligned in Agisoft to generate 10 band orthophotos covering all the experiments. Details on the spectral properties of the 10 bands of the sensor are described in Table S2.6. Based on these bands, four VIs were calculated. DVI, SAVI and EVI (see Table 3.2 for full names and equations) are commonly used VIs to estimate the LAI of wheat (W. Li et al., 2023) while SAVI was also shown to be correlated with above-ground biomass (F. Wang et al., 2022). NDVI was calculated as a reference to the emissivity (Diaz et al., 2021) of the plants. The same masks as for the TIR images were used to mark ROIs on the multispectral orthomosaics. The 50^{th} percentile (median) was used to aggregate VI values within single ROIs to single values with a Python 3.8 (van Rossum, Guido and Drake, Fred L., 2009) script for subsequent analysis (for more details, see Supplementary Materials section S2.15).

VIs were recorded on multiple dates and the VIs were correlated to CT that was measured at the date closest to the VI recording.

Table 3.2: Multispectral VIs used to approximate biomass (DVI, SAVI, EVI) and LAI (SAVI) and as a reference to the emissivity (NDVI).

Index	Full name	Formula		Reference
DVI	Difference Vegetation Index	$DVI = NIR_{842} - Red_{668}$	(3.6)	Tucker Tucker, 1979
EVI	Enhanced Vegetation Index	$EVI = 2.5 \cdot \frac{NIR_{842} - Red_{650}}{NIR_{842} + 6 \cdot Red_{650} - 7.5 \cdot Blue_{444} + 1}$	(3.7)	Huete et al. Huete et al., 2002
NDVI	Normalized Difference Vegetation Index	$NDVI = \frac{NIR_{842} - Red_{668}}{NIR_{842} + Red_{668}}$	(3.8)	Rouse et al. Rouse et al., 1974
SAVI	Soil Adjusted Vegetation Index	$SAVI = 1.5 \cdot \frac{NIR_{842} - Red_{650}}{NIR_{842} + Red_{650} + 0.5}$	(3.9)	Huete Huete, 1988

3.2.4.5 Estimate the spatial trend

The spatial trend of the plots p across the field in row $c_{(p)}$ and column $c_{(p)}$ direction $\hat{\theta}_{r_{(p)}c_{(p)}}$ was estimated as the difference between the plot-wise CT estimates after a temporal correction $\hat{\theta}_p^{t\_c}$ and after a temporal and spatial correction $\hat{\theta}_p^{ts\_c}$,

$$\hat{\theta}_{r_{(p)}c_{(p)}} = \hat{\theta}_p^{t\_c} - \hat{\theta}_p^{ts\_c} = \hat{\theta}_{ikpr_{(p)}c_{(p)}n} - \hat{\theta}_{ikpn} \ . \tag{3.10}$$

Assuming the consistency of spatial effects between flights, these plot-wise spatial trends would be correlated between flights. As a larger number of observations per plot is assumed to increase the repeatability of the estimations (Treier et al., 2024), spatial trends were calculated for all flights individually, but also for all flights within a campaign simultaneously. With at least two flights per campaign, this was increasing the number of observations per plot at least two-fold.

3.2.5 Geometric effects

As shown in Table 3.1, multiple sources of variance have a geometric effect on CT readings. They can be caused by vignetting, viewing geometry-related effects, atmospheric effects, and geometric emission and reflectance patterns (*i.e.* BRDF).

Two different methods were applied to account for geometric effects. In a first approach, the covariance of the residuals e_{ijknp} of the mixed models (Eq. 3.1) with geometric covariates was examined by PLSR with the R-package PLS (Mevik and Wehrens, 2007).

The linear relations between geometric covariates (Table S2.5) and residuals were visually identified in an exploratory data analysis and where necessary, trigonometric transformations were applied to angular covariates for linearization. Covariates with an apparent linear relationship to the residuals (Table 3.3) were included in the PLSR model. In addition, the interaction between the longitudinal distance in the direction of the sun and the sine of the elevation angle of the drone was part of the PLSR analysis, as it describes the path of light from the sun to the drone. The inclusion of the two terms without interaction does not describe the path adequately as positions in front and behind the drone in the direction of the sun get the same values.

Table 3.3: Covariates with evident trends were identified among all orignial covariates and transformations were applied to linearize the trends. Several trends can describe the same spatial dimension (e.g. Lateral in direction of sowing rows).

Linearized covariates	**Dimension**	**Transformation**	**Name in Model**
Sine of the elevation angle of the drone	Elevation of the drone	Sine	Drone-Elevation-sin
Lateral distance in direction of sowing rows	Lateral in direction of sowing rows	None	RowDir-lat-Dist
Absolute lateral distance in direction of sowing rows		Absolute value	RowDir-lat-Dist-abs
Cosine of lateral angle in direction of sowing rows		Cosine	RowDir-lat-Angl-cos
Absolute value of lateral angle in direction of sowing rows		Absolute value	RowDir-lat-Angl-abs
Longitudinal distance in direction of sowing rows	Longitudinal in direction of sowing rows	None	RowDir-lon-Dist
Absolute longitudinal distance in direction of sowing rows		Absolute value	RowDir-lon-Dist-abs
Cosine of longitudinal angle in direction of sowing rows		Cosine	RowDir-lon-Angl-cos
Absolute value of longitudinal angle in direction of sowing rows		Absolute value	RowDir-lon-Angl-abs
Lateral distance in direction of the sun	Lateral in direction of the sun	None	SunDir-lat-Dist
Absolute lateral distance in direction of the sun		Absolute value	SunDir-lat-Dist-abs
Cosine of lateral angle in direction of the sun		Cosine	SunDir-lat-Angl-cos
Absolute value of lateral angle in direction of the sun		Absolute value	SunDir-lat-Angl-abs
Longitudinal distance in direction of the sun	Longitudinal in direction of the sun	None	SunDir-lon-Dist
Absolute longitudinal distance in direction of the sun		Absolute value	SunDir-lon-Dist-abs
Cosine of longitudinal angle in direction of the sun		Cosine	SunDir-lon-Angl-cos
Absolute value of longitudinal angle in direction of the sun		Absolute value	SunDir-lon-Angl-abs
Interaction between longitudinal distance in direction of the Sun and sine of the elevation angle of the drone	Interaction SunDir-Drone-Elevation	None	Interact.-SunDir-Drone
Trigger timing	Time	None	Trigger-time
Total distance between drone and plot	Distance	None	Dist-tot

The PLSR coefficients were calculated for each covariate and each flight to determine which covariate explained the most of the variance of the residuals e_{ijknp} from the pre-processing model (Eq. 3.1). Several linearized covariates in the PLSR model described the same spatial dimension (Table 3.3). With the aims of avoiding redundancy and simplifying the model, the

model was reduced to contain only relevant dimensions. Relative PLSR coefficient magnitudes $\beta_{\mathrm{rel,i}}$ were calculated within each flight and each covariate as:

$$\beta_{rel,i} = \frac{|\beta_i|}{\sum_{i=1}^{n}|\beta_i|}, \tag{3.11}$$

where β_{i} denotes the PLSR coefficient of the i^{th} of n covariates. To determine the least descriptive covariates, the medians of relative magnitude of the covariates β_{i} over all flights j were calculated.

$$\beta_{med,i} = med\{|\beta_{rel,i;j}|\} \tag{3.12}$$

Covariates with the lowest median were skipped in a supervised backward feature elimination until the most descriptive transformation types and dimensions were left in the model (similar to methods summarized in Mehmood et al., 2012).

In a second approach to account for geometric effects, a generalized ex ante vignetting correction was applied as described in Treier et al. (2024). A generalized vignetting correction image was created in an indoor experiment, with its pixel values representing a mean vignetting effect as relative temperature difference within an image under controlled conditions. The pixel values of the correction image were then subtracted from the corrsponding pixels of all TIR images (for more details, see Supplementary Materials section S2.16).

Subsequent analysis with mixed models and PLSR analysis was performed on TIR images with and without vignetting correction.

3.2.6 Reference measurement and complementary experiments to better understand phenotypic variability, viewing geometry and thermal drift as sources of CT variance

The mixed model allowed estimation of the contribution of genotypes, experimental treatment regimens, spatial trends, and thermal drift to the overall variance. With the PLSR models, the contribution of viewing geometry to the overall variance was examined. To demonstrate the relationship between CT and the phenotypic variability of genotypes and treatment regimens, reference measurements were made on wheat phenotypes similar to Das, J. Christopher, Apan, Roy Choudhury, et al. (2021). CT was compared with grain yield, FCC, plant height, flag leaf rolling, flag leaf senescence, and multispectral indices as approximations of LAI and above-ground biomass (Table 3.2) by means of Pearson correlation. Complementary experiments were conducted to demonstrate the impact of apparent soil cover and wind on TIR readings qualitatively.

3.2.6.1 In-field reference measurements of phenotypic traits

Grain yield was measured with a combine harvester. The water content of the grain was determined with a Dickey-John GAC 2100 grain moisture tester within 24 hours after harvest and the grain yield per ha was noralized at 15 % gravimetric water content.

Plant height was measured with a measuring rod in five randomly chosen spots within each plot, and the mean taken as plot-wise plant height. It was measured from the soil to the tip of the ears without considering awns.

With dry conditions, leaf rolling was observed in season 2022 and visually rated in the field according to Pask et al. (2012). Leaf rolling ratings ranged from 0 to 3 where 0 corresponded to no rolling, 1 to a loosely rolled leaf ($<$ 33 % of leaf rolled), 2 to a moderately rolled leaf (34 - 66 % rolled) and 3 to a tightly rolled leaf ($>$ 67 % rolled). Flag leaf rolling was compared with the CT measurement performed on a date closest to the rolling scoring date, and the CT differences between the groups were examined with a Wilcoxon signed-rank test.

On the second flight date of EuVar21, senescence had already progressed. Therefore, flag leaf senescence ratings are presented for both EuVar21 measurements dates but not for the other trials. Flag leaf senescence was rated according to E. A. Chapman et al. (2021) and the ratings correspond to the proportions of senescent yellow leaf area of the flag leaf. 0 % corresponds to a fully green leaf and 100 % corresponds to a fully senescent leaf.

3.2.6.2 Qualitative demonstration of impact of apparent soil cover

A handheld calibrated high-resolution thermal camera (VarioCAM High Definition, Jenoptik, Jena, Germany) was used to demonstrate the influence of apparent soil cover qualitatively. This camera also included an RGB sensor which allowed a comparison of visible color images with thermal images of the very same scene.

3.2.6.3 Multi-view analysis of FCC from RGB data to demonstrate the correlation with CT

To examine the relationship between apparent CT and apparent canopy cover, the FCC was estimated based on RGB images as proposed by Deery, Rebetzke, Jimenez-Berni, James, et al. (2016). On 6 June 2022, a flight with a DJI Air 2S drone (SZ DJI Technology Co. Ltd., China) was performed in both experiments. The flight height was 20 m and the speed was limited to $3\,\mathrm{m\,s^{-1}}$. The front overlap was 65 % and the side overlap was 85 %. These settings resulted in a GSD of $\approx 5.5\,\mathrm{mm}$. While such a GSD may be considered too large for a very detailed examination of apparent soil cover, it is sufficient to demonstrate general trends.

Images were saved in 8-bit JPEG format and 16-bit DNG raw format. The DNG files were transformed to TIFF file format in Python 3.8 (van Rossum, Guido and Drake, Fred L., 2009). Using the interactive image analysis tool Ilastik (Berg et al., 2019), pixels of the TIFF images were segmented into three classes: green plant, senescent plant, and background. With these classes, FCC could be calculated as:

$$FCC = \frac{PN_{\text{green plant}} + PN_{\text{senescent plant}}}{PN_{\text{green plant}} + PN_{\text{senescent plant}} + PN_{\text{background}}} \tag{3.13}$$

where PN denotes the number of pixels of a specific class in an area of interest. Multiple plot-wise FCC values were fitted with the same mixed model in ASReml-R as CT (Eq. 3.1) but replacing CT by FCC. Adjusted means for plot-wise FCC were estimated, and the FCC residuals were analyzed with respect to viewing geometry.

3.2.6.4 Geometric patterns of atmospheric effects

TIR readings are also affected by atmospheric effects which depend on the path length between the sensor and the target (Jimenez-Berni, P. J. Zarco-Tejada, et al., 2009; Meier et al., 2011). To demonstrate the geometric nature of this effect, a simple data simulation was performed. Assuming a perfect nadir orientation of the sensor, the point directly below the drone is closer to the drone than points toward the edges of the image, *i.e.* the path length between sensor and plot is increased, which increases attenuation of TIR radiation and decreases transmittance of the atmosphere. Taking a simplified assumption of an attenuation of $0.001\,\mathrm{K\,m^{-1}}$ through the atmosphere (Meier et al., 2011), a theoretical attenuation effect was calculated at two flight heights (40 m and 300 m).

3.2.6.5 Fan experiment to determine the influence of wind

Kelly et al. (2019) and Yuan and Hua (2022) described a strong relation between temporal drift of TIR measurements and wind on the sensor. To confirm this link for the sensor used, a

fan experiment was set up, inspired by these two studies. The sensor was placed indoors in a dim environment at room temperature, pointing at a uniform hard foam PVC sheet. A fan and a lamp were used to cool and heat the sensor respectively. The apparent temperature of the PVC sheet and the standard deviation of the pixel-wise temperature were analyzed To examine whether sudden and strong temperature gradients have a sustained influence on subsequent TIR readings, warm and hot disturbance objects (hands at body temperature and a water cooker with boiling water) were introduced into the scene several times for several seconds each (for more details, see Supplementary Materials section S2.17).

3.2.7 Treatment deflation for correlation estimates

Strong treatment effects can be dominant and mask genotype effects, especially when values are compared by correlations, and the main driver of correlation is a treatment effect. To avoid inflated correlations of possibly dominant treatment effects, correlations were calculated on original data, on data after temporal and spatial correction, and on data after a treatment effect correction. The treatment effects were corrected for by subtracting the mean treatment effects from the plot-wise values after temporal and spatial correction.

3.2.8 Correction of reference measurement and correlation with CT

In-field reference measurements (yield, plant height, FCC, multispectral indices) were fitted with mixed models as done with CT. A model similar to Eq. 3.1 but without a temporal component was fitted in ASReml-R to correct for spatial trends. CT values before spatial correction ($\hat{\theta}_p^{mean}$ & $\hat{\theta}_p^{t-c}$) were correlated with uncorrected reference measurements. CT after spatial and temporal correction $\hat{\theta}_p^{ts-c}$ was correlated with spatially corrected reference measurements and treatment deflated CT $\hat{\theta}_p^{t-defl}$ was correlated with treatment deflated reference measurements.

3.3 Results

3.3.1 Percentile choice to aggregate pixel values into uncorrected data

For EuVar21, EuVar22 and SwiVar21, the 50$^{\text{th}}$ percentile (median) was chosen to aggregate all pixel values within a ROI into a single value. For EuVar22, the biomass in the non-fertilized part of the experiment was low, leading to large proportions of visible soil in the thermal images. Therefore, the 25$^{\text{th}}$ percentile was chosen as it better represented CT, containing fewer background signal from the soil (Fig. S2.8). The resulting uncorrected plot-wise CT estimates $\hat{\theta}_p^{mean}$ (Fig. S2.9 & Fig. S2.15) contained strong temporal and spatial trends.

3.3.2 Correcting for temporal and spatial trends

The mixed model (Eq. 3.1) allowed the estimation of the impact of sources of variance not related to viewing geometry. Fig. 3.3a shows an example of the temporal trends $f_{\text{spl}}(\lambda_j)$ estimated for the three flights of the SwiVar22 campaign flown on 2022-06-14 at 13:00. All three flights of the campaign were processed with the mixed model at once. The color of the line indicates the motion of the drone in the direction of the main flight path. The pattern of increasing and decreasing temperature seemed to be switching with the direction of motion of the drone, but this trend did not seem to be persistent, as it can be seen especially with the third flight, where the patterns of temperature and flight direction did not coincide anymore. Temporal trend estimates for all flights can be looked up at Fig. S2.10 and Fig. S2.16 for EuVar and SwiVar respectively. The resulting estimates after removing temporal trends $\hat{\theta}_p^{t-c}$

(Eq. 3.3), still contain strong spatial patterns that are not consistent within campaigns (*e.g.* Fig. 3.3d for the first flight of the same campaign as in Fig. 3.3a, Fig. S2.11 and Fig. S2.17 for all estimates $\hat{\theta}_p^{t-c}$ of EuVar and SwiVar, respectively).

3.3.3 Estimating the effect of experimental treatments

After correcting for temporal and spatial trends (Eq. 3.4), plot estimates $\hat{\theta}_p^{ts-c}$ containing genotype, treatment, and plot effects could be derived. When looking at $\hat{\theta}_p^{ts-c}$ for the same flight as Fig. 3.3d, a strong treatment effect was evident between the left and right sides of the experiment, where the cooler left side corresponded to the fertilized part of the experiment and the hotter right part to the unfertilized part (see Fig. S2.12 and Fig. S2.18 for all estimates $\hat{\theta}_p^{ts-c}$ of EuVar and SwiVar, respectively).

Mean treatment effects were estimated for all flights of EuVar (Fig. 3.3b, Fig. S2.14) and SwiVar (Fig. 3.3c, Fig. S2.20) as deviation from the mean experiment temperature. Within both experiments, the treatment effects were consistent for the two years, but stronger in 2022. However, for EuVar, the treatment effects were small, with a maximum difference of ∼0.15 °C in 2021 and ∼0.32 °C in 2022. The "minimal" regimen featured the lowest temperature, followed by the "maximal" and "medium" regimen. The differences between the cooler fertilized and the warmer non-fertilized treatment regimen of SwiVar were larger. In 2021, the maximum difference was around ∼0.38 °C while for 2022, strong treatment effects were observed with a maximum difference of approximately ∼4.8 °C.

3.3.4 Estimating the effect of genotypes and genotype-treatment interactions

When also removing mean treatment effects (Eq. 3.5), estimates were corrected for spatial, temporal, and main treatment effects. On an experiment scale, estimates $\hat{\theta}_p^{t-defl}$ did not contain strong spatial trends or treatment effects anymore and appeared relatively flat. The variance between the plot-wise estimates $\hat{\theta}_p^{t-defl}$ as seen in Fig. 3.3f corresponded to genotypic effects and genotype-treatment interactions without the main treatment effects (see Fig. S2.13 and Fig. S2.19 for all estimates $\hat{\theta}_p^{t-defl}$ of EuVar and SwiVar respectively).

3.3.5 Impact of correction for non-geometric trends on variance of estimates

Confounding sources of variance, mainly temporal and spatial trends, contributed significantly more to total variance than experimental sources of variance related to the phenotypes.

When correcting plot-wise CT estimates for temporal effects ($\hat{\theta}_p^{t-c}$), temporal and spatial effects ($\hat{\theta}_p^{ts-c}$) and finally also deflating treatment effects ($\hat{\theta}_p^{t-defl}$), the variance of the adjusted plot estimates was constantly decreasing (Fig. 3.4a). The variance of $\hat{\theta}_p^{t-defl}$, which still comprised genotypic variance, variance of genotype-treatment interactions, and plot effects, was orders of magnitude smaller than the initial variance of uncorrected plot estimates $\hat{\theta}_p^{mean}$. The mean variance decreased from 2.74 K^2 to 0.09 K^2 for EuVar21 and from 8.40 K^2 to 0.42 K^2 for EuVar22. For SwiVar21, variance decreased from 2.75 K^2 to 0.02 K^2 and from 7.68 K^2 to 0.32 K^2 for SwiVar22.

The greatest variance reduction occurred with the temporal and spatial correction after which the variance was below 0.5 K^2, except for SwiVar22. The variance was similar for $\hat{\theta}_p^{ts-c}$ and $\hat{\theta}_p^{t-defl}$ for all experiments but for SwiVar22, where the variance decreased a lot by treatment deflation, indicating a mild treatment effect for EuVar21, EuVar22 and SwiVar21 but a strong treatment effect for SwiVar22.

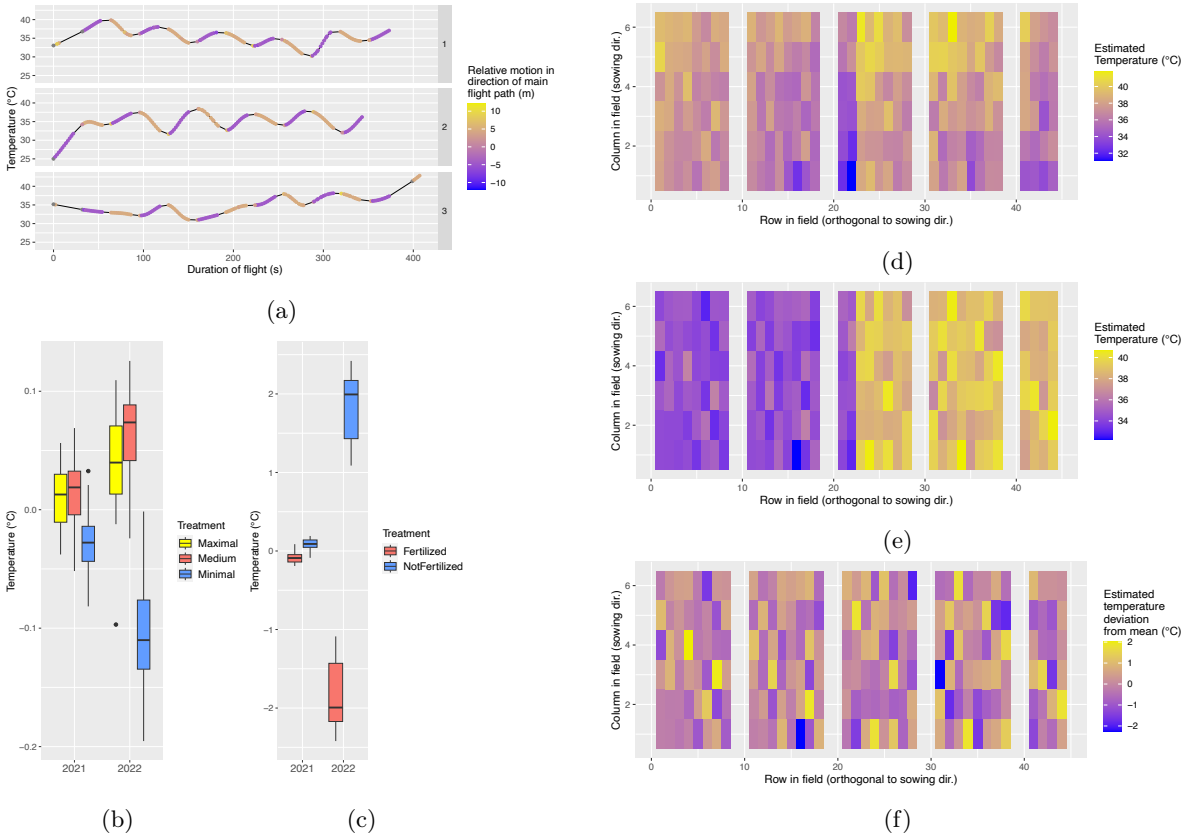


FIGURE 3.3: Sources of CT variance not related to viewing-geometry: Thermal drift of TIR measurements for the three flights of the campaign on 2022-06-14 at 13:00 was contextualized with the motion in the direction of the main flight path (a). The three rows are the three individual flights within the campaign. The colors indicate the motion in the direction of the main flight path. Purple indicates flights in one direction, and yellow indicates flights in the opposite direction of the flight path grid. For gray points, thermal drift was estimated on the basis of the mixed model, while there was no corresponding measurement of motion along the main flight path. For the estimation of the trends, all three flights were included in the same mixed model (Eq. 3.1). The box plots indicate the mean treatment effects for all flights in both years for EuVar (b) and SwiVar (c). After correcting the first flight of the campaign shown in (a) for temporal trends, the adjusted estimates $\hat{\theta}_p^{t-c}$ (Eq. 3.3) still contain significant, apparently spatial trends (d). After correction CT estimates for temporal and spatial trends (Eq. 3.4), plot estimates $\hat{\theta}_p^{ts-c}$ contained the genotype, treatment, and plot effects (e). When also subtracting mean treatment temperatures ($\hat{\theta}_p^{t-defl}$, Eq. 3.5), just genotype- and plot-effects were left and the interaction effect between treatment and genotype (f).

3.3.6 Impact of correcting CT for non-geometric trends on correlation between CT and phenotypic traits

Yield, plant height, four multispectral indices (DVI, EVI, NDVI, SAVI) and in 2022 also FCC were measured as phenotypic reference traits, as they represent possible physiological sources of CT variance for EuVar (Figs. S2.21 & S2.22) and SwiVar (Figs. S2.23 & S2.24). In EuVar21, senescence ratings were performed. Flag leaf rolling was rated in 2022 as an indicator of drought stress. In-field reference measurements were compared with CT of corresponding flights by Pearson correlation. Uncorrected CT values were correlated with the uncorrected reference measurements. Corrected CT was correlated with corrected reference measurements and treatment deflated CT was correlated with treatment deflated reference measurements. For yield, plant height, and FCC, a general overview of the correlations with CT is presented

in Fig. 3.4b - Fig. 3.4d. For each trial in each year, two flights conducted at two distinct dates were analyzed for each experiment before treatement deflation (Figs. 3.5a, c, e, g) and after (Figs. 3.5b, d, f, h).

3.3.6.1 Correlation between CT and yield

Yield at 15 % gravimetric water content was correlated with CT in conditions with and without water limitation and in the presence of weaker and stronger treatment effects. Significant correlations tended to be more consistent over all flights after applying different corrections.

Correlations were increased by the different corrections in the wet year 2021 for the relatively heterogeneous set of genotypes of EuVar21. Uncorrected CT $\hat{\theta}_p^{mean}$ was significantly correlated with yield only for 7 out of 22 flights and correlations were negative and weak to moderate (Fig. 3.4b). After correction ($\hat{\theta}_p^{ts-c}$), correlations were weak to strong and significant for all 22 flights ($p \leq 0.01$).

For the same genotypes in the dry year (EuVar22), uncorrected CT for 15 out of 17 flights was significantly and negatively correlated with yield with weak to strong correlations. After temporal and spatial correction, only 6 flights showed a weak significant correlation with yield. Therefore, the correlation between CT and EuVar22 yield was mainly driven by spatial trends. For both trials of EuVar, deflation of treatments ($\hat{\theta}_p^{t-defl}$) had little effect.

For the less heterogeneous genotypes of SwiVar, the trends were similar for both years. Initially, SwiVar21 and SwiVar22 showed a very broad range of correlations between yield and uncorrected CT values $\hat{\theta}_p^{mean}$. After correction ($\hat{\theta}_p^{ts-c}$), more correlations were significant and mostly negative, except for SwiVar21, where two correlations were positive. For SwiVar22, all 32 flights were significantly and negatively correlated with yield. However, after deflating the treatment effects ($\hat{\theta}_p^{t-defl}$), correlations were no longer significant for SwiVar in both years.

The differences between the data with and without vignetting correction were small, except for the $\hat{\theta}_p^{ts-c}$ values of EuVar22, where correlations with yield were relatively random. To have a more robust estimate of the reliability of these correlations, CT was also estimated based on all flights within campaigns (Fig. S2.25) and then correlated with yield. The general pattern of correlations was similar to that based on individuals flights.

Correlations of selected flights (Fig. 3.5) are in accordance with this general pattern with strongest and most highly significant correlations for EuVar21 ($p \leq 0.001$). The correlation in SwiVar was strongly driven by treatment effects, and the correlations were no longer significant after deflating treatment effects.

3.3.6.2 Correlation between CT and plant height

Significant correlations between CT and plant height were negative for all flights (Fig. 3.4c). For all experiments, the correlations became stronger and more flights became significantly correlated with plant height after the corrections. After correcting for temporal, spatial and treatment effects, all flights were significantly correlated with plant height except four EuVar21 flights. Deflating treatment effects did not change the correlations much for EuVar, but led to more negative correlations for the less heterogeneous genotypes of SwiVar in both years, but especially during the hot season of SwiVar22.

Looking at selected flights (Fig. 3.5), the correlation between CT and plant height was weaker and less significant in the trial with heterogeneous genotypes during the wet year (EuVar21), compared to all other trials, which showed all highly significant correlations ($p \leq 0.001$), except for SwiVar22, where this was the case only after treatment deflation.

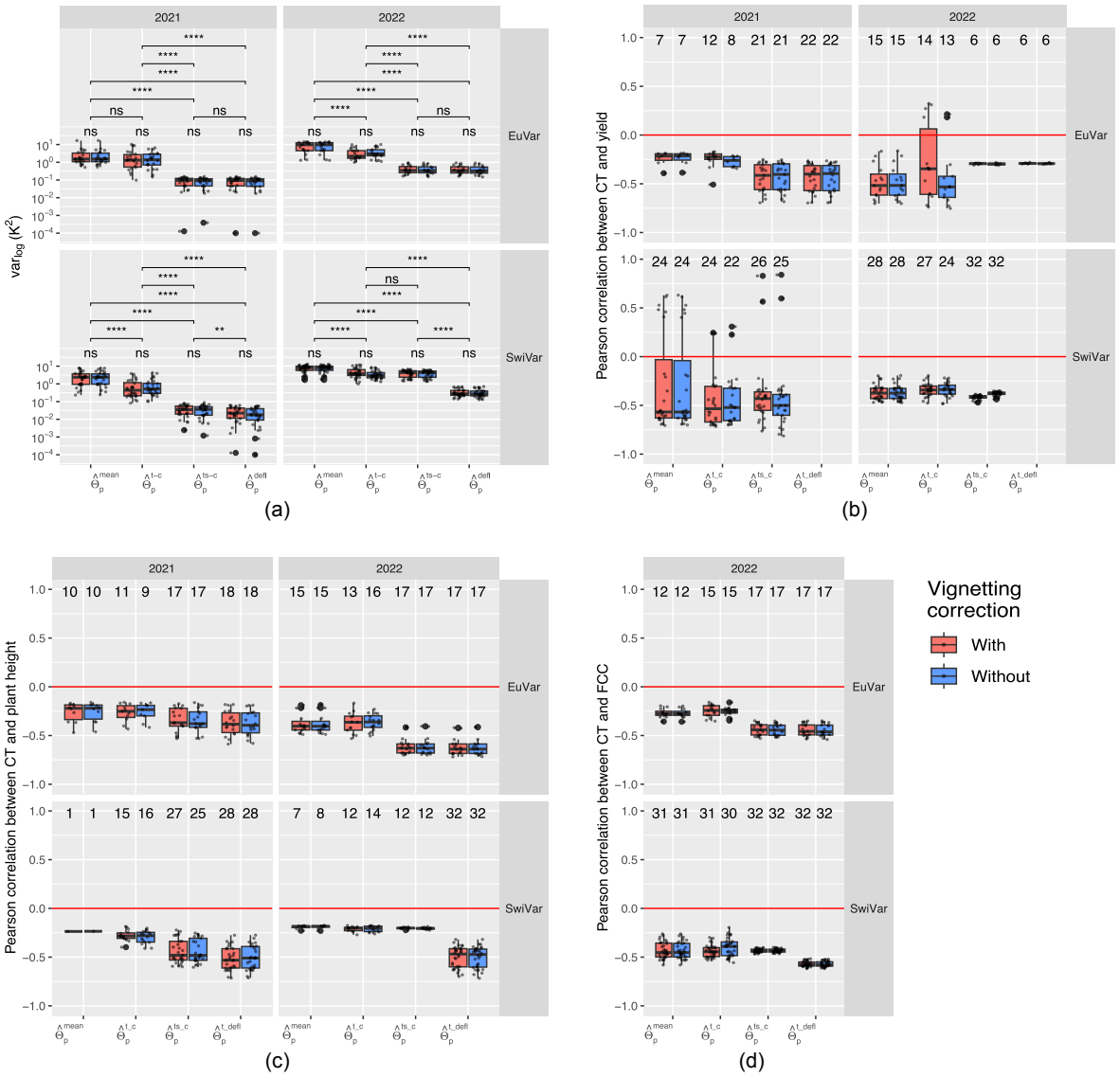


FIGURE 3.4: Variance of CT estimates after different corrections steps and relationship between CT and in-field reference measurements: (a) Comparison of the variance of uncorrected plot-wise estimates (Eq. 3.2) over all flights with CT estimates after correcting with the mixed model (Eq. 3.3, Eq. 3.4 & Eq. 3.5) Significant differences between correction steps are indicated based on a pair-wise t-test. Significance levels: ns: $p > 0.05$; *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$; ****: $p \leq 0.0001$. Note that a logarithmic scale is used! The CT estimates without and with correction were also correlated to in-field reference measurements, namely (b) yield, (c) plant height and (d) FCC. Just correlations significant at $p \leq 0.01$ are shown. The number above the boxplots indicates the number of significant correlations included in the respective box plots.

3.3.6.3 Correlation between CT and FCC

As for plant height, the correlations with FCC became more significant and stronger with the corrections applied (Fig. 3.4d & Fig. 3.5). For EuVar22 and for SwiVar22, CT of all flights was significantly correlated with FCC after temporal and spatial correction. For EuVar22, treatment deflation did not much change the correlations. For SwiVar22, correlations became stronger with treatment deflation, indicating a genotypic effect as the driver of the correlation between CT and FCC, partially masked by a strong treatment effect.

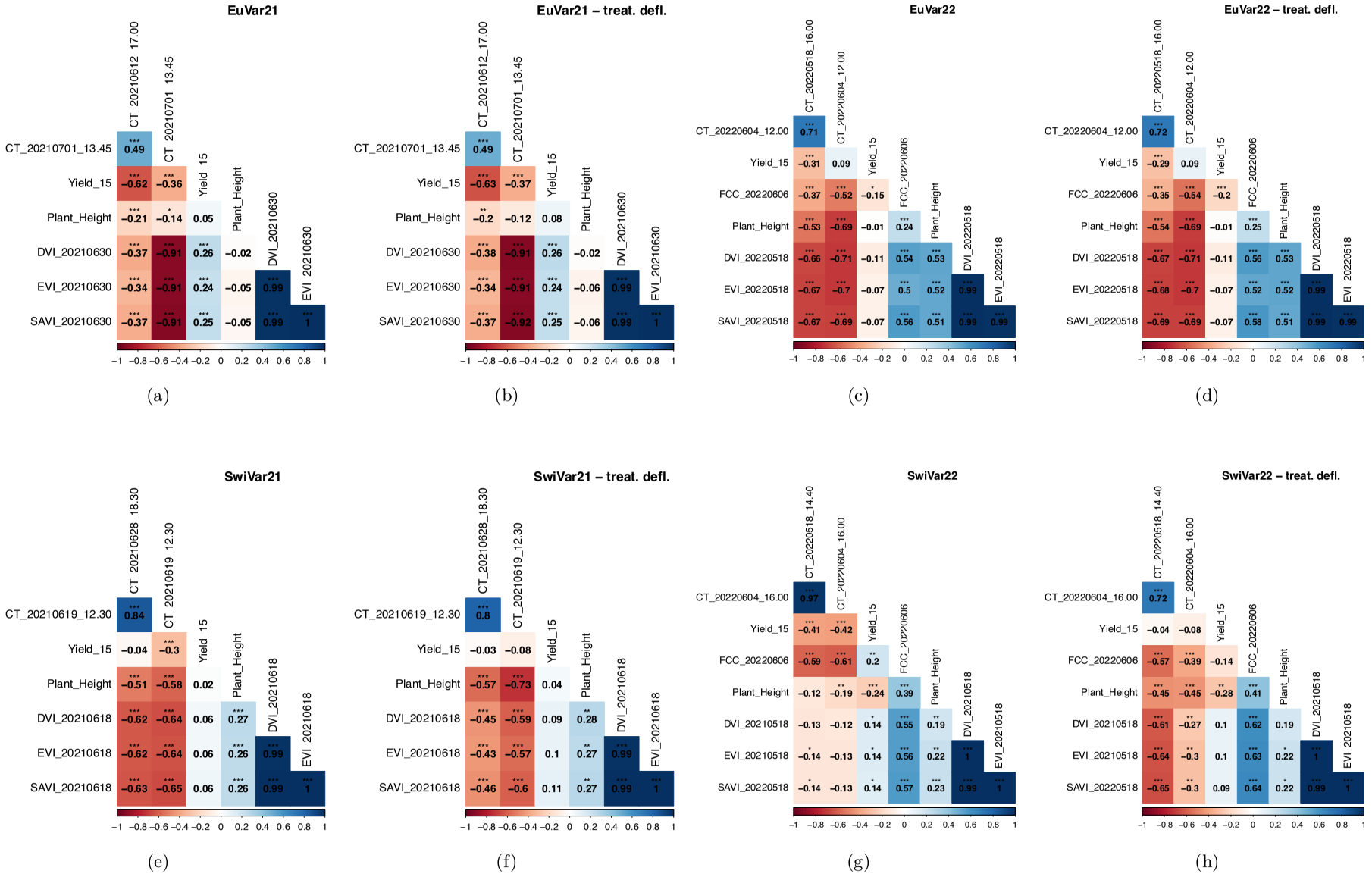


FIGURE 3.5: Pearson correlation of genotypic CT with genotype specific in-field reference measurements for EuVar (a - d) and SwiVar (e - h). For each experiment in each year, CT of two distinct dates was correlated with yield at 15 % gravimetric water content, plant height and three multispectral indices DVI, EVI and SAVI. Correlations were calculated on CT after a spatial and temporal correction according to Eq. 3.5 (a, c, e, g) and after deflating treatment effects on CT estimates as well as on reference measurement by subtracting mean treatment effects (b, d, f, h). Dates and flight times are indicated for CT measurements and dates for multispectral measurements and FCC estimates. FCC estimates were just done in 2022.

3.3.6.4 Correlation between CT and multispectral vegetation indices

VIs were negatively correlated with CT for all trials (Fig. 3.5) and the correlations were highly significant ($p \leq 0.001$) except for SwiVar21 before treatment deflation ($p > 0.01$). Correlations were always higher with the CT measurements taken closer to the date of the VI measurements.

3.3.6.5 Impact of flag leaf rolling on CT

When grouping CT estimates according to flag leaf rolling ratings of the dry year 2022, significant differences of CT could be observed for some flights. For EuVar22 flights on 2022-06-10 at 12:00 (Fig. 3.6a), CT was significantly different between leaf rolling rating groups for CT estimates before and after applying a treatment deflation on CT values ($\hat{\theta}_p^{ts-c}$ & $\hat{\theta}_p^{t-defl}$). For SwiVar flights on 2022-06-10 at 13:00 (Fig. 3.6b), differences were significant before treatment deflation ($\hat{\theta}_p^{ts-c}$) but not after ($\hat{\theta}_p^{t-defl}$) and lower flag leaf rolling ratings were associated with higher temperatures. The differences were only significant after treatment deflation for the flights on 2022-06-17 at 16:40 (Fig. 3.6c) but not before. The differences between the flag leaf rolling rating groups after treatment deflation were generally small ($< 0.40\,\mathrm{K}$). For most other dates, differences were not significant (Figs. S2.32 - S2.34).

3.3.6.6 Impact of senescence on CT

Flag leaf senescence was just rated for the two dates of EuVar21. The senescence ratings for 2021-06-11 were compared with the CT of 2021-06-12 at 17.00 (Fig. 3.6d) but the correlation was not significant. The senescence ratings for 2021-07-02 were strongly correlated ($r = 0.69$, $p \leq 0.001$) with the CT of 2021-07-01 at 13.45 (Fig. 3.6e).

3.3.7 Phenotypic correlations between reference measurements

Correlations between in-field reference measurements with CT were discussed above, yet possible correlations between reference measurements as summarized in Fig. 3.5 must also be considered.

Plant height and yield were never correlated except for weak but significant correlations in SwiVar22 prior to treatment deflation ($p \leq 0.01$).

Yield was only weakly correlated with VIs for EuVar21 and for SwiVar22 before treatment deflation ($p \leq 0.001$), but significant correlations were always weaker than correlations between yield and CT for corresponding dates.

FCC and yield showed a weak but significant correlation ($p \leq 0.001$) in EuVar22 and in SwiVar22 before treatment deflation ($p \leq 0.01$).

3.3.8 Spatial CT trends in the field

Based on estimates of single flights, the spatial field trend estimates $\hat{\theta}_{r_{(p)}c_{(p)}}$ were not consistent. The sign of the correlations between flights changed randomly, (Fig. S2.26 - Fig. S2.31). Spatial trend estimates based on all flights within campaigns appeared random for EuVar21 (Fig. 3.7a, Figs. S2.36a & S2.38) but more consistent for EuVar22, SwiVar21, and SwiVar22. For EuVar22 (Fig. 3.7b, Figs. S2.36b & S2.39) spatial field trends of campaigns were positively correlated except for the campaign on 2022-06-11 at 15:15 and correlations were highly significant. SwiVar21 flights (Fig. 3.7c, Figs. S2.37a & S2.40) showed moderate to very strong correlations within the 2021-06-19 flights. Within 2021-06-28, the correlations were positive and negative, while the positive correlations were stronger and more significant. The correlations between flights on 2021-06-19 and 2021-06-28 were positive for 16 out of 20 correlations and were weak to strong and highly significant in most cases. The four negative correlations were very weak to

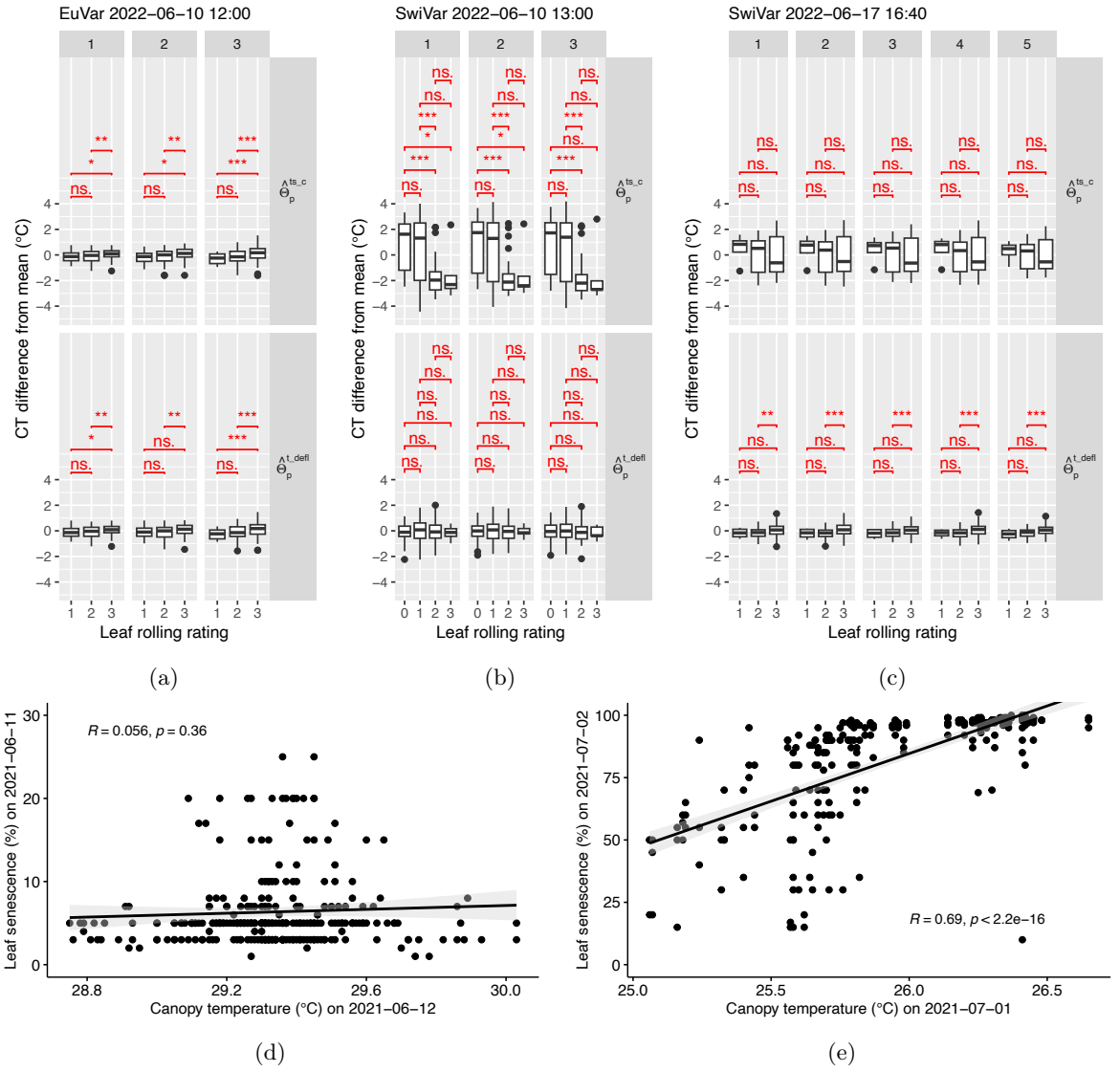


FIGURE 3.6: Impact of flag leaf rolling and senescence on CT. Corrected CT differences from mean were grouped for campaigns on specific dates and flight times by their flag leaf rolling rating for EuVar on 2022-06-10 (a) and SwiVar on 2022-06-10 (b) and on 2022-06-17 (c) before ($\hat{\theta}_p^{ts-c}$) and after ($\hat{\theta}_p^{t-defl}$) applying a treatment deflation on CT estimates. The numbers above the individual columns indicate the flight number of the flights within the campaigns of CT measurements. For EuVar on 2022-06-09 and SwiVar on 2022-06-17, all ratings were larger than 0. Leaf rolling ratings were conducted on the same day as flights or the day before. The significance of differences between groups of leaf rolling ratings was highlighted in red. Significance levels: ns: $p > 0.05$; *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$. Senescence ratings of EuVar21 for 2021-06-11 were compared with the CT of 2021-06-12 at 17.00 (d) and the senescence ratings for 2021-07-02 were compared with the CT of 2021-07-01 at 13.45 (e).

weak and significant at $p \leq 0.001$ just in two cases. Within SwiVar22 (Fig. 3.7d, Figs. S2.37b & S2.41), the correlations ranged from strong to very strong ($p \leq 0.001$) within days and from moderate to strong between different days. Weaker correlations were often not significant at $p \leq 0.05$. Two correlations were negative but significant at $p \leq 0.001$.

The variance of the spatial trend estimates within flights $var(\hat{\theta}_{r_{(p)}c_{(p)}})$ was much stronger in 2022 compared to 2021 for both trials (Fig. 3.7e). The mean $var(\hat{\theta}_{r_{(p)}c_{(p)}})$ was 1.09 K^2 for EuVar21 and increased to 2.87 K^2 for EuVar22. The mean $var(\hat{\theta}_{r_{(p)}c_{(p)}})$ was lower in SwiVar

but also increased from 0.32 K^2 for SwiVar21 to 0.76 K^2 for SwiVar22.

3.3.9 PLSR modeling of TIR residuals to better understand geometric sources of variance of apparent CT

3.3.9.1 TIR residuals and geometric trends

After pre-processing with the mixed model in ASReml, the residuals were analyzed for geometric patterns. Looking, for example, on the residuals of the flight of the SwiVar campaign on 2021-06-19 at 12:30 (Figs. 3.8a - 3.8c), a gradient along the lateral "distance in direction of sowing rows" (Fig. 3.8a) can be seen. The dimensions "distance in direction of sun" (Fig. 3.8b) and "distance on the sensor" (Fig. 3.8c) showed very similar patterns and the main difference was a rotation around the origin of the respective dimensions. For the first flight of the SwiVar campaign on 2022-06-18 at 11:40 (Figs. 3.8d - 3.8f), distinct patterns can be seen with respect to the dimensions "distance in direction of sowing rows" (Fig. 3.8d), "distance in direction of sun" (Fig. 3.8e) and "distance on the sensor" (Fig. 3.8f). The residuals were more positive below the camera and more negative with more oblique viewing geometries and patterns were very similar again between the dimensions with a rotation around the origin. Although these patterns were not always the same between the flights, they were always very similar between the three dimensions of one flight. Also, after vignetting correction, the patterns remained very similar to patterns before vignetting correction (not shown).

The theoretical atmospheric effect was almost zero for a flight height of 40 m (Fig. 3.8g) but became larger at a flight height of 300 m (Fig. 3.8h). The pattern at flight height 300 m was very similar to the geometric trends at 2022-06-18 (Figs. 3.8d - 3.8f) but also to the vignetting effect (Fig. S2.3). While the real atmospheric effect could not be described within this study, this demonstrates the point-symmetric nature of this effect but also its negligible order of magnitude at low flight heights.

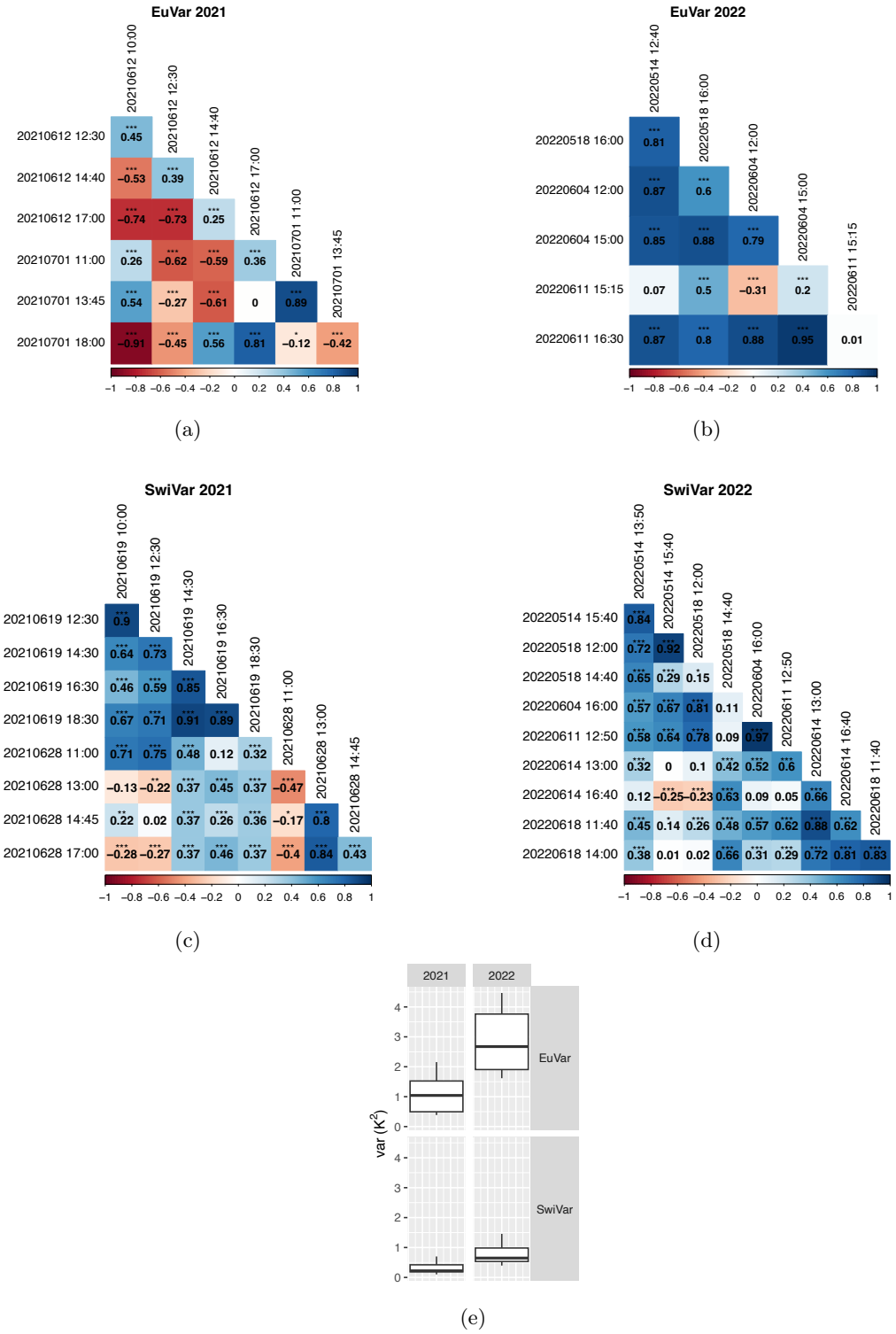
3.3.10 PLSR modeling of geometric CT trends

The residuals e_{ijknp} of the mixed model (Eq. 3.1) were used as input of the PLSR model. Table 3.4 summarizes how much of the variance of e_{ijknp} within single flights could be explained using geometric covariates in PLSR.

Of the 20 initial covariates included in the PLSR models (Table 3.3), 9 were selected in a supervised selection for use in further processing. The relative PLSR coefficient magnitudes β_{i} of the selected covariates are shown in Fig. 3.8i. The four covariates "RowDir-lat", "RowDir-long", "SunDir-lat" and "SunDir-lon" were the most important in PLSR, followed by "Interact.-SunDir-Drone". The absolute values of the four covariates ("RowDir-lat-abs", "RowDir-long-abs", "SunDir-lat-abs" and "SunDir-lon-abs") were less important in PLSR for most flights with values around 0 %. However, for some flights, especially in 2021, they reached values of up to 10 %.

The median values of the explained variance ranged from 20.3 % to 59.2 % when just including 9 covariates and not applying vignetting correction. They were generally highest in SwiVar21 while they were lowest in EuVar21. EuVar22 and SwiVar22 showed intermediate values. When only using 9 instead of 20 covariates, the explained variance was 4.0 % lower on average. The explained variance without ex ante vignetting correction was on average 10.9 % higher compared to data with vignetting correction applied. The differences without and with vignetting correction were greater for SwiVar than for EuVar.

Fig. 3.8j compares the residual variance of mixed models and PLSR to initial variance of CT values and variance of initial CT values corresponds to 100 %. The proportion of variance explained with mixed models was always larger when ex ante vignetting correction was applied,



(a) (b) (c) (d) (e)

FIGURE 3.7: Pearson correlations between estimates of spatial trends $\hat{\theta}_{r_{(p)}c_{(p)}}$ for individual campaigns. Spatial trends were estimated according to Eq. 3.10 for (a) EuVar21, (b) EuVar22, (c) SwiVar21 and (d) SwiVar22. Estimates were based on all flights within individual campaigns. The variance of estimates of spatial trends is summarized in (e). Significance levels: *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$.

while the variance of the initial CT values was very similar (Fig. 3.4a). This holds also true for the variance explained after PLSR but the differences between data with and without vignetting correction became smaller. The mean proportion of residual variance after mixed models ranged from 2.98 % to 9.61 %. After PLSR, the mean proportion of residual variance ranged from 2.46 % to 3.51 %, *i.e.* by combining mixed models and PLSR, 97.54 % to 96.49 % of initial CT variance could be explained on average. Details for the reduction in variance of single flights are shown in Fig. S2.42 - Fig. S2.45.

TABLE 3.4: Explained variance of residuals e_{ijknp} by PLSR fitting after pre-processing with the mixed model (Eq. 3.1). PLSR fitting was done with all 20 lineralized covariates and a reduced set of nine selected covariates. Mean and median values were calculated over all PLSR models of the two experiments EuVar and SwiVar for data with and without vignetting correction (VC) over the two years.

	Number of covariates		**Explained variance of residuals (%)**			
			2021		2022	
			without VC	with VC	without VC	with VC
EuVar	20	mean	30.9	20.8	50.8	42.9
		median	30.4	21.5	47.5	39.2
	9	mean	24.4	17.8	48.6	41.1
		median	20.3	18.6	45.9	37.6
SwiVar	20	mean	62.6	45.3	51.3	40.9
		median	65.3	43.9	56.0	45.6
	9	mean	57.6	41.9	47.9	37.7
		median	59.2	41.1	52.3	42.3

3.3.11 Reference measurement to better understand the sources of variance in apparent CT

3.3.11.1 Fan experiment to determine the influence of wind

The fan experiment showed a strong reaction of the sensor to heating and cooling (Fig. 3.9). The apparent temperature of the PCV sheet dropped immediately by more than 20 °C upon switching on the lamp and rose again to a temperature of about 10 °C below the previous temperature. During the next 15 min, it slowly increased. As soon as the fan was turned off, the temperature rose by more than 30 °C and immediately decreased again and continued to decrease for 5 min until the fan was turned off and the temperature dropped again until the fan was turned on again. The same pattern was repeated three times until the lamp was finally turned off and the temperature stabilized anew. Strong temperature gradients between monitored objects themselves did not cause any drift. The introduction of warm and hot objects did increase the standard deviation of the pixel-wise temperature as long as the objects were within the FOV but did not appear to cause a drift of the apparent temperature or an increased standard deviation for any longer than the period during which disturbance objects were present inside the FOV.

3.3.11.2 Qualitative demonstration of impact of apparent soil cover on CT

For most situations, soil was warmer than the vegetation which was especially evident when looking into the rows perpendicularly (*e.g.* Figs. 3.10a & 3.10b). From an oblique viewing

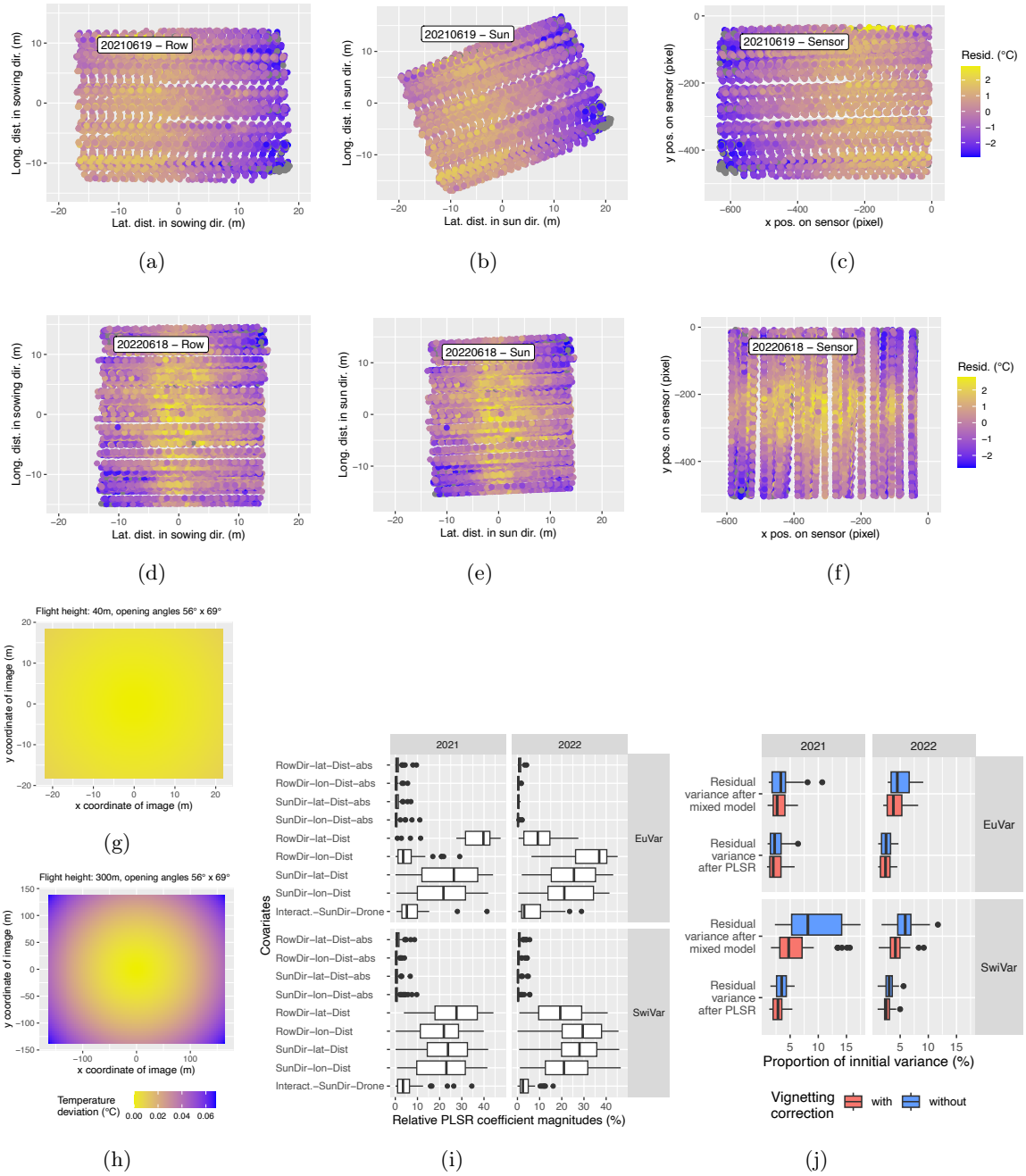
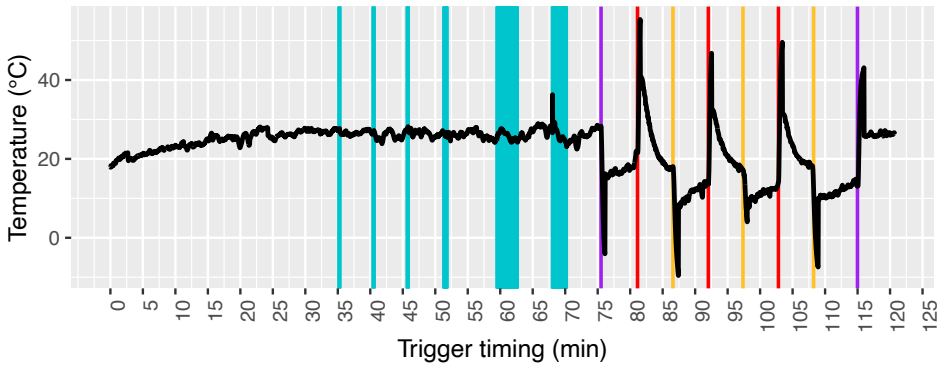
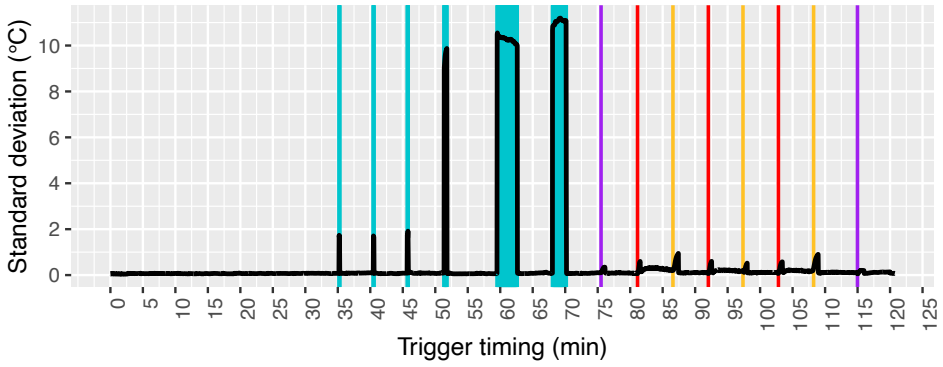


FIGURE 3.8: Geometric trends of CT estimates of first flights of SwiVar campaigns on 2021-06-19 at 12:30 (a - c) and on 2022-06-18 at 11:40 (d - f). CT residuals of the mixed model (Eq. 3.1) are plotted with respect to lateral and longitudinal distance of the plot seen from the drone in sowing row direction (a & d), sun direction (b & e) and the position of the plot center on the focal plane array of the TIR sensor, *i.e.* the x/y coordinates of the thermal images (c & f). A theoretical atmospheric effect is shown for two different flight heights (g) 40 m and (h) 300 m. (i) shows the PLSR coefficients of the 9 selected linearized geometric covariates which indicate the relative importance of the covariates in PLSR modelling to explain the variance of the CT residuals after the mixed models. (j) summarizes the variance after the mixed models (multiple for each plot in each flight) and after PLSR modelling expressed as % of initial variance.



(a)



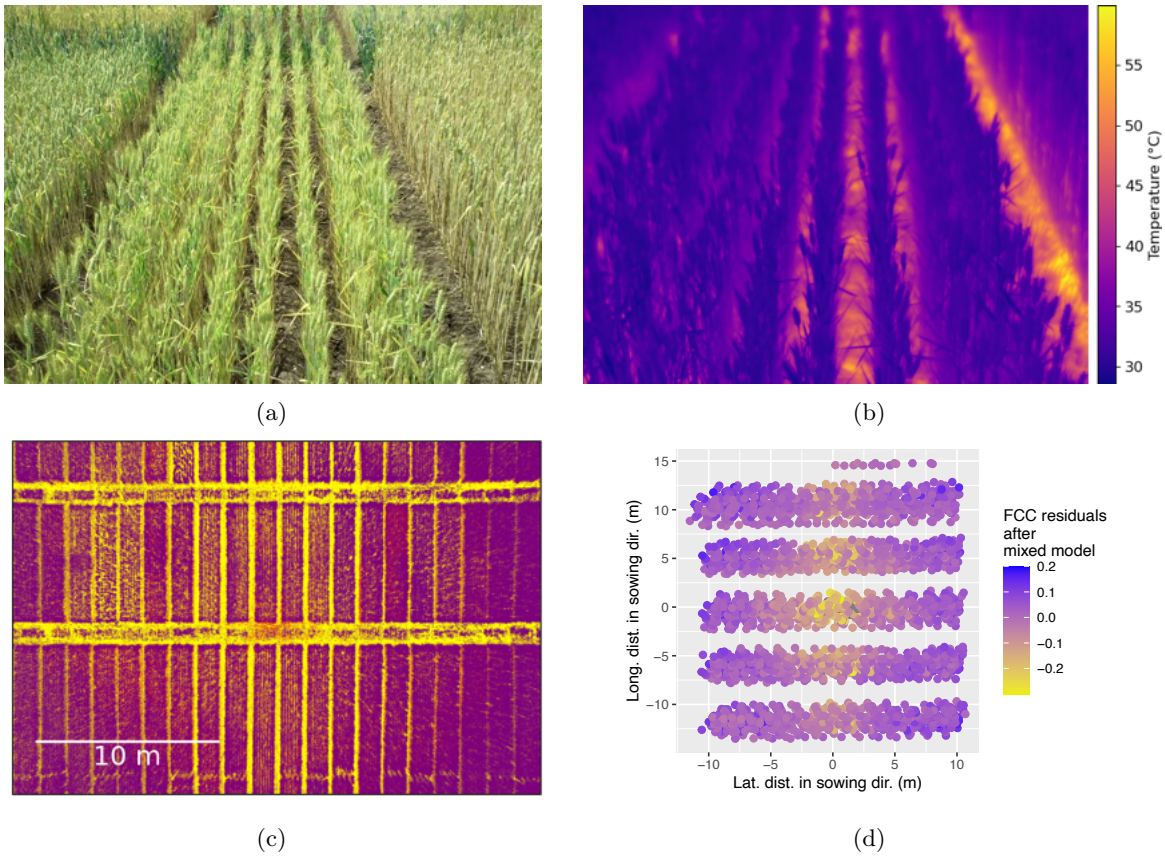
(b)

FIGURE 3.9: TIR drift (a) and standard deviation of pixels-wise temperature on the PVC sheet (b) during the fan experiment. During the stabilization period, warm objects were introduced into the FOV three times (first three vertical blueish shadings), and then hot objects were introduced into the FOV for three times (subsequent three larger shadings). At about 75 min, the heating lamp was turned on (first vertical purple line). The fan was then turned on (red lines) and off (yellow lines) three times before the lamp was turned off (second vertical purple line).

angle, the FCC decreased and so did the average apparent CT in the respective area.

3.3.11.3 Multi-view residuals of FCC from RGB data to demonstrate viewing-geometry dependency of CT

The apparent FCC showed a distinct pattern with a lower apparent FCC in the center and a higher apparent FCC toward the edges of the images (Fig. 3.10c) which is related to the more oblique viewing angles. After fitting the FCC values for design factors in a mixed model (Eq. 3.1, but for CT instead of FCC), the residuals showed a distinct pattern with regard to position relative to row direction (Fig. 3.10d). They were lowest when following a line parallel to row direction directly below the drone (lateral distance in the direction of sowing = 0). When diverging perpendicularly from this line in both directions (*i.e.* with increasing lateral distance perpendicular to the direction of sowing), the residuals became more positive, *i.e.* FCC increased. A similar yet less distinct effect could be observed along this line with increasing residual values when diverging from the position on the soil directly below the drone (with increasing longitudinal distance parallel to direction of sowing). Areas with low FCC coincided with warm areas, and spatial trends were often very similar between the two traits (cf. Fig. 3.10d and Figs. 3.8d - 3.8f).



(a) (b) (c) (d)

FIGURE 3.10: FCC trends in relation to viewing geometry: The same scenery is shown on an RGB image (a) and a TIR image (b). This shows how the soil is warmer than the plants. To demonstrate how the apparent fractional canopy cover (FCC) changes with viewing geometry, RGB images were labeled in Ilastik software to segment images into plant (purple) and background (yellow) (c). The resulting images were analyzed by the multi-view method to get FCC for each plot in each image. The FCC values were fitted with a mixed model (Eq. 3.1, but for CT instead of FCC) for design factors. The residuals of the model are shown in (d) in relation to the position of the plot relative to the sowing row direction for SwiVar22.

3.4 Discussion

This study used the multi-view approach (Treier et al., 2024) to discuss the manifold sources of variance in airborne thermal imaging, based on data from two very different wheat variety testing trials followed over two seasons, characterized by very contrasting meteorological conditions. The discussion of the different sources is structured according to the primary type of correction (Table 3.1).

3.4.1 Temporal correction of CT

Temporal trends contributed the most to the total variance of CT estimates. Fig. 3.3a illustrated the magnitude of temporal trends, which can be several times larger than genotype-specific differences (*e.g.* Kelly et al., 2019; Treier et al., 2024; Z. Wang et al., 2023). Temporal correction reduced the variance of CT estimates the most (Fig. 3.4a) which is in line with Z. Wang et al. (2023). This demonstrates the importance of proper handling of temporal trends in thermal measurements, as has been highlighted in several publications (*e.g.* Kelly et al., 2019; Malbéteau et al., 2021; Treier et al., 2024; Z. Wang et al., 2023; Yuan and Hua, 2022). Z. Wang et al. (2023) elaborated on the distinction between thermal drift and the

temporal variation of land surface temperature (LST). Thermal drift is caused by the thermal camera when the temperatures of FPA, lens, and camera body change. The wind on the sensor cools them and exposure to sunlight as well as the sensor's electronic heats them, leading to fluctuation temperature readings even when facing toward a target with an actual constant temperature (*e.g.* Aragon et al., 2020; Kelly et al., 2019; Messina and Modica, 2020; Treier et al., 2024; Z. Wang et al., 2023). This interaction was confirmed for the sensor used in this study with a fan experiment (Fig. 3.9). In accordance with findings in Kelly et al. (2019), the warming of the sensor led to a decrease in the apparent temperature of the target and vice versa. As internal processes of TIR cameras are proprietary information of the manufacturers, the reasons for this are difficult to determine (Budzier and Gerlach, 2015; Kelly et al., 2019). The thermal signal reacted within seconds after a change of wind conditions (fan) or thermal radiation (heating lamp). In contrast to thermal drift, temporal variation corresponds to actual changes in the temperature of a given target that can be caused by wind, changing air temperature, VPD, solar illumination, changing water status of the plant and the plants physiological response to such changes (*e.g.* Idso et al., 1981; Perich et al., 2020; Rebetzke et al., 2013; M. P. Reynolds, Pask, et al., 2012; Z. Wang et al., 2023). The impact of temporal variation was reduced in this study by flying in weather conditions that were rather stable throughout single flights (Kelly et al. (2019); Figs. S2.6 & S2.7). Nevertheless, also in stable conditions, LST changes, but these changes are comparably slow and if measurements are taken within a short interval, *e.g.*, within 30 min, the temporal variation in LST is relatively low (Z. Wang et al., 2023). A typical flight time in this study was 7 to 9 min, a 3 flight campaign lasted about 25 min, and therefore a large proportion of temporal trends can be assumed to be thermal drift, and temporal variation contributed relatively little to total variance of CT estimates.

3.4.2 Spatial correction of CT

Thermal imaging was proposed to estimate spatial field heterogeneity caused, for example, by variability of soils, soil water content, or soil-borne pathogens, and to improve the interpretability of other phenotypic measurements (*e.g.* Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Messina and Modica, 2020). In contrast to hand-held infrared thermometers, many experimental plots and larger areas can be measured simultaneously and repeatedly in a short period with airborne thermography. Handheld infrared thermometers are also prone to thermal drift, but with just one measurement taken at a time, the temporal and spatial trends are challenging to separate from each other in a statistical analysis (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016). Revisiting the same spot multiple times in a short interval (*e.g.* 30 min) improves the estimation of the real relative temperature of the spot and thus of spatial trends, since the temporal variation of the CT can be assumed to be relatively small and temporal trends are mainly thermal drift (Z. Wang et al., 2023). When working with uncorrected images in orthomosaic approaches, each plot is measured multiple times. The temporal and spatial effects are reduced by leveling them out in orthomosaic blending. Perich et al. (2020) accounted for the remaining temporal and spatial variance together in a mixed model, and they stated that it remains challenging to unravel the two. Multi-view offers an opportunity to alleviate this limitation as measurements are analyzed individually (Treier et al., 2024), but as shown in this study, a spatial trend estimate based on a single flight showed little reliability (Fig. S2.28 - Fig. S2.31). As claimed by Z. Wang et al. (2023), increasing the number of observations per plot led to a more consistent estimation of the spatial trends (Fig. 3.7a - Fig. 3.7d). To further improve the estimation of spatial trends, it is proposed to conduct at least two flights over the field with different flight

paths, orthogonal to each other. This reduces the probability of artifacts due to the repeated occurrence of similar temporal patterns when following the same flight plan.

In 2022, the spatial trend $\hat{\theta}_{r_{(p)}c_{(p)}}$ was more pronounced than in 2021. Thus, the variance of the spatial trend was greater in 2022 than in 2021 (Fig. 3.7e), indicating a stronger expression of the spatial trend. 2021 was a wet year and a sufficient water supply can be assumed throughout the growing season. The spatial trends were therefore relatively weak. Such weak trends are more difficult to reproduce, as little differences in the estimation lead to different trends. In such homogeneous conditions, the multi-view approach might fail to detect the weak spatial trends reliably. At the same time, the correct estimation of weaker trends is also less important because their impact on final results becomes negligible. 2022 was hot and dry, and spatial trends in water status were observed in the field. A more pronounced spatial trend can be estimated more easily and reliably.

However, simultaneous accounting for temporal and spatial trends was shown to lead to highly consistent CT estimates even when based only on a single flight (Treier et al., 2024).

3.4.3 CT variance reduction by temporal and spatial trends

After the temporal and spatial correction, the experimental effects remained, *i.e.*, the effects of genotypes and treatments, as well as the effects of viewing geometry.

The added variance of these effects was much smaller than the initial variance of the temperature estimates, with the exception of SwiVar22, which showed a strong treatment effect (Fig. 3.4a). Agricultural research is usually interested in the effects of genotypes and treatments. This shows the importance of reducing the effects of unwanted sources of variance. Only through the appropriate consideration of large confounding influences, more subtle effects actually under observation within an experimental setup can truthfully be estimated (Damm et al., 2022).

3.4.4 Correlation of CT with in-field reference measurements of phenotypic traits

In accordance with the literature (Rebetzke et al., 2013; M. P. Reynolds, Pask, et al., 2012), yield and CT were negatively correlated in conditions without water limitation. This was only the case in 2021 as 2022 was a hot and dry season. The correlations were stronger and more significant in EuVar21 compared to SwiVar21. While the effect of fertilizer application in SwiVar21 was rather small, it appeared to be the main driver of the correlation between corrected CT and yield. In the EuVar trial, a relatively diverse set of European genotypes was tested, while in SwiVar, varieties of the Swiss variety list and candidates for registration in the variety list were tested. It can be assumed that the phenotypic variability between the varieties was greater in EuVar than in SwiVar. With more pronounced differences between estimates, stronger correlations are more easily achieved.

There was a consistently negative correlation between CT and plant height. This could in part be caused by effects related to canopy architecture, *e.g.* increased LAI and a stronger exposure to wind (*e.g.* Z. Wang et al., 2023), but also by genetic co-locations of quantitative trait loci for CT and plant height (*e.g.* Rebetzke et al., 2013). The correlations were more pronounced in SwiVar after treatment deflation, indicating a masking effect of fertilizer treatment on the genotypic correlation between CT and plant height.

Although just measured in 2022, the trends for FCC were similar to those of plant height, with stronger correlations after treatment deflation. The constant correlation between FCC and plant height also indicates that they can be interlinked. Furthermore, with decreasing FCC, the effect of mixed pixels can be expected to increase, especially if the GSD is larger

than the size of the plant organs (H. Jones and Sirault, 2014), shifting the CT estimate toward the temperature of the soil background.

The flag leaf rolling is a protective mechanism of wheat to reduce transpiration losses. It reduces the amount of incident radiation intercepted by the plant and traps air within the leaf, reducing the VPD at the border layer (Pask et al., 2012). It was used as an indicator of the level of drought and heat stress to which the wheat was exposed. Although CT differences between groups of different flag leaf rolling ratings were significant for some dates, these differences remained relatively small ($< 0.40\,\mathrm{K}$) after treatment deflation. Differences before treatment deflation were large for SwiVar on 2022-06-10 (Fig. 3.6b) and lower flag leaf rolling ratings were associated with higher CT estimates, which is counter-intuitive. To understand this, the interaction between CT estimates, water use, and above-ground biomass must be analyzed. In 2022, it was evident from field observations that the above-ground biomass in the unfertilized part of SwiVar was much lower than in the fertilized part. The lower biomass was confirmed by reference measurements, as FCC but also multispectral indices that approximated above-ground biomass and LAI were lower in the unfertilized part (Fig. S2.24). At the same time, flag leaves expressed stronger rolling in the fertilized part compared to the unfertilized part (Fig. S2.35), indicating the plants experienced a stronger water deficit in the fertilized part (Pask et al., 2012). The lower biomass presumably led to a lower total transpiration in the unfertilized part and saved soil water, which in turn allowed plants to maintain unrolled leaves longer into the season compared to the fertilized part, where available water was exhausted earlier. This illustrates well the complex interactions between phenotypes, water status, transpiration, and CT. At the same time, this highlights the importance of environments for the contextualization of the expression of CT as a trait. In 2021 almost the same set of genotypes was sown as in 2022 and the treatments were identical, but led to a much more pronounced treatment effect in 2022 with lower FCC, above-ground biomass and LAI.

The correlation between CT estimates and reference measurements was strongest between CT and multispectral vegetation indices (Fig. 3.5). This correlation was strongest in EuVar21, when the correlations between CT and yield were also strongest. CT was often negatively correlated with yield and plant height, but yield and plant height were not correlated except for a weak correlation in SwiVar22. The impact of the treatments on correlations was small for EuVar21, EuVar22 and SwiVar21. These results support the findings of Pask et al. (2012), Rebetzke et al. (2013) and Roche (2015), that CT and yield are especially correlated when conditions are not water limited.

Multiple sources of phenotypic variability, genetic or related to treatment, are associated with CT, and this must be taken into account in the analysis (Maes and Steppe, 2012; Rebetzke et al., 2013). It was demonstrated how correlations can be driven or masked by the treatment effect. For example, yield was only correlated with CT before treatment deflation in SwiVar22 but the genotypic correlation between plant height and CT only became evident after treatment deflation. The correlation of CT with plant height and FCC was consistently stronger than the correlation with yield, except for EuVar21. This might indicate that a plant height effect was masking the yield effect on CT in many cases. It remains unclear why plant height and CT showed a weaker correlation in EuVar21 but the FCC measurement of 2022 were consistently correlated with plant height. FCC was not estimated in 2021 but canopies were observed to be very dense in this season. This might have led to saturated FCC with values near 1 (*i.e.* 100 % canopy cover), which might have reduced the effect of plant height on CT, unmasking the correlation between CT and yield.

Although FCC and CT were correlated, the treatment effect of FCC in SwiVar22 was not as evident as for CT (cf. Fig. S2.24g and Fig. S2.15). This was possibly caused by saturation of the FCC where the canopy appears largely closed even on the unfertilized part, with an

FCC near 1, but is still less dense than the canopy of the fertilized part. Through the less dense canopy, the soil background could have a larger impact on CT (Das, S. C. Chapman, et al., 2021; Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Pask et al., 2012), making the nadir-oriented measurements appear hotter compared to the more oblique measurements (Perich et al., 2020). The interactions of CT and soil-background can change with increasing temperatures throughout the day. The soil may be cooler than the plant in the morning and warmer later in the day (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016).

For the second EuVar21 flight date on 2021-07-01, senescence had progressed for some genotypes while it was still in early stages for other genotypes (Fig. 3.6e). The strong correlation between senescence ratings and CT underlines the importance of considering phenology in the timing of CT estimates (Anderegg, Kirchgessner, et al., 2024; Lopes and M. P. Reynolds, 2010; Rebetzke et al., 2013). However, the 2021 season was characterized by frequent precipitation, and days with optimal conditions for CT estimates (no clouds, little wind) were rare. For logistical reasons, it was therefore not possible to conduct the second measurement day earlier and with a less pronounced senescence. Such meteorological and logistical constraints avoiding optimal measurement timing are a common problem in agricultural research, breeding, and variety testing. However, CT measurements during intermediate leaf senescence stages also showed similar correlation patterns, notably with yield, and with measurements taken earlier in the season (Treier et al., 2024). Although measurements taken at the same phenological stage are optimal, this is indicating that conclusions drawn from CT show a certain robustness, even when the sample population shows some phenological heterogeneity, *e.g.* in cases where measurement before onset of senescence is not possible.

The correlation with yield was always stronger for CT than for the multispectral indices (DVI, EVI, SAVI). Now, the indices were chosen as approximate measurements of above-ground biomass and LAI and not yield. In addition, correlation with NDVI was not shown, yet NDVI was closely associated with the indices used (Figs. S2.21 - S2.24). Nevertheless, this underscores the potential of airborne CT for yield prediction in remote sensing, also for temperate climates.

3.4.5 Estimating geometric effects by PLSR modeling

Geometric effects on CT within one image were cited to be as high as 3.5 °C (Perich et al., 2020) and the range of residual values with geometric patterns were larger than 4 °C in this study (Fig. 3.8). The geometric patterns of the residuals were in some cases point-symmetric (*e.g.* Fig. 3.8d - 3.8f) and sometimes looked similar to those of vignetting (Fig. S2.3) and path-length dependent atmospheric effects (Fig. 3.8h) or FCC (Fig. 3.10d). These three effects were very similar in shape and they are all possible causes for these patterns, however, they cannot be disentangled further with this method. The causes and effects of vignetting are well presented in literature (*e.g.* Aasen, Honkavaara, et al., 2018; Yuan and Hua, 2022; Kelly et al., 2019). Atmospheric effects might be negligible when flown at low altitudes (Künzer and Dech, 2013; Messina and Modica, 2020), however, at higher altitudes they might become important (Jimenez-Berni, P. J. Zarco-Tejada, et al., 2009), as shown in Figs. 3.8g & 3.8h. This study assumed an oversimplified length-dependent model. For higher altitudes, the attenuation could be estimated based on MODTRAN radiative transfer models (Jimenez-Berni, P. J. Zarco-Tejada, et al., 2009; Jimenez-Berni, P. Zarco-Tejada, et al., 2009; Maes, Pashuysen, et al., 2011; Maes, Huete, et al., 2017). In addition to flight height, the strength of the atmospheric effect on the measured temperature depends primarily on atmospheric pressure, air temperature, and humidity (Meier et al., 2011; Schläpfer et al., 2022). FCC residuals showed a similar spatial pattern as CT residuals after processing with a mixed model (Eq. 3.1) and it is likely that FCC also contributed to CT variance, where CT associated with a lower FCC appeared higher. FCC therefore affected the genotypic variability of CT as was shown with correlation between

CT and FCC, but also the residual FCC pattern. Geometric effects on CT can be expected to be more pronounced for canopies with lower FCC, as their apparent FCC changes from low to almost closed canopy for oblique viewing geometries. In contrast, for almost closed canopies with an almost saturated FCC towards 1, this change is very limited (Fig. 3.10c). Like plant height, FCC is a structural trait of the wheat canopy, and structural traits interact with CT. Other structural traits not considered in this study but with a potential impact on CT include LAI or leaf angle (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Maes and Steppe, 2012; P. Zarco-Tejada, González-Dugo, L. Williams, et al., 2013).

Often, the residuals also contained more axisymmetric and continuous trends. Such trends could be caused by BRDF or unilaterally warmed spikes. However, such trends usually feature a gradient parallel to the principal plane of the sun (Bai et al., 2023; Perich et al., 2020). This was not always the case (see, *e.g.*, Fig. 3.8a - 3.8c). This could possibly be caused by an interaction of sowing row direction and incident sunlight, where the spacing between the sowing rows allows light to penetrate the canopy and warm the plant from one side, but not from the other (*e.g.* H. G. Jones, Stoll, et al., 2002). Another possible explanation is the camera orientation not being perfectly nadir. With a slightly tilted camera, some geometric effects would still be concentric with the image center (*e.g.* vignetting), while other effects like FCC would not align with the center of the image anymore. The concentric and eccentric patterns would then combine into a less point-symmetric pattern with a more continuous appearance.

In PLSR modeling, covariates without absolute value transformation are better suited to describe continuous effects, while covariates after absolute value transformation rather correspond to point-symmetric effects. Based on PLSR coefficient magnitudes, continuous effects (initial covariates without absolute value transformation) were generally more important in explaining residual variance than point-symmetric effects (absolute covariate values) and PLSR coefficient magnitudes for point-symmetric effects were close to zero for most cases (Fig. 3.8i).

PLSR modeling allowed the explanation of a significant proportion of residual variance (Table 3.4) as geometric effects. It should be noted that the proportion of variance that can be explained by PLSR also depends on the magnitude of the initial variance. However, in this study no clear correlation between initial variance and variance explained by PLSR could be shown. Yet, it is hypothesized that the relatively large proportion of residual variance explained by PLSR in SwiVar2021 was due to the relatively low overall variance in this trial, which increased the proportion of residual variance in overall variance. The proportion of explained variance was consistently greater on data without vignetting correction, indicating that vignetting correction and PLSR were reducing initial variance of the same geometric dimensions, *i.e.* PLSR was also modeling vignetting. The proportion of variance that was accounted for by vignetting correction could therefore not be explained by PLSR, which was decreasing the proportion of residual variance explainable by PLSR.

However, the contribution of residual variance to total variance was relatively small (Fig. 3.8j). Especially point-symmetric effects such as vignetting and FCC seemed to have little impact on total variance, as demonstrated by the low importance of absolute coefficient values in PLSR modeling. The relatively small impact of vignetting correction was also supported by the low difference of the proportion of residual variance explained by PLSR between the data with and without vignetting correction. This difference in explainable residual variance was only 10.9 % and it is hypothesized that this percentage is also an approximation of the total importance of vignetting correction. Furthermore, vignetting correction had little impact on total variance (Fig. 3.4a) but also on the correlation of CT with other phenotypic traits (Figs. 3.4b - Fig. 3.4d). The contribution of residual variance to total variance might

vary depending on the cropping system under observation. A row crop with a larger inter-row spacing or a poor plant development associated with a lower FCC might feature more pronounced FCC patterns and therefore stronger geometric trends of CT. Kelly et al. (2019) and Perich et al. (2020) report that such geometric effects are more important when analyzing CT based on single images. When CT analysis uses multi-view or orthomosaics, plot estimates are based on multiple images or selected for most nadir-oriented views, both reducing the geometric impact on plot-wise estimates.

3.4.6 Unexplained residual variance of CT

The sequential application of mixed models and PLSR models could explain a large proportion of variance. But there will always remain unexplained residual variance and though the contribution of residual variance to total variance might be negligible, some possible causes of residual variance are mentioned in the following. Residual variance could be caused by non-geometric non-uniformity effects that neither the vignetting correction nor the PLSR could account for. Also, non-continuous effects impacting CT like temporal CT inconsistencies due to gusts might not be accounted for as well as the sensor noise beyond thermal drift, *i.e.*, dark signal noise (Aasen, Honkavaara, et al., 2018). The canopy may also feature holes, caused, for example, by heterogeneous emergence, damage from rodents, or previous sampling events, which could have different impacts on CT estimates depending on viewing geometry (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016).

3.4.7 Emissivity and CT variance

An important determinant of CT variance that is often ignored in airborne thermography of crops is emissivity. Emissivity compares the TIR radiation emitted by a surface with the TIR radiation emitted by a black body at the same temperature (Fuchs and Tanner, 1966; Jacob et al., 2004; Messina and Modica, 2020). Two objects of different materials can have the same temperature, but if they have different emissivities, they appear to have different temperatures in thermal images. Messina and Modica (2020) summarizes multiple factors that influence emissivity: color, chemical composition, surface roughness, moisture content, field of view, viewing angle, spectral wavelength, etc. (Campbell and Wynne, 2011; Jacob et al., 2004; J. Jensen, 2009). The emissivities cited in the literature vary, but in general, for healthy leafy vegetation, an emissivity of 0.99 can be assumed (Diaz et al., 2021), where for dry vegetation, emissivity from 0.88 to 0.94 were reported. Water has an emissivity of 0.99 and dry soil an emissivity of around 0.92 (Jacob et al., 2004; Lillesand et al., 2015; Meier et al., 2011). Stressed vegetation generally has a lower emissivity than healthy vegetation, and plant emissivity is highly sensitive to water content (Chandel et al., 2022). Diaz et al. (2021) assumed an emissivity of 0.99 when the NDVI of the respective pixel was above 0.5. NDVI was below 0.5 for some measurements at the last measurement date of EuVar21 (Fig. S2.21c) and SwiVar21 (Fig. S2.23c). Therefore, different emissivities would have to be assumed for different plots of the same measurement flight. This would come with the necessity of estimating the correct emissivity for the specific plot, which can lead to large differences in CT estimates. For example, when an object has a temperature of 20 °C, under the assumption of an emissivity of 0.99, it would appear to be 293.15 K*0.99 = 290.22 K or 17.07 °C. Assuming an emissivity of 0.98, the apparent temperature would be 293.15 K*0.98 = 287.29 K or 14.14 °C. The difference between the two emissivity assumptions of only 0.01 corresponds to 2.93 °C, which is about the range of genotype-specific differences in the experiments of this study and is therefore far too large to study genotype-specific differences of CT.

To avoid the introduction of errors by estimating erroneous emissivity values for individual plots, it might thus be more appropriate to assume a constant emissivity for all measurements,

when no absolute CT values are needed. The absolute value of CT is particularly important for physiological investigations, where absolute values are needed to approximate physiological quantities such as transpiration rate or gas exchange. If, on the other hand, relative CT is compared, the absolute value plays a lesser role. For this study, for example, an emissivity of 1 was assumed. A stressed vegetation, in most situations, would have a higher CT and at the same time a lower emissivity. The effect of assuming a too high emissivity would thus lead to a too low estimate of temperature on the thermal image, and the question remains whether differences of apparent CT on thermal images arise from differences in CT or from a varying emissivity.

In addition, emissivity might also be affected by FCC and LAI. The emissivity of soil can be significantly lower than the emissivity of healthy vegetation, and low FCC, or low LAI, even at a relatively high FCC, might impact the emissivity of a plot, biasing the CT estimates. Cheng and S. Dong (2024) demonstrated for satellite data that the error of emissivity estimates is lower when the emissivity of the soil background is closer to the emissivity of the vegetation, and when the LAI of the vegetation is higher. Sobrino et al. (2005) explored the dependence between emissivity and viewing angle and described that the level of the angular dependency is related to LAI.

However, measuring emissivity in the field is a very tedious task that cannot be easily implemented (Almawazreh et al., 2025). It must be measured at night (Sugita et al., 1996), or by shielding the vegetation with boxes to exclude environmental radiation from the surroundings (Rubio et al., 1997). Thus, in many field studies, the emissivity is ignored (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016; Deery, Rebetzke, Jimenez-Berni, Bovill, et al., 2019; Perich et al., 2020; Anderegg, Aasen, et al., 2021) while other assume a fixed emissivity (often 1), as in this study (Al Masri et al., 2017; Mahlein et al., 2019; Almawazreh et al., 2025).

For satellite-based estimates of LST, model-based approaches to determine emissivity were proposed (Sobrino et al., 2005; Meng et al., 2017; Cheng and S. Dong, 2024), *e.g.* based on NDVI estimates. To the best of the authors knowledge, there are no similar studies for drone-based CT estimates. Yet, the study of Treier et al. (2024) provides the tool to estimate CT in dependence of viewing geometry. In addition, Roth, Aasen, et al. (2018) used the multi-view approach to determine the LAI of soybean. These two approaches could be combined with emissivity estimates to promote a more robust understanding of the interaction of CT, emissivity, viewing geometry, and LAI.

3.5 Conclusions

Canopy temperature is affected by manifold sources of variance which interact with each other. Multiple sources of variances were reviewed based on extensive field data and by using the previously suggested multi-view approach in this study. Experimental sources of variance (genotypes and treatments) were impacted by meteorological conditions in the growing season. To reveal the relation between CT and other traits, corrections for confounding sources of variance (*e.g.* thermal drift, spatial trends, geometric effects) were applied. Temporal trends were consistently the most important confounding source of variance, followed by spatial trends. Estimation of spatial trends and their disentanglement from temporal trends remain a challenge, but a path to improved estimation of the spatial trends by flying multiple times with different flight paths was proposed. Phenotypic relationships can be masked or result from artifacts of random but concurrent instantaneous trends. After correction for disturbing trends, the correlation between phenotypic traits was accentuated. Not applying such corrections might thus entail misleading conclusions on phenotypic relationships with CT. Plant height and FCC were shown to be important phenotypic drivers of CT in many situations and were more correlated with CT than yield, except for well-watered conditions and a diverse set of

genotypes. However, CT was constantly more correlated with yield than multispectral proxy measurements of above-ground biomass and LAI. Although other CIs may be better suited to estimate yield, this highlights the potential of CT to enhance in-season yield estimates in temperate climates, for example, to avoid losing all the information of an experiment due to a hail storm close to harvest. Flag leaf rolling had a relatively small but significant impact on CT. Complex interactions of above-ground biomass, flag leaf rolling as drought symptom, water use by the canopy, and CT were demonstrated. Treatment effects can be considerable and modify other phenotypic traits and their interaction with CT. Geometric trends were shown to have distinct patterns for flights and campaigns, but they explained a relatively low proportion of total variance. Temporal, spatial, genotypic, treatment related and geometric effects together explained the largest part of the initial variance, leaving just a small proportion unexplained. It is hypothesized that many insights on the sources of variance of uncalibrated airborne thermography that were gained in this study are transferable to other crops and other climatic conditions (especially hotter). In cooler conditions, the correlation between CT and yield might be limited due to lower transpirational demands of the plants, leading to lower genotype specific differences of CT. As the study was conducted with wheat, a row crop with relatively large inter-row spaces, following the rationales outlined in this study should also lead to meaningful results in the analysis of other crops with low FCC. At the same time the rather ephemeral character of CT and its strong interaction with the environment should always be kept in mind, as they entail a limited transferability of CT information between different environments. Nevertheless, within the different environments in this study, multi-view thermography served as a means to foster a comprehensive and empirically backed understanding of variance components in drone-based CT estimates. This facilitates the planning, conduct, and interpretation of drone-based CT screenings in variety testing and breeding.

Data availability

The data will be made available upon reasonable request.

Authors' contribution

Simon Treier: conceptualization, methodology, software, formal analysis, visualization, writing – original draft. Lukas Roth: conceptualization, supervision, methodology, review & editing. Juan M. Herrera: project administration, funding acquisition, conceptualization, supervision, methodology, acquisition, writing – review & editing. Achim Walter, Nicolas Vuille-dit-Bille, Lilia Levy Häner, Helge Aasen, Andreas Hund, : writing – review & editing.

Funding

This study was financed by Agroscope and the work of Simon Treier was in part supported by the two H2020 projects InnoVar and Invite.

Acknowledgments

We thank Johanna Antretter, Fernanda Arelmann Steinbrecher, Ulysse Schaller, Matthias Schmid and Julien Vaudroz for rating of phenology; Nicolas Widmer and his team as well as Yann Imhoff for field management; Margot Visse-Mansiaux for support in setting up the experiments.

4 Comparison of PhenoCams and drones for lean phenotyping of phenology and senescence of wheat genotypes in variety testing

Simon Treier[1,2], Nicolas Vuille-dit-Bille[1], Margot Visse-Mansiaux[1], Frank Liebisch[3], Helge Aasen[4], Lukas Roth[2], Achim Walter[2], Juan M. Herrera[1]

1 – Production Technology & Cropping Systems Group, Agroscope, Route de Duiller 60, 1260 Nyon, Switzerland

2 – ETH Zürich, Institute of Agricultural Sciences, Universitätstrasse 2, 8092 Zürich, Switzerland

3 – Water Protection and Substance Flows Group, Agroecology and Environment Division, Agroscope, Reckenholzstrasse 191, 8046 Zürich, Switzerland

4 – Earth Observation of Agroecosystems Team, Agroecology and Environment Division, Agroscope, Reckenholzstrasse 191, 8046 Zürich, Switzerland

After reviewers' feedback, this chapter is currently under revision for a resubmission to The Plant Phenome Journal.

Abstract

In variety testing and breeding of wheat (*Triticum aestivum* L.), it is crucial to know the timing of phenological stages and the senescence behavior of genotypes to select for locally adapted varieties. Knowing the timing of phenological stages also allows for a more meaningful interpretation of measurements such as yield, quality or disease ratings. In the presence of stresses, only a combined characterization of phenology and environmental conditions will permit to unravel stress resistance and stress avoidance. Capturing these traits visually in the field is very time-consuming. Here, a semimobile PhenoCam setup was used to track phenology and senescence from ear emergence to full maturity. PhenoCams mounted on field masts took images of wheat plot trials on a daily basis. In a partial least squares regression (PLSR) analysis, temporal features of multiple vegetation indices were combined in one model to track phenology and senescence. The method was compared with visual reference methods and repeated drone flights with a multispectral camera. Achieved Pearson's correlation between visual reference methods and PhenoCam predictions was stronger than 0.8, often stronger than 0.9, for most stages. An economic analysis showed that PhenoCams are economically interesting, especially to observe remote experimental sites. Thus, PhenoCams offer a cost-effective replacement of visual ratings of phenology and senescence, in the context of multi-environment trials.

4.1 Introduction

In variety testing, breeding, and research of bread wheat genotypes (*Triticum aestivum* L.), it is crucial to know the timing of phenological stages and the senescence behavior of the individual genotypes:

The temporal characterization of plant development allows selecting genotypes better adapted to local climates and soils. For example, early flowering behavior allows plants to escape early summer drought and heat stress during the sensitive stage around meiosis (*e.g.* Rogger et al., 2021), whereas later flowering behavior allows plants to escape late frosts around meiosis (Langer et al., 2014). As the climatic conditions in Central Europe are changing, the use of adapted wheat genotypes can be a strategy to mitigate adverse effects on yield and reduce production risks (Holzkämper et al., 2015; Rogger et al., 2021). According to Asseng, Ewert, Rosenzweig, et al. (2013), wheat yield is more prone to uncertainty with increasing levels of CO_2 and temperature. For every 1 °C increase in temperature above the temperature optimum, there is an estimated decrease in wheat yield of 6 %, and yield becomes more variable in space and time (Asseng, Ewert, Martre, et al., 2015). Consequently, a diversity of wheat germplasm must be maintained and developed to provide a diverse set of adaption strategies to grow wheat in future climate conditions (Kahiluoto et al., 2019).

In field experiments, genotypes with a different phenological development may not be exposed to the same stresses in the same year. Knowing the timing of phenological stages thus allows for a more meaningful interpretation of other measurements such as yield. For example, low radiation at jointing, booting (Jia et al., 2021), young microspore stage (10 - 12 days before heading, H. Yang et al., 2020), ear emergence (Welbank et al., 1968), anthesis (Ford and Thorne, 1975), and throughout grain filling (Jia et al., 2021) can significantly reduce yield, mainly due to a lower number of grains per ear and, therefore, a reduced sink size for carbon accumulation. Low radiation can also reduce photosynthesis (Mu et al., 2010) and damage the photosynthetic system (H. Yang et al., 2020). However, this relationship is not straightforward and a moderate radiation reduction can also increase the yield, depending on the genotpye (H. Li et al., 2010). The young microspore stage is generally sensitive to stresses (B. Dong et al., 2017; H. Yang et al., 2020), and heat or drought during anthesis also adversely affect yield (Farooq et al., 2014; Mahrookashani et al., 2017). So, in years with low radiation, heat, or drought conditions, some genotypes might have avoided adverse conditions by an earlier or later phenological development.

With regard to the characterization of disease resistance of different genotypes, specific weather conditions are conducive to various wheat diseases during specific phenological stages (e.g. Ferrigo et al., 2016). Only a combined characterization of phenology and environmental conditions, *i.e.* a thorough envirotyping, will permit a disentanglement of disease resistance from disease avoidance due to different phenological development.

As a complement to standardized agronomic measurements, such as yield, baking quality, overwintering, plant height, thousand kernel weight or disease ratings (WBF, 2021), Vegetation indices (VI), obtained from spectral measurements, are increasingly being used to estimate crop productivity. VIs were shown to be best correlated with yield at specific phenological stages, usually shortly after flowering in the case of wheat. In this context, knowing the phenology is also critical for comparing VIs (e.g. Longchamps and Philpot, 2023; Naito et al., 2017; D. Wang et al., 2022). It allows for a temporal normalization of spectral measurements as the spectral signatures depend not only on genotypes but also on the phenological stages.

Similarly to phenology, the senescence behavior of wheat was shown to be a selection criterion for higher-yielding genotypes (Hund et al., 2019). Genotypes showing a late onset of senescence followed by a rapid progression of senescence produced higher grain yields under water-limited conditions. The so-called stay-green behavior combines a prolonged

photosynthetic activity with a rapid and efficient translocation of nutrients and sugars from other plant organs to the grain (Anderegg, Yu, et al., 2020; Cao et al., 2021; J. T. Christopher, Veyradier, et al., 2014; J. T. Christopher, M. J. Christopher, et al., 2016). In contrast, maintaining a green canopy late into the growing season, but without a rapid and efficient translocation of sugars and nutrients, might be associated with lower yield in the absence of water-limited conditions (Anderegg, Yu, et al., 2020; Kipp et al., 2014).

Knowing the end of senescence is important, as the varieties in breeding and variety testing trials are not mature and senescent at the same time, but are typically all harvested on the same date. A genotype that has been senescent for, *e.g.* ten days before harvest, but remains in the field in humid conditions might have low grain quality and higher loads of mycotoxins due to black head molds (Hershman, 2011; Lorenz, 1986; Poursafar et al., 2016). Also, particularly when humid conditions occur in combination with cooler temperatures, the breaking of seed dormancy could lead to pre-harvest sprouting (Gao et al., 2013; Zhou et al., 2017), and consequently to the degradation of starch, lipids, and proteins in grains (Yan et al., 2023). Genotypes senescent for a longer period before harvest might also be prone to grain shedding (Aasen, Kirchgessner, et al., 2020).

Finally, it is also important to know the maturity behavior of a genotype to plan optimal crop rotations (Montazeaud et al., 2016). For example, a wheat genotype with early maturity might allow a following legume cover crop to develop more biomass (Blackshaw et al., 2010). In double-crop systems with wheat and soybean as widely used in the U.S., an earlier wheat harvest can be followed by an earlier soybean sowing, increasing both growth and yield of the latter (Parvej et al., 2020).

For all these reasons, it is crucial not only to perform an adequate envirotyping but also to associate information on environmental conditions with the phenological characterization of genotypes (Costa-Neto et al., 2023; Elmerich et al., 2023) for a comprehensive view of differences in yield and quality.

Estimating the timing of phenological and senescence stages visually in the field requires frequent field visits of experts during the period when these stages usually occur. This is very time-consuming and therefore expensive (Montazeaud et al., 2016; Velumani et al., 2020), especially as breeding and variety testing trials are usually conducted in several locations to account for genotype by environment (G×E) interactions. These visual assessments also suffer from observation-bias in case the assessments are done by different experts. To overcome the drawbacks of visual field ratings, methods are being developed to screen the progression of plant development in a more automated and objective manner. Adamsen et al. (1999) used a digital camera to describe wheat senescence 20 years ago. Sadeghi-Tehran et al. (2017) used digital images generated with a field scanner to detect wheat heading and flowering. Burkart et al. (2018) extracted simple VI dynamics from single images taken with a drone 100 m above a barley field throughout the growing season and then compared these dynamics with the timing of the phenological stages. In the study of Anderegg, Yu, et al. (2020), the senescence dynamics of more than 300 winter wheat varieties were tracked with a radio spectrometer. The authors state that compared with spectral tracking, visual assessment remains the gold standard method as it showed a closer correlation with yield than the VIs derived with the spectral methods, but the latter offer the potential for up-scaling to very large breeding trials, where visual ratings are no longer feasible. J. T. Christopher, Veyradier, et al., 2014; J. T. Christopher, M. J. Christopher, et al., 2016 and Montazeaud et al., 2016 used a hand-held Greenseeker to measure the normalized difference vegetation index (NDVI) and applied dynamic models to describe the stay-green properties of wheat such as delayed onset of senescence and an accelerated senescence rate. Cao et al. (2021) compared the ability of more expensive drone-based multispectral cameras with cheaper drone-based RGB cameras to track senescence and stay-green. They concluded that while multispectral sensors allow

for a more accurate characterization of senescence parameters (*e.g.* onset, midpoint and conclusion of senescence and senescence rate)s, cheaper RGB sensors also allow for tracking senescence behavior. Whereas these approaches showed promising results, radio spectrometers and multispectral drone-based sensors are expensive. In addition, for drone-based approaches, the images have to be processed with specific photogrammetric software, and all approaches still need frequent field visits.

By monitoring plants with fixed-position cameras that take images at a high frequency (typically several times a day), the need for frequent visits to the study site can be overcome. Such fixed-position systems were applied to derive information on dynamic traits of plants (e.g., timing of phenological stages), yet, most of these PhenoCam studies focused on forests, ecology, or ecophysiology of larger systems. Typically, these studies are based on camera platforms, installed above tree canopies (e.g. Hella Ellen Ahrends et al., 2009; T. F. Keenan et al., 2014; Andrew D. Richardson, Jenkins, et al., 2007; Andrew D. Richardson, Braswell, et al., 2009), inside canopies (e.g. Kurc and Benton, 2010) or opportunistically profit from webcams pointing at relevant vegetation (e.g. Graham et al., 2010; Ide and Oguma, 2010). Andrew D Richardson et al. (2013) and Andrew D. Richardson, Hufkens, et al. (2018) based their work on the PhenoCam Network (`https://phenocam.nau.edu/`).

In agriculture, there is plenty of research that describes protocols for obtaining information on crop state (e.g. Adamsen et al., 1999; Hunt, Doraiswamy, et al., 2013), morphology (e.g. Hasan et al., 2019) and performance (e.g. T. Jensen et al., 2007; Gracia-Romero et al., 2017; Yue et al., 2019; H. Wang et al., 2020) from digital images derived from different sources. The use of fixed-position digital repeat PhenoCams is gaining interest in agriculture too. Naito et al. (2017) used PhenoCams mounted on masts at 8 m above rice fields, which combined RGB and NDVI images and took images daily from late vegetative stage to dough stage. They estimated traits related to rice yield such as shoot biomass and grain weight under different nitrogen treatments. Bhatti et al. (2024) installed an NDVI PhenoCam at 6 m above ground for gap filling of satellite-based NDVI time series. Thereby, they improved the classification of crops on satellite maps. This is one example in which PhenoCams were used to bridge the spatial and temporal gap between satellite data and close-up images (Browning et al., 2017; Andrew D. Richardson, 2019).

A limited number of studies apply PhenoCams to track phenology in agricultural experiments. Most of these phenology studies featured one genotype of one species per image, and the literature on the application of digital repeat photography in the context of variety characterization is sparse (Aasen, Kirchgessner, et al., 2020). Taylor and Browning (2021) used opportunistic images of the PhenoCam network to estimate different phenological stages of corn, wheat / barley, soybean, and alfalfa. Guo et al. (2022) tracked maize phenology in RGB images from masts of different heights. Liu et al. (2022) installed RGB timelapse cameras on sticks 1.5 m above the canopy to estimate the effects of cropping systems on crop phenology. On wheat, Zhu et al. (2016) established a fixed-position digital repeat imaging workflow on three varieties sown in three environments. In their approach, cameras were installed 5 m above the ground and high-resolution images were analyzed with computer vision algorithms to detect ears upon emergence. Velumani et al. (2020) installed 47 fixed-position cameras in four different environments. With each camera only covering a relatively small area of one single variety, they generated high-resolution images that were analyzed with deep learning algorithms to detect heading and flowering.

To our knowledge, only Aasen, Kirchgessner, et al. (2020) used PhenoCams to describe the timing of the phenological stages of nine soybean genotypes where multiple genotypes appeared on one image. Brocks, Bendig, et al. (2016) and Brocks and Bareth (2018) mounted a pair of two RGB PhenoCams on a platform, 10 m above the ground and applied stereo vision to create 3-D surface models to estimate above-ground biomass on nine barley cultivars but

not to track plant development.

Typically, for such PhenoCam examinations, images are taken throughout the growing season at high temporal resolution, *i.e.* daily to several times per day (Aasen, Kirchgessner, et al., 2020). The greenness, or generally the dynamic of the color changes of the plant canopy is then tracked with the help of VIs (Hufkens, Trevor F. Keenan, et al., 2016; Andrew D. Richardson, Braswell, et al., 2009), such as the green chromatic coordinate (GCC) and the VI dynamics are analyzed (e.g. Ide and Oguma, 2010; Migliavacca et al., 2011; Browning et al., 2017).

The platforms mentioned so far were of rather stationary nature or limited in height to several meters above ground ($\sim 6\,\mathrm{m}$). Field trials, especially in the context of multi-location trials, are usually conducted on different fields in subsequent seasons to allow for adequate corp rotation. Therefore, a PhenoCam should be ideally mounted on a mobile mast.

With an increasing distance between the camera and the plot, a better spatial resolution of the camera is needed. This was limiting in the past, but in recent years, digital imaging technology has improved a lot, increasing the signal-to-noise ratio and spatial resolution of cameras. At the same time, solar-powered cameras became available, and the data storage capacity increased. This development made compact and fully autonomous time lapse cameras commercially available that can work for weeks up to months without any intervention. Using such cameras for digital repeat imagery, mounted on mobile yet relatively high masts and in combination with the application of bespoke image analysis protocols allows to derive information on plants with a high temporal resolution at relatively low hardware costs and without the need of frequent field visits. Such a PhenoCam setup might therefore offer the potential to breeders and examination offices to get continuous information on dynamics of crop growth and senescence with high temporal resolution and precision (Aasen, Kirchgessner, et al., 2020).

Previous studies often focused on one VI at a time to track plant development and senescence (e.g. Anderegg, Yu, et al., 2020; Cao et al., 2021). Now, the progression of phenology and senescence usually leads to different changes in plants at different stages. For example, chlorophyll breakdown might be difficult to visually detect at the beginning of senescence, where 50 % of leaf chlorophyll can be lost before visual yellowing and chlorosis (E. A. Chapman et al., 2021). At the same time chlorophyll breakdown is a very dominant visual feature during later stages of senescence. It might be well tracked with visible band VIs like GCC but also multispectral VIs like NDVI, which are both related to chlorophyll absorbance. Anthocyanins, carotenoids, and sometimes colorless chlorophyll breakdown products (A. Fischer and Feller, 1994; Hörtensteiner, 2006) or changes in water content can lead to a change of the spectral signature beyond changes in greenness. Therefore, VIs that focus not only on greenness and chlorophyll, but on changes in other spectral bands in the visible and non-visible spectrum could confer complementary information on the development of the plant (Anderegg, Yu, et al., 2020; Cao et al., 2021). As an example of a VI that reflects the dynamics of the breakdown of two pigment types, the plant senescence reflectance index (PSRI) uses chlorophyll/carotenoid ratio (Anderegg, Yu, et al., 2020; Merzlyak et al., 1999).

VIs might also be combined. For example, Anderegg, Yu, et al. (2020) used a random forest regression based on multiple VI-derived senescence dynamics parameters to predict yield and grain protein content. Guo et al. (2022) integrated textural and spectral dynamics of RGB images into a single analysis to track the phenology of maize. Longchamps and Philpot (2023) applied pairs of normalized difference metrics of two VIs, one primarily related to chlorophyll concentration, one primarily related to water content, to track the phenology of corn and soybean.

The scope of this research is to establish and evaluate a PhenoCam-based lean phenotyping workflow to monitor wheat phenology and senescence. Overall, the hypothesis is tested that

such a workflow can become superior to conventional approaches in variety testing. Therefore, in detail, (I.) a mobile mast-based PhenoCam setup will be introduced, suitable for high-throughput field phenotyping in the context of wheat-variety testing. (II.) A method to pre-process and analyze multiple VI dynamics at once will be suggested, to predict the timing of different stages of plant development. (III.) The PhenoCam method will be compared with different types of visual reference field ratings and VIs based on drone images (RGB and multispectral). (IV.) The cost effectiveness of the different methods will be analyzed in a simple economic calculation example.

4.2 Materials and Methods

This study was carried out on a wheat variety testing trial over three consecutive seasons (2020-2021, 2021-2022, 2022-2023). During this period, the fields were observed with PhenoCams and visual field ratings were collected as reference measurements. As technical benchmark methods, additional drone flights were conducted. The methods were compared to each other in terms of performance and cost. A conceptual overview of the study is provided in Fig. 4.1.

4.2.1 Field experiments

The winter wheat variety testing experiment (Fig. 4.2a) was sown at Agroscope agricultural research station, Changins, Switzerland [46°23′55.4″N 6°14′20.4″E, 425 m.a.s.l., the World Geodetic System (WGS) 84]. The soil of the experimental site is a shallow Calcaric Cambisol (Baxter, 2007; Cárcer et al., 2019). The trial consisted of 30 modern registered European winter wheat varieties and is further referred to as the EuVar trial. The same varieties were sown in three different treatment regimes for the three seasons. In the "maximal" regimen, one growth regulator and one fungicide treatment were applied. In the "medium" regimen, there was only the growth regulator application and not the fungicide application. In the "minimal" regimen, neither a growth regulator nor a fungicide was applied. Tables S3.1 and S3.2 give a detailed overview of the different treatments. Fertilizers and herbicides were applied in three splits and at equal rates to all treatments according to the Proof of Ecological Performance (PEP) certification guidelines (Swiss Federal Council, 2013), which represent a minimal standard of good practice for agriculture in Switzerland. Each variety-treatment combination was repeated on three plots. Within single plots, a wheat genotype was sown in eight rows, with a spacing of 15 cm between them, resulting in an observable canopy of about 1.25 m x 6.7 m each. Within blocks of 3 by 10 plots, the genotypes were randomly distributed and these blocks were randomly nested within three treatment replicates. Each replicate contained three blocks, and each block was treated with one of the three treatments. The 270 plots of the experiment span 27 rows (which followed the tractor track direction) and 10 columns (Fig. S3.1). In total, the experiments were 79 m long (in tractor track direction) and about 55 m wide. The first two seasons of this experiment were first described by Treier et al., 2024, but the main characteristics are also described here for clarity.

4.2.2 PhenoCam setup

Tikee PRO 2 / 2+ (Enlaps SAS, Montbonnot-Saint-Martin, France) solar powered autonomous time lapse camera systems (Fig. 4.2b) were installed in the field on TekMast VMS-21-M mobile field masts (Teksam Company NV, Genk, Belgium) at 12 m above ground (Fig. 4.2a). The masts were stabilized with ropes from three sides and the anchor pins were reinforced with ground screws (Fig. S3.2). Each camera system carried two cameras with CMOS RGB sensors (4608 x 3456 pixels) which had a fixed horizontal angle of 90° between them. According to

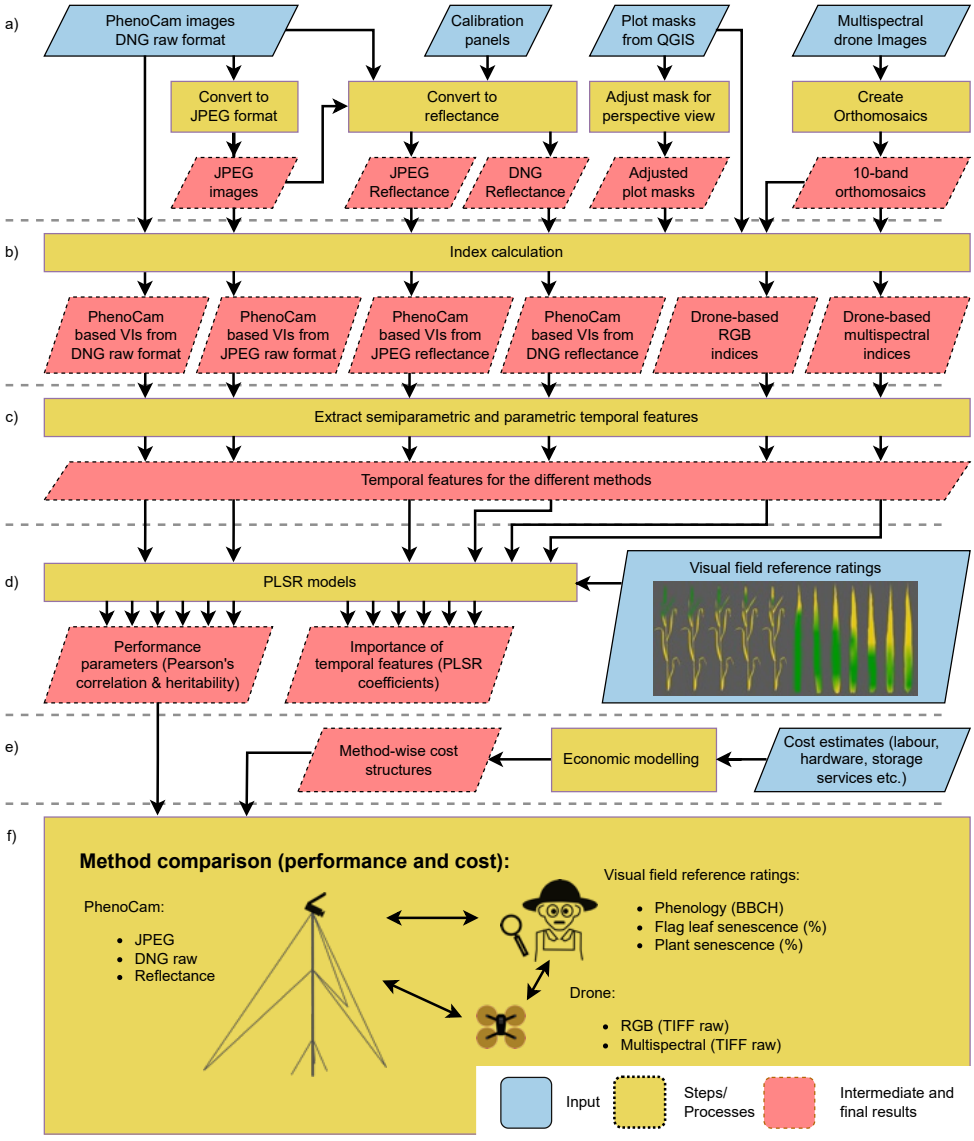


FIGURE 4.1: Overview of the workflow of the study: (a) Image data was acquired with PhenoCams and converted to different image- and data types. Drone images were aligned into orthomosaics. For PhenoCams and drone data, plot masks were created. (b) From all plots on all images, various vegetation indices (VIs) were calculated. (c) From VI values, semiparametric and parametric temporal features were derived. (d) Visual field reference ratings were carried out for three types of ratings (phenology, flag leaf senescence and plant senescence). VI based temporal features were then used to predict the timing of visual ratings in partial least squares regression (PLSR) models. These models also allowed to determine the most relevant temporal features for PLSR prediction. (e) To also compare the cost structure of the different methods, conceptual economic modeling was carried out. (f) The methods were compared to each other in terms of performance and cost.

full width at half maximum specifications, the spectral sensitivity of the sensors was highest from 430 nm to 500 nm for the blue spectral band (B), from 475 nm to 600 nm for the green band (G) and from 580 nm to 660 nm for the red band (R). The two cameras of one system together covered an angle of 220° horizontally and overlapped for central parts of these 220° regions (Fig. 4.2d). Two masts with two camera systems each (eight cameras in total) were set up on the narrow side of the experiment at a distance of 30 m (in 2022) to 45 m (in 2021 and 2023) from each other, but only four cameras covered different parts of the EuVar experiments. These four cameras were oriented from north-east to south-west. With a vertical opening angle of 90°, the cameras were installed obliquely pointing toward the ground, covering at least the region from the base of the masts to the edge of the fields in the direction of the horizon.

The masts were installed in the field shortly after sowing and uninstalled when the wheat was mature, one day before harvest or after harvest. The cameras were programmed to take images every 2 hours in the period from 7:00 to 17:00 every day to cover the period of daylight from spring onward. Images were saved in DNG raw format on SD (Secure Digital) cards plugged into the cameras. As the Tikee PRO 2 camera model allowed for a maximal memory size of 128 GB, the cards had to be replaced once during the duration of the experiment, for which the masts had to be lowered. This was done about two weeks before the expected heading date.

4.2.3 Image file format

Images were saved in 16-bit DNG raw format. This format is data-heavy (34 MB/image), and to test whether the lighter 8-bit JPEG (Joint Photographic Experts Group) format (15.9 MB/image) also allows for similar quality, DNG images were transformed to JPEG format in Python.

4.2.4 Multispectral measurements

In parallel to mast recordings, the trials were also monitored with an airborne MicaSense RedEdge-MX Dual multispectral camera (MicaSense Inc., Seattle, Washington, USA). The camera was carried by a DJI Inspire 2 drone (SZ DJI Technology Co. Ltd., China). The flight height was 60 meter in 2021 and 40 meter in 2022 and 2023, resulting in a ground sampling distance (GSD) of 3.98 cm and 2.71 cm, respectively. The side overlap was set to 80 %, the flight speed was limited to $5\,\mathrm{m\,s^{-1}}$ and an image was taken at an interval of 2 s in 2021 and 1 s in 2022 and 2023, resulting in a front overlap of approximately 80 % for the two flight configurations. Images of a calibrated MicaSense reflectance panel were taken at the beginning and the end of each flight. Flights were conducted throughout the growing seasons. From shortly before heading (BBCH 59; Lancashire et al., 1991) to the end of senescence, the flights were flown at higher temporal intervals of weekly to several times a week (Fig. 4.3). The images were saved in a raw TIFF image format.

Agisoft Metashape Professional software (Agisoft LLC, St. Petersburg, Russia) was used to align images, to generate 10 band orthomosaics (Fig. 4.2d) covering the whole experiment. Details on the spectral properties of the 10 bands of the sensor are described in Table S3.3. The reflectance panel used for calibration featured a QR code and Agisoft provides the functionality to detect this code and conduct a calibration of the targets autonomously.

4.2.5 Mask creation

To define regions of interest (ROI) on images, plot masks (Fig. 4.2c) were created for each plot appearing on each camera in each year. First, orthogonal masks were created based on

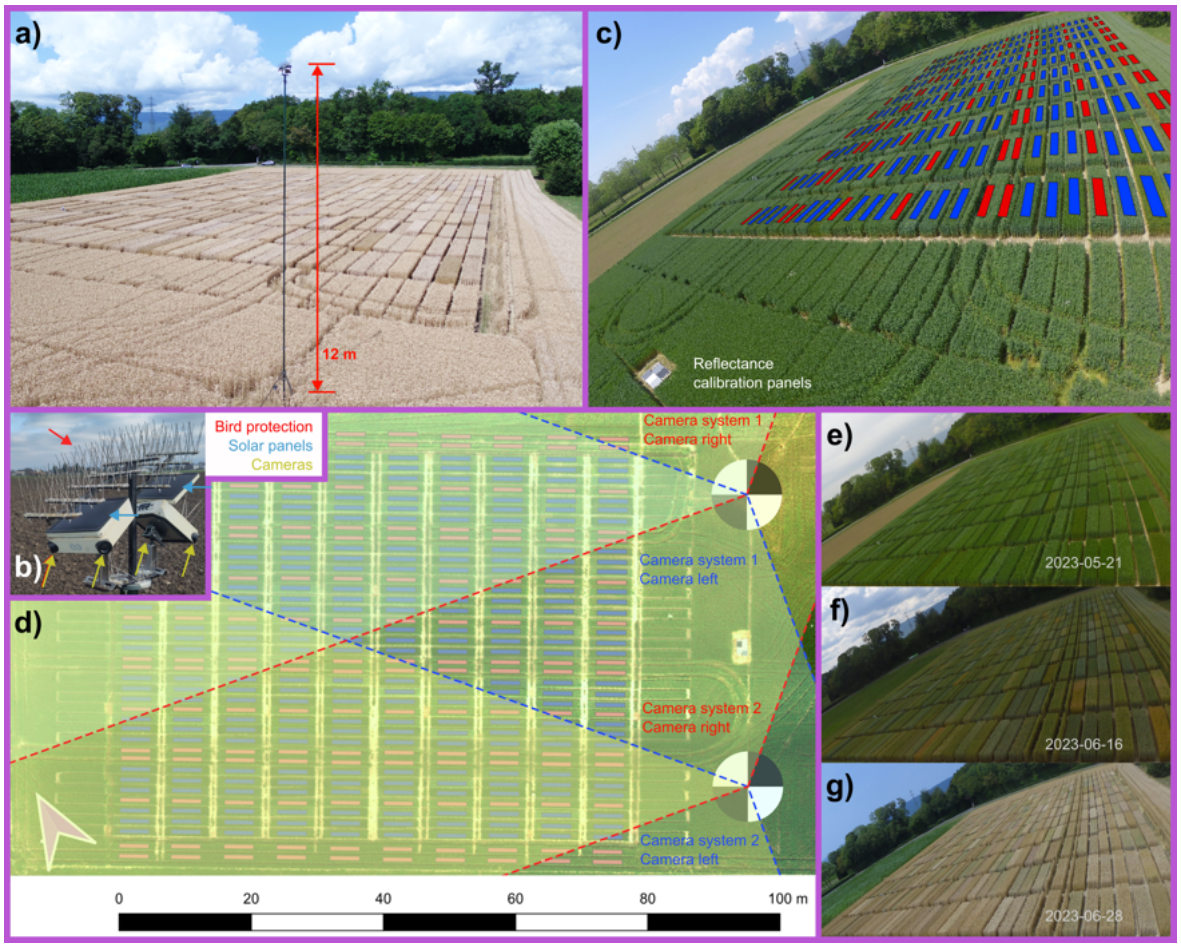


FIGURE 4.2: Overview on the PhenoCam setup and data. Masts (a) were installed in the field carrying camera systems (b) to create at least five images a day (*e.g.* c & e - g). (d) Two masts with two camera systems pointing toward the experiment of interest were installed at the narrow side of the experiment, partially covering the same plots from different angles as indicated by the opaque circle segments. In the back of (d), an orthomosaic is shown as it was created for every drone flight. Plot masks were created and adjusted for perspective view for PhenoCam images (c) or adjusted to field plots for drone-based orthomosaics (d). The colors of the masks in (c) and (d) indicate weather a plot was part of the experiment (blue) or a border plot to separate different treatments (red).

a CSV file, specifying row and column position of the plot and corresponding meta information (*e.g.* genotypes, treatments, etc.) of the plots. This was achieved with a Python 3.8 script (van Rossum, Guido and Drake, Fred L., 2009), using the "ogr" module of the "GDAL" library (GDAL/OGR Contributors, 2024) and defining approximate plot dimensions in the image coordinate system (that is, pixel coordinates) directly in the script. Then a homography transform was applied to the shape coordinates to achieve a perspective view. The homography matrix was estimated based on four corresponding points between orthogonal masks and perspective images, using a Python script provided by Socretquuliqaa Lee (`https://gist.github.com/Socret360/bcefb0f95cfc20800ea3409f40b8bb58`). The transformed coordinates were calculated as the dot product of the orthogonal shape coordinates with the homography matrix. The masks were then manually adjusted in QGIS (QGIS Development Team, 2022) to match a base image. To account for border effects in the field and for inaccuracies of referencing and superimposition of different images, buffers were applied to masks. As a consequence of perspective, the masks had very different sizes, and shape

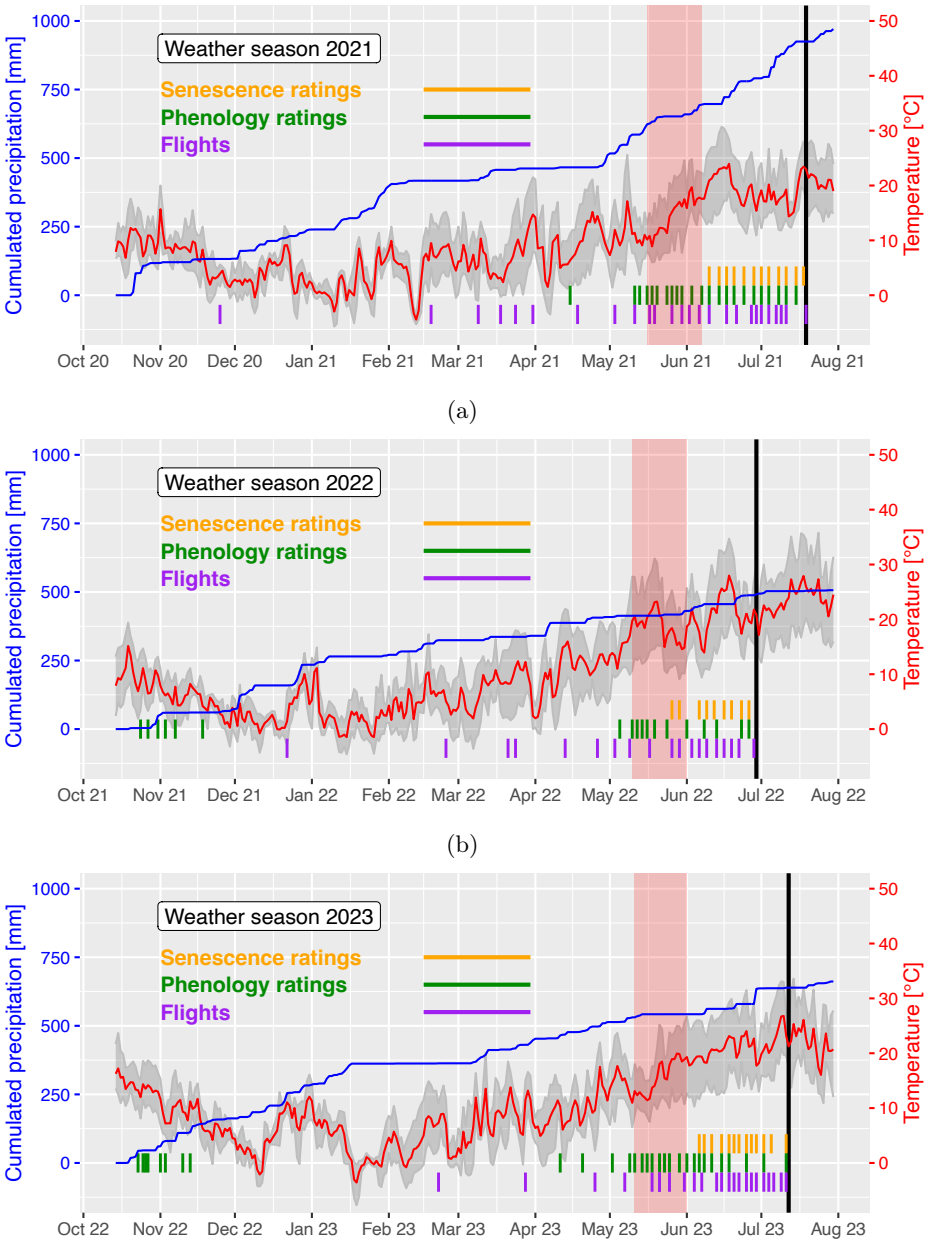


FIGURE 4.3: Overview on measurements and meteorological conditions in the three experimental seasons. (a) - (c) show the general weather conditions during the growing seasons 2021 to 2023 from sowing until after harvest. Red shows the mean air temperature, and the shades indicate daily temperature minima and maxima. The vertical purple lines indicate the dates of flights, the green and orange lines the dates of phenology and senescence ratings respectively. Cumulative precipitation is shown as a rising blue line. During the period shaded in red, heading was observed in the field. Harvest dates are marked by black lines.

buffers between masks were adjusted based on individual visual judgment, resulting in shapes corresponding to approximately 50 % of the surface of the plots. The masts shook slightly in the wind, resulting in differences in the position and orientation of the cameras. Thus, the masks needed adjustment over time. To that end, well-illuminated reference images were selected throughout the growing season, which were taken between 10:30 and 12:30. The masks were then manually adjusted for these reference images in QGIS based on the masks from the base image, saved to GeoJSON format again, and used as reference masks for the respective reference image and all subsequent images until a new reference image was available.

Georeferenced masks for drone data analysis (Fig. 4.2d) were created similarly as masks for PhenoCams, but in the Swiss CH1903+ / LV95 - EPSG:2056 coordinate reference system (CRS) and without homography transform. Border buffers of 25 cm and up to 1 m were left on plot width and length, respectively.

4.2.6 Index calculation and index value extraction

A large number of color VIs were proposed for different applications in agronomy and were compiled and compared in various publications (*e.g.* Anderegg, Yu, et al., 2020; Anderegg, Tschurr, et al., 2023; Cao et al., 2021; Hasan et al., 2019; Hunt, Doraiswamy, et al., 2013; W. Li et al., 2023; D. Wang et al., 2022). In this study, a multitude of described VIs were calculated from color bands (Fig. 4.1b) in the visual RGB color space (Table 4.2) and for multispectral drone images also from the infrared range (Table 4.3).

Each pixel within each plot featured a value for each VI. To aggregate these values to single values per plot and VI, zonal statistics were applied by calculating the mean, the 50^{th} percentile (or median), the 90^{th}, the 98^{th} and the 99^{th} percentile of the pixel values within the individual masks. Higher percentiles were included as for most VIs, higher values are associated with plants and choosing a high percentile can help to avoid exposed soil within plots that affects VI values (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016). The 98^{th} and the 99^{th} percentile are close to maximal values (*i.e.* the 100^{th} percentile), but not as sensitive to artifacts and disturbances.

4.2.7 Extraction of temporal features

The values of these VIs often follow characteristic dynamics throughout the growing season (*e.g.* Figs. 4.2e - 4.2g). For very early stages such as emergence (BBCH 09; Lancashire et al., 1991) or the three-leaf stage (BBCH 13), no corresponding temporal behavior of VIs, *e.g.* a sudden and pronounced increase in VI values, could be observed based on visual inspection. Therefore, the analysis conducted in this study focused on the stages from heading (BBCH 59) to senescence. Heading occurs typically after mid-May or around 210 days after sowing (DAS) and to reduce the amount of data to be handled, as well as to simplify the automatic feature extraction, the data was limited to the relevant period. Thus, to extract semiparametric (Fig. 4.1c) temporal features from VIs, only VI values from one month before expected heading, *i.e.* 180 DAS, up to one or two days before harvest were considered. To derive smooth VI dynamics, either the rolling mean, a Savitzky-Golay filter, spline smoothing, and locally estimated scatter plot smoothing (loess) were applied to the data (*e.g.* Bhatti et al., 2024; Guo et al., 2022; Hufkens, Melaas, et al., 2019; Klosterman et al., 2014). The maximal and minimal values of the smoothed dynamics from the four different smoothing types were defined as 100 % and 0 %, respectively, and two temporal features were extracted as the time in DAS when the value reached 10 % and 2 %, respectively, similar to J. T. Christopher, Veyradier, et al. (2014), where 10 % was defined as conclusion of senescence. An overview of the temporal feature types is presented in Table 4.1. For VIs with increasing values toward maturity and senescence,

the VI values were reflected over the DAS axis (*i.e.* the x-axis) so that the maximum always appeared earlier than the minimum.

TABLE 4.1: Overview and description of temporal features. The temporal features were calculated for all VIs and data aggregation methods (*i.e.* mean and different percentiles).

Temporal feature	**Description**	**Feature class**
$D1_{LocMax\_1}$	1st derivative of Gompertz, 1st local maximum	Gompertz derivative
$D2_{LocMax\_1}$	2nd derivative of Gompertz, 1st local maximum	
$D2_{LocMin\_1}$	2nd derivative of Gompertz, 1st local minimum	
$D3_{LocMax\_1}$	3rd derivative of Gompertz, 1st local maximum	
$D3_{LocMax\_2}$	3rd derivative of Gompertz, 2nd local maximum	
$D3_{LocMin\_1}$	3rd derivative of Gompertz, 1st local minimum	
$D4_{LocMax\_1}$	4th derivative of Gompertz, 1st local maximum	
$D4_{LocMin\_1}$	4th derivative of Gompertz, 1st local minimum	
$D4_{LocMax\_2}$	4th derivative of Gompertz, 2nd local maximum	
$D4_{LocMin\_2}$	4th derivative of Gompertz, 2nd local minimum	
$Loess_{Max}$	Loess* smoothed curve at maximum	Loess smoothing with threshold
$Loess_{Min}$	Loess* smoothed curve at minimum	
$Loess_{0.02}$	Loess* smoothed curve at 2 % of max-min range	
$Loess_{0.1}$	Loess* smoothed curve at 10 % of max-min range	
$Rolling_{0.02}$	Rolling mean at 2 % of max-min range	Rolling mean smoothing with threshold
$Rolling_{0.1}$	Rolling mean at 10 % of max-min range	
$Sav.Gol_{0.02}$	Savitzky–Golay smoothed curve at 2 % of max-min range	Savitzky–Golay smoothing with threshold
$Sav.Gol_{0.1}$	Savitzky–Golay smoothed curve at 10 % of max-min range	
$Spline_{0.02}$	Smoothing spline at 2 % of max-min range	Spline smoothing with threshold
$Spline_{0.1}$	Smoothing spline at 10 % of max-min range	
$Spline_{Max}$	Smoothing spline at maximum	
$Spline_{Min}$	Smoothing spline at minimum	

*locally estimated scatter plot smoothing (Loess)

These semiparametric approaches allow capturing very dynamic seasons, but do not imply growth dynamics and are more prone to overfitting if the measurement noise is very high (Roth, Rodríguez-Álvarez, et al., 2021). Thus, in addition, parametric temporal features were derived from Gompertz models (Fig. 4.1c, *e.g.* Anderegg, Yu, et al., 2020; E. A. Chapman et al., 2021). First, for each measurement, the accumulated thermal time from sowing was calculated in growing degree days (GDD) as

$$T_{therm,h} = \sum_{h=1}^{n} \begin{cases} \frac{T_{max,h}+T_{min,h}}{2 \cdot 24} - \frac{T_{base}}{24}, & \text{if } T_{min,h} > T_{base}, \\ 0, & \text{if } T_{min,h} \leqslant T_{base}, \end{cases} \tag{4.1}$$

where $T_{max,h}$ and $T_{min,h}$ are the maximum and minimum temperatures of the n^{th} hour h after sowing. T_{base} was assumed to be 0 °C (G. McMaster, 1997). 24 hourly means sum up to the GDD of one day.

TABLE 4.2: Red-green-blue (RGB) vegetation indices (VI). For some VIs, the literature provides multiple, sometimes significantly different, formulas. For excess red index (ExR), excess green minus excess red (ExGR), and triangular greenness index (TGI), this is indicated as subscript at the end of the index names. The numbers behind the band names in the formulas indicate the wavelengths of the spectral bands used.

Index	**Full name**	**Formula**		**Reference**
ExB	Excess blue index	$\frac{(1.4. \cdot Blue_{444}) - Green_{531}}{Red_{650} + Green_{531} + Blue_{444}}$	(4.2)	Lu et al. (2019); Mao et al. (2003); Xu and C. Li (2022)
ExG	Excess green index	$(2 \cdot Green_{531}) - Red_{650} - Blue_{444}$	(4.3)	Woebbecke et al. (1995)
ExR_b	Excess red index	$(1.4 \cdot Red_{650}) - Blue_{444}$	(4.4)	Meyer, Mehta, et al. (1998)
ExR_g	Excess red index	$1.4 \cdot Red_{650} - Green_{531}$	(4.5)	Cao et al. (2021); Lu et al. (2019); Meyer, Mehta, et al. (1998); Xu and C. Li (2022); J. Zhang et al. (2021)
$ExR_{g\text{-}norm}$	Normalized excess red index	$\frac{(1.4 \cdot Red_{650}) - Green_{531}}{Red_{650} + Green_{531} + Blue_{444}}$	(4.6)	Lu et al. (2019)
$ExGR_{Meyer}$	Excess green minus excess red	$ExG - ExR_b = 2 \cdot Green_{531} - 2.4 \cdot Red_{650}$	(4.7)	Meyer and Neto (2008)
$ExGR_{Zhang}$	Excess green minus excess red	$ExG - ExR_g = 3 \cdot Green_{531} - 2.4 \cdot Red_{650} - Blue_{444}$	(4.8)	J. Zhang et al. (2021)
$ExGR_{Lu}$	Excess green minus excess red	$ExG - ExR_{g-norm} = (2 \cdot Green_{531}) - Red_{650} - Blue_{444} - \frac{(1.4 \cdot Blue_{444}) - Green_{531}}{Red_{650} + Green_{531} + Blue_{444}}$	(4.9)	Lu et al. (2019)
GBRI	Green-blue ratio index	$\frac{Green_{531}}{Blue_{444}}$	(4.10)	Sellaro et al. (2010)
GLI	Green leave ratio	$\frac{2 \cdot Green_{531} - Red_{650} - Blue_{444}}{2 \cdot Green_{531} + Red_{650} + Blue_{444}}$	(4.11)	Louhaichi et al. (2001)
GRRI	Green-red ratio index	$\frac{Green_{531}}{Red_{650}}$	(4.12)	Tucker (1979)
IKAW	Kawashima Index	$\frac{Red_{650} - Blue_{444}}{Red_{650} + Blue_{444}}$	(4.13)	Kawashima (1998)

Continued on next page

Table 4.2 – continued from previous page

Index	Full name	Formula		Reference
MGRVI	Modified green-red vegetation index	$\frac{(Green_{531})^2 - (Red_{650})^2}{(Green_{531})^2 + (Red_{650})^2}$	(4.14)	Bendig et al. (2015)
MNVI	Meyer-Neto vegetation index	$2 \cdot Green_{531} - 2 \cdot Blue_{444} - 2.4 \cdot Red_{650}$	(4.15)	Jin et al. (2017); Meyer and Neto (2008)
NGBDI	Normalized green–red difference index	$\frac{Green_{531} - Blue_{444}}{Green_{531} + Blue_{444}}$	(4.16)	Meyer and Neto (2008); Xu and C. Li (2022)
NGRDI	Normalized green–blue difference index	$\frac{Green_{531} - Red_{650}}{Green_{650} + Red_{444}}$	(4.17)	Meyer and Neto (2008)
RBRI	Red-blue ratio Index	$\frac{Red_{650}}{Blue_{444}}$	(4.18)	Hasan et al. (2019); Segal (1982); Sellaro et al. (2010)
RGBVI	Red-reen-blue vegetation index	$\frac{(Green_{531})^2 - (Blue_{444} \cdot Red_{650})}{(Green_{531})^2 + (Blue_{444} \cdot Red_{650})}$	(4.19)	Bendig et al. (2015)
TGI_{fixed}	Triangular greenness index (simplified)	$Green_{531} - 0.39 \cdot Red_{650} - 0.61 \cdot Blue_{444}$	(4.20)	Kavaliauskas et al. (2023); Hunt, Daughtry, et al. (2011) Segarra et al. (2023)
TGI	Triangular greenness index	$-0.5 \cdot \left[(\lambda_{650} - \lambda_{444})(Red_{650} - Green_{531}) - (\lambda_{650} - \lambda_{531})(Red_{650} - Blue_{444})\right]$	(4.21)	Hunt, Daughtry, et al. (2011)
VARI	Visible atmospherically resistant index	$\frac{Green_{531} - Red_{650}}{Green_{531} + Red_{650} - Blue_{444}}$	(4.22)	Gitelson, Kaufman, et al. (2002)
RCC	Red chromatic coordinate	$\frac{Red_{650}}{Red_{650} + Green_{531} + Blue_{444}}$	(4.23)	Gillespie et al. (1987)
GCC	Green chromatic coordinate	$\frac{Green_{531}}{Red_{650} + Green_{531} + Blue_{444}}$	(4.24)	Gillespie et al. (1987)
BCC	Blue chromatic coordinate	$\frac{Blue_{444}}{Red_{650} + Green_{531} + Blue_{444}}$	(4.25)	Gillespie et al. (1987)
R	Red	Red_{650}	(4.26)	-
G	Green	$Green_{531}$	(4.27)	-

Continued on next page

Table 4.2 – continued from previous page

Index	Full name	Formula		Reference
B	Blue	$Blue_{444}$	(4.28)	-

TABLE 4.3: Multispectral vegetation indices. The numbers behind the band names in the formulas indicated the wavelengths of the spectral bands used.

Index	Full name	Formula		Reference
ARI1	Anthocyanin reflectance index	$\frac{1}{Green_{560}} - \frac{1}{RedEdge_{705}}$	(4.29)	Gitelson, Merzlyak, and Chivkunova (2001)
ARI2	Anthocyanin reflectance index	$NIR_{842} \cdot \left(\frac{1}{Green_{560}} - \frac{1}{RedEdge_{705}}\right)$	(4.30)	Gitelson, Merzlyak, and Chivkunova (2001)
DVI	Difference vegetation index	$NIR_{842} - Red_{668}$	(4.31)	Tucker (1979)
EVI	Enhanced vegetation index	$2.5 \cdot \frac{NIR_{842} - Red_{650}}{NIR_{842} + 6 \cdot Red_{650} - 7.5 \cdot Blue_{444} + 1}$	(4.32)	Huete et al. (2002)
NDRE	Normalized difference red edge index	$\frac{NIR_{842} - RedEdge_{717}}{NIR_{842} + RedEdge_{717}}$	(4.33)	Gitelson and Merzlyak (1994); Barnes et al. (2000); Tang et al. (2022)
NDVI	Normalized difference vegetation index	$\frac{NIR_{842} - Red_{668}}{NIR_{842} + Red_{668}}$	(4.34)	Rouse et al. (1974)
$NDVI_{717}$	Normalized difference vegetation index	$\frac{RedEdge_{717} - Red_{668}}{RedEdge_{717} + Red_{668}}$	(4.35)	Rouse et al. (1974)
$PSRI_{705}$	Plant senescence reflectance index	$\frac{Red_{650} - Green_{531}}{RedEdge_{705}}$	(4.36)	Merzlyak et al. (1999)
$PSRI_{717}$	Plant senescence reflectance index	$\frac{Red_{668} - Green_{560}}{RedEdge_{717}}$	(4.37)	Merzlyak et al. (1999)
$PSRI_{740}$	Plant senescence reflectance index	$\frac{Red_{650} - Green_{531}}{RedEdge_{740}}$	(4.38)	Merzlyak et al. (1999)
SAVI	Soil adjusted vegetation index	$1.5 \cdot \frac{NIR_{842} - Red_{650}}{NIR_{842} + Red_{650} + 0.5}$	(4.39)	Huete (1988)
SR	Simple ratio	$\frac{NIR_{842}}{Red_{668}}$	(4.40)	Birth and McVey (1968); Jordan (1969)

Then, data was selected for the relevant growth period by just considering data from 180 DAS to harvest. VI values were reflected over the DAS axis again where necessary, but this time ensuring that the minimum appeared earlier than the maximum. The data was then smoothened with a loess function to extract the minimum (LoessMin) and maximum (LoessMax) of the the smoothened VI data. The original VI values that appeared before LoessMin were set to the LoessMin value, and the values appearing after LoessMax were set to the LoessMax value. These restricted VI values should ensure that the model captures the main slope and is not dampened by high VI values before LoessMin or a possible decrease in VI after LoessMax. In addition, the data was shifted in such a way that LoessMin was 0 and the restricted VI data started to increase around 0 GDD. This translation of the restricted VI data ensured that the value range was suitable for fitting a Gompertz model,

$$I = ae^{-be^{-ct}}, \tag{4.41}$$

with the package "nls.multstart" (Padfield and Matheso, 2020) in R (R Development Core Team, 2022). In the model, I represents the VI value at time t. a is the asymptote and was restricted to values from 0.9 to 1.1 $\cdot$ LoessMax. b is a location parameter that mainly affects the starting point of the curve. Parameter c impacts the slope and the starting point. The Gompertz parametrization allowed for a monotonously increasing dynamic, from which temporal features were derived by calculating the first four derivatives of the fitted Gompertz model. Local minima and maxima of the derivatives were determined and the timing of the local minima and the maxima as well as of LoessMin and LoessMax were extracted. The timing was then transformed from GDD back to DAS to use the same temporal unit as for the semiparametric method. These procedures were performed individually for the mean and the different percentiles used for aggregation of the VI values for each VI in each plot.

This procedure was also adopted for drone-based VI. As for the visual ratings, VI values were smoothed with a penalized smoothing spline in the "pspline" package, with degrees of freedom set to two thirds of the number of measurements, and interpolated for single days. From daily VI values, semiparametric and parametric temporal features were then extracted as for PhenoCam data, just without the different smoothing approaches.

4.2.8 In-field calibration panels

Color VIs can be based on raw digital numbers (DN) of images, or reflectance values can be derived from DNs with calibration panels. To this end, five calibration panels were installed in the field (Fig. 4.2c) throughout the growing season (just from 2021-05-26 for the first year). Reflectance values of the individual panels were measured in the field on 2023-06-28 (black: 5.2 %, dark gray: 10.3 %, gray: 16 %, light gray: 24 %, white: 50.3 %) with a field portable spectroradiometer PSR + 3500 (Spectral Evolution Inc., Haverhill, Massachusetts, USA). The wheat and weeds around the panels were removed periodically to guarantee a direct line of sight from the PhenoCams onto the panels. If panels were heavily stained, *e.g.* after a heavy rainfall, they were cleaned. To avoid sustained staining, the panels were mounted on a wooden structure about 15 cm above the ground from season 2022 onward.

The reflectance panels needed to be detected in the single images. To avoid another manual adjustment as for the plot masks, a semi-autonomous pipeline was developed in Python 3.8 (van Rossum, Guido and Drake, Fred L., 2009). The coordinates of an approximate region where the panels were to be found within the images were provided with the images in a CSV file. Within these regions, the five panels were detected by searching for areas of homogeneous textures. To that end, the variance of the Laplacian transform was calculated and an Otsu threshold was applied using the packages "OpenCV" (Bradski and Kaehler, 2000) and "NumPy" (Harris et al., 2020). From the resulting filtered variance images, connected components that fell within a specific pixel size range were selected. The minimum and maximum pixel size was determined for each camera-year combination individually with respect to the size of panels in the images. If five connected components were found, they were ordered by intensity and correlated with the reflectance values of the panels. If the coefficient of determination r^2 exceeded 0.88, the calibration was considered valid and the DN values were transformed into reflectance values using the empirical line method. 0.88 was selected as threshold, as it allowed a very strong correlation ($r > 0.938$) between DN and reflectance without being too restrictive and disregarding too many values, thereby decreasing the temporal resolution of reflectance-based VI values too much. In total, 14051 images were taken and for 3929 images, calibration data was directly available. For the 3929 images of the second camera of the same camera system, the calibration equation was taken from the first image. For 2592 images where no calibration was available from the same camera system, the calibration equation of an image of another camera was used if taken within the same period (± 15 min), and with the same ISO setting and exposure time. For 3601 images, no suitable calibration information was found and no reflectance was calculated for those cases, reducing the temporal resolution of reflectance-based VI values to 81 % compared to DN-based VI values.

4.2.9 Visual field reference ratings

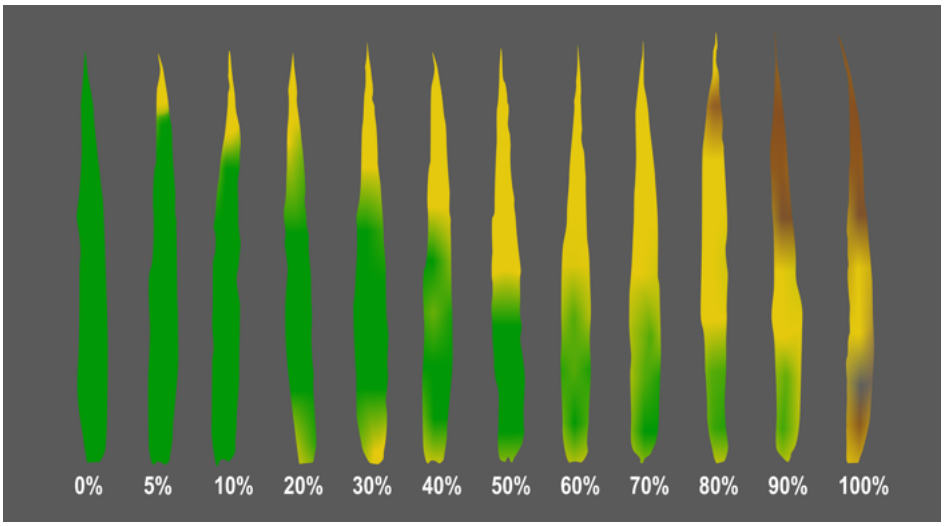
Three types of in-field reference ratings were conducted in parallel to PhenoCam observations and flights (Fig. 4.1d), and recorded using the Field Book app (Rife and Poland, 2014): Phenology ("BBCH"), flag leaf senescence ("SenLeaf") and plant senescence ("SenPlant").

Phenology was rated according to the BBCH scale (Lancashire et al., 1991). The rating interval was two to four days during the heading period and decreased to approximately weekly toward complete maturity.

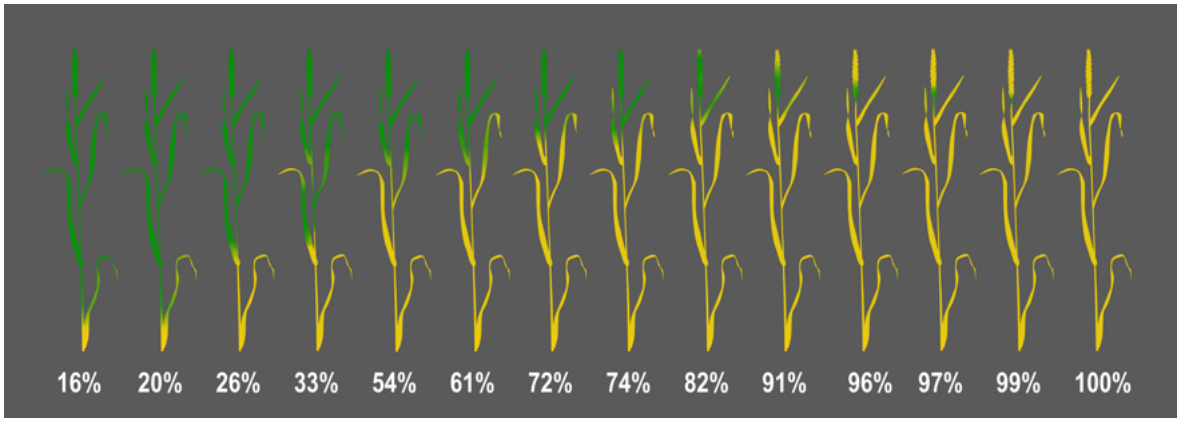
The senescence rating of the flag leaf was carried out according to the scale of Pask et al. (2012), where 0 % corresponds to a fully green leaf and 100 % to a fully senescent leaf (Fig. 4.4a).

The plant senescence rating (Fig. 4.4b) was inspired by the plot senescence of Anderegg, Yu, et al. (2020) and the peduncle and ear senescence rating of E. A. Chapman et al. (2021). The plot rows were opened manually and the senescence of the whole plant was rated from 0 % (fully green plant) to 100 % (completely senescent plant). Field ratings were performed mostly in 5 % steps except for very late ratings from 95 % onward. The last 5 % were rated in smaller steps and mainly related to changes around the ears bases and peduncles. 100 % ratings was only achieved when the ears base as well as the peduncles were completely senescent.

The single rating events for phenology and both types of senescence are visualized in Fig. 4.3. All visual ratings were in DAS. To allow for daily resolution, the values were smoothed with a penalized smoothing spline in the R package "pspline" (Ripley and Ramsey, 2024), with degrees of freedom set to two thirds of the number of measurements, and interpolated for single days. The rating of a specific reference level was given the DAS value of the day on which the interpolated value exceeded a specific reference level for the first time.



(a)



(b)

FIGURE 4.4: Senescence rating scales. (a) Flag leaf senescence, according to Joshi et al. (2007) and Pask et al. (2012). (b) Plant senescence scale. The percentage depicts senescent proportion of all pixels. Figure (a) was inspired by an image of the John Innes Centre and the University of Nottingham.

4.2.10 Partial least squares regression (PLSR) models to predict phenology and senescence from temporal features

As the temporal features used as predictors were expected to be highly correlated, PLSR was preferred over other approaches such as random forest regression. Random forests are prone to overfitting when using highly correlated data (*e.g.* Gregorutti et al., 2016) and PLSR was shown to produce more generalizable predictions than random forests (*e.g.* Lee et al., 2017). Recursive feature elimination (RFE) can be applied to increase generalizability and the risk of overfitting (Gregorutti et al., 2016; Darst et al., 2018), however, PLSR does not need a lot of computational capacity, whereas repeated RFE can be very computationally intensive.

Feature selection

The temporal features extracted previously for each plot were used as input data in a PLSR analysis (Fig. 4.1d) with the R package "PLS" (Mevik and Wehrens, 2007). Because many VIs were tested based on multiple aggregation percentiles and the mean aggregation, the number of features as predictor variables in PLSR initially exceeded the number of observations. Although PLSR analysis is suitable for this situation (Carrascal et al., 2009), the number of temporal features was reduced based on the magnitudes of the relative PLSR coefficients $\beta_{\mathrm{rel,i}}$, which were calculated for each temporal feature type as:

$$\beta_{rel,i} = \frac{|\beta_i|}{\sum_{i=1}^{n}|\beta_i|}, \tag{4.42}$$

where β_{i} denotes the PLSR coefficient of the i^{th} of n temporal features. PLSR started with the full set of temporal features available for all plots. The features with the lowest $\beta_{\mathrm{rel,i}}$ were skipped in a backward feature elimination until the most predictive features were left in the model similar to methods summarized in Mehmood et al., 2012. At the beginning, 200 temporal features were dismissed at each iteration. With the decrease in features, the number of dismissed features continuously decreased. When *e.g.* 345 features were left, 5 temporal features were dismissed at each iteration, just 2 features when 280 features were left and just 1 feature for the last iterations from 100 features down to 15 features. The PLSR used 10 components and 10 segments for cross-validation. With each model, the different levels of the reference rating types were predicted using the full set of observations and were correlated with the reference rating. With this procedure, it was determined that the correlations were relatively stable above 200 temporal features but started to weaken below, and the number of temporal features was set to 200 for the next step.

100 times repeated cross validation

For each level of each type of reference measurement (BBCH, SenLeaf, SenPlant), in the previous step, a set of 200 temporal features was selected. This reduction in features allowed for a computationally efficient cross-validation. For each reference level, the data was now randomly split by two split approaches. For the first approach, the PLSR models were trained with 75 % of randomly chosen observations and tested with the remaining 25 % of the observations. For the second approach, PLSR models were trained with the observations of 19 genotypes and tested on the observations of 11 genotypes. The maximum number of components of the PLSR model was set at 15, and the optimal number of components in the range of 1 to 15 was selected for each model individually with the "selectNcom" function. As model accuracy metric, Pearson's correlation between predictions and reference measurements was used. This procedure was repeated 100 times for each reference level, with the split into training and validation data repeated each time. The correlation coefficients and standard

deviation of the coefficients within 100 repetitions were used to characterize the quality of the PLSR models.

Importance estimation of VIs and feature types

Temporal features were extracted on the basis of multiple semiparametric and parametric methods to capture dynamics, further named feature types. In addition, temporal features were based on various zonal statistics (mean and multiple percentiles) for pixel-value aggregation, and on various VIs. To estimate the importance of the different feature types, aggregation methods and VIs, the sums of magnitudes of the relative PLSR coefficients were calculated within three reference classes. The classes were generically defined for the three reference types as early, when reference levels, *i.e.* values that the different reference types could assume, ranged from 10 to 25, intermediate for values 40 to 70, and late for values 85 to 100. To achieve robust importance estimates, coefficients were summed up for the different reference values of the reference classes for all iterations of cross-validation and the different groups of comparison. Comparison groups were either feature types, aggregation methods or VIs. $\beta_{rel,sum,ref.class,group}$ was the sum of the 100 relative PLSR coefficients $\beta_{rel,i}$ of the 100 repetitions k of cross-validation (only 75 %/25 % train/test data split) for all reference levels j within one reference class, and one comparison group,

$$\beta_{rel,sum,ref.class,group} = \sum_{j=1}^{n} \sum_{k=1}^{100} \sum_{i=1}^{200} \beta_{rel,i,j;k}, \tag{4.43}$$

where $\beta_{rel,i,j;k}$ was the relative PLSR coefficient of the i^{th} out of 200 temporal features, of the k^{th} out of 100 repetitions of the j^{th} reference level within a reference class. Finally, coefficient sums were normalized to the range from 0 to 1.

4.2.11 Heritability of predicted values

In addition to Pearson's correlation, heritability of PLSR-predicted values was calculated as quality criterion, using the R package "SpATS" (Rodríguez-Álvarez et al., 2018). This package allows to provide information on the location of measurements within the experiments as row and column coordinates to a mixed model , which are then used for a spatial correction. After spatial correction, the generalized heritability according to Oakey et al. (2006) could be calculated.

In generalized heritability, the effective dimensions ED_g are divided by the difference between the number of genotypes m_g and the number of zero eigenvalues ζ_g:

$$H^2_{genral.} = \frac{ED_g}{(m_g - \zeta_g)}, \tag{4.44}$$

with

$$ED_g = (m_g - 1)\frac{\sigma_g^2}{(\sigma_g^2 + \frac{\sigma_e^2}{r})}. \tag{4.45}$$

Heritability was calculated for every tenth iteration in the previous cross-validation for each year individually.

TABLE 4.4: Equations used for cost estimations. The cost of the methods was calculated by summing the equations in this table as indicated with × in the "Method" columns. The terms used in the equations are described in Table 4.5.

Description	Equation	Method			
		Visual Rating	Drone RGB	Drone Multi-spectral	Pheno-Cam
Staff cost round trips	$n_{measurement} \cdot Cost_{rating,staff} \cdot (2 * t_{drive}) \cdot n_{days,rating}$	×	-	-	-
Vehicle cost round trips	$n_{measurement} \cdot Cost^{\text{-}dist} \cdot dist^{\text{-}t} \cdot (2 * (t_{drive} - 0.4)) \cdot n_{days,rating}$	×	-	-	-
Staff cost rating	$n_{measurement} \cdot Cost_{rating,staff} \cdot t_{rating}$	×	-	-	-
Cost drone and sensor	$Cost_{drone,sensor}$	-	×	-	-
Initial processing cost	$n_{seasons} \cdot (Cost^{t}_{tech,staff} + Cost^{t}_{comput.}) \cdot t_{proc.init.}$	-	×	×	×
Staff cost round trips	$n_{measurement} \cdot Cost^{t}_{tech,staff} \cdot (2 * t_{drive})$	-	×	×	-
Vehicle cost round trips	$n_{measurement} \cdot Cost^{\text{-}dist} \cdot (2 * (t_{drive} - 0.4)) \cdot dist^{\text{-}t}$	-	×	×	-
Staff cost drone piloting	$n_{measurement} \cdot Cost^{t}_{tech,staff} \cdot t_{flight}$	-	×	×	-
Storage cost images	$n_{measurement} \cdot Cost^{GB}_{storage} \cdot Size_{data,images}$	-	×	×	-
Storage cost photogrammetry	$n_{measurement} \cdot Cost^{GB}_{storage} \cdot Size_{data,photogrammetry}$	-	×	×	-
Cost image handling	$n_{measurement} \cdot (Cost^{t}_{tech,staff} + Cost^{t}_{comput.}) \cdot t_{proc.}$	-	×	×	-
Computation cost processing	$n_{measurement} \cdot Cost^{t}_{comput.} \cdot t_{comput.}$	-	×	×	-
Cost drone	$Cost_{drone}$	-	-	×	-
Cost drone sensor	$Cost_{sesnor}$	-	-	×	-
Cost field masts	$Cost_{masts}$	-	-	-	×
Cost time lapse cameras	$Cost_{cameras}$	-	-	-	×
Initial processing cost	$n_{seasons} \cdot (Cost^{t}_{tech,staff} + Cost^{t}_{comput.}) \cdot t_{proc.Init}$	-	-	-	×
Staff cost round trips	$n_{seasons} \cdot n_{visits,PhenoCam} \cdot n_{persons} \cdot Cost^{t}_{tech,staff} \cdot (2 * t_{drive})$	-	-	-	×
Staff cost vehicle	$n_{seasons} \cdot n_{visits,PhenoCam} \cdot Cost^{\text{-}dist} \cdot (2 * (t_{drive} - 0.4)) \cdot dist^{\text{-}t}$	-	-	-	×
Staff cost setup & dismounting	$n_{seasons} \cdot n_{visits,PhenoCam} \cdot Cost^{t}_{tech,staff} \cdot t_{setup,dismounting} \cdot n_{person}$	-	-	-	×
Storage cost images	$n_{measurement} \cdot n_{cameras} \cdot 35_{images} \cdot Cost^{GB}_{storage} \cdot Size_{data}$	-	-	-	×
Cost image handling	$n_{measurement} \cdot n_{cameras} \cdot 35_{images} \cdot (Cost^{t}_{tech,staff} + Cost^{t}_{comput.}) \cdot t_{proc.,image}$	-	-	-	×
Computation cost processing	$n_{measurement} \cdot n_{cameras} \cdot 35_{images} \cdot Cost^{t}_{comput.} \cdot t_{comput.}$	-	-	-	×

4.2.12 Method cost comparison

For a schematic comparison of the economic cost of the different methods (Fig. 4.1e), different cost components were estimated based on personal experience. The cost components were, *e.g.*, material costs, staff cost, operation / processing cost, transportation cost, but also continuous data storage costs (*e.g.* Q. Huang et al., 2024; Marinello, 2023). A detailed listing of the components is shown in Table 4.4. The values of the components are dependent on the number of measurements $n_{measurement}$ and the costs were estimated for the four methods "Visual Rating", "Drone RGB", "Drone Multispectral" and "PhenoCam". The total costs for one method correspond to the sum of the different components applicable for the different methods, as indicated by × in the "Methods" columns in Table 4.4. 15 measurements were assumed to correspond to one season. Some costs were associated with field visits, which were necessary for every measurement of "Visual Rating", "Drone RGB", "Drone Multispectral", but just twice for the "PhenoCam" method for setup and dismounting. 35 PhenoCam images were taken each week. Once a week was about the average measurement frequency of visual ratings and drone flights, although this frequency can vary from three times a week to biweekly, depending on the phenological stage. Thus, 35 PhenoCam images were assumed to correspond to one measurement of the other methods. For each measurement or field visit, it was assumed that two times 0.4 h would be needed to load and unload the equipment to/from the car for each round trip. Otherwise, a travel speed of $80\,\mathrm{km\,h^{-1}}$ was assumed, which was relevant for the calculation of the travel costs, depending on the distance covered. Calculations were conducted for two scenarios with 700 plots and 1400 plots, respectively, and three different

TABLE 4.5: Explanation of terms in cost estimation equations of Table 4.4. The terms are grouped by methods. Terms in method "Universal" are used in two or more methods. Where applicable, values used in the cost estimation are provided for two scenarios (700 plots & 1'400 plots) and different methods. Values are in Swiss francs (CHF). In December 2024, one CHF corresponded to 1.06 € and 1.11 $ (`www.xe.com`).

Method	**Term**	**Description**	**Value (if fixed)**	
			700 plots	1400 plots
Universal	$n_{measurements}$	Number of measurements.	-	-
	$n_{seasons}$	Number of seasons, corresponds to $n_{measurements}$ divided by 15 and rounded down.	-	-
	$dist^{-t}$	Distance covered within one hour of drive.	80 km h^{-1}	80 km h^{-1}
	t_{drive}	Time to get to the experimental site (one-way). For the time of round trips, this time is multiplied by two.	-	-
	$(t_{drive} - 0.4)$	The negative offset of 0.4 h penalizes the distance covered during the first hour of driving for loading and unloading the equipment.	-	-
	$Cost^{-dist}$	Cost per km driven.	0.6 CHF km^{-1}	0.6 CHF km^{-1}
	$Cost^{-t}_{tech,staff}$	Cost of one technical staff for one hour.	78 CHF h^{-1}	78 CHF h^{-1}
	$Cost^{-t}_{comput.}$	Cost of one hour computing.	3 CHF h^{-1}	3 CHF h^{-1}
	$Cost^{-GB}_{storage}$	Cost to store one GB for 10 years.	2.76 CHF GB^{-1}	2.76 CHF GB^{-1}
	$Size_{data,images}$	Size of image data per measurement generated with drones or of single images for the PhenoCams.	5 GB*/17 GB**/0.016 GB***	10 GB*/34 GB**/0.016 GB***
	$Size_{data,photogrammetry}$	Just applies to drone flights: Size of photogrammetric projects of the drone measurements per measurement.	11 GB*/18 GB**	22 GB*/36 GB**
	$t_{proc.init.}$	Time for initial processing, e.g. creating georeferenced image masks, setting up the processing pipeline etc.	7 h*/8 h**/16 h***	14 h*/16 h**/32 h***
	$t_{proc.}$	Processing time after initial processing.	1 h*/3 h**/1 s***	2 h*/6 h**/1 s***
	$t_{comput.}$	Computation time of data per measurement (drones) or per image (PhenoCam).	2 h*/5 h**/1 min***	4 h*/10 h**/1 min***
	t_{flight}	Time needed to cover 700 and 1'400 plots respectively with droneflights.	2 h	4 h
$Cost_{rating}$	$n_{days,rating}$	Number of days to complete the rating for one measurement, but not necessarily full days. Determines the number of round trips per measurement.	1	2
	$Cost_{rating,staff}$	Cost of one rating staff for one hour.	61 CHF h^{-1}	61 CHF h^{-1}
	t_{rating}	Time needed for one measurement of all plots.	5 h	10 h
$Cost_{RGB,drone}$	$Cost_{drone,sensor}$	Cost of drone with integrated camera system, e.g. DJI Mavic 3 Pro (SZ DJI Technology Co. Ltd., China).	1'700 CHF	1'700 CHF
$Cost_{Multispec.,drone}$	$Cost_{drone}$	Cost of drone that can carry a Micasense RedEdge-MX DUAL sensor, e.g. DJI Matrice 350 RTK (SZ DJI Technology Co. Ltd., China).	9'300 CHF	9'300 CHF
	$Cost_{sesnor}$	Cost of Micasense RedEdge-MX DUAL sensor (MicaSense Inc., Seattle, Washington, USA).	9'500 CHF	9'500 CHF
$Cost_{PhenoCam}$	$n_{days,setup,dismounting}$	Number of days to either set up or dismounting the PhenoCams.	1	2
	$n_{visits,PhenoCam}$	Number of times to visit the experimental site for the setup and for the dismounting of the PhenoCams.	2	2
	$n_{cameras}$	Number of cameras (two cameras per system, four per mast in our setup).	8	16
	$n_{persons}$	Two people are needed for the setup and dismounting of the PhenoCams.	2	2
	$Cost_{masts}$	Cost of two or four Teksam field masts.	24'000 CHF	48'000 CHF
	$Cost_{cameras}$	Cost of 4 or 8 autonomous time lapse camera systems (8 and 16 cameras in total), e.g. Enlaps Tikee 3 Pro+. (Enlaps SAS, Montbonnot-Saint-Martin, France)	6'800 CHF	13'600 CHF
	35_{images}	35 images of the PhenoCam correspond to one measurement of the other methods.	-	-
	$t_{proc.,image}$	Time needed to handle and process one image.	1 s	1 s
	$t_{setup,dismounting}$	Time needed to setup or dismounting the cameras.	6 h	12 h

*RGB drone, **Multipsectral drone, ***PhenoCam

distances of the experiments of the research station, which changed the equipment needed and the time necessary for traveling and rating or flying. The terms used and assumed values for the two plot number scenarios are shown in Table 4.5. Calculations were performed for 1 to 90 measurements or 0 to 6 seasons. The storage cost values were estimated to be 0.023 CHF GB/month based on the official price listing of standard google cloud storage on Swiss-based servers (Google, 2024), assuming a storage duration of 10 years.

4.2.13 Weather data recording

The air temperature and daily precipitation were obtained from a Meteoswiss (Federal Office of Meteorology and Climatology, `https://www.meteoswiss.admin.ch`) weather station which was located about 800 m from the experimental site at Changins [46°24′3.7″N 6°13′39.6″E, 458 m.a.s.l., WGS 84].

4.3 Results

4.3.1 Mast setup

PhenoCams were installed at 12 m above the ground, this was a compromise since the first tests started at 20 m, but after observing the behavior of the masts under windy conditions, the masts were lowered to 12 m, which allowed a stable and continuous operation throughout the season. The ropes loosened over time due to constant back and forth in the wind and had to be tightened manually from time to time.

The footprint of the mast ropes occupied a circular area with a diameter of 20 m. This was a significant obstacle to agricultural treatment operations in the field. In addition, experiments could not be carried out in the area between the anchorages of the stabilizing ropes.

The calibration panels needed to be cleaned regularly, as they were visited by animals (Fig. S3.3), which left footprints, or were polluted by splashing water during rains. The field fauna also nibbled on the ropes of the masts. Although the wheat and weeds around the calibration panels were generously removed, panels were shaded by the plants, especially in spring in the morning or evening.

To protect the camera systems from birds, spikes had to be installed to prevent them from using the cameras as a vantage point for hunting mice. Without spikes, bird droppings would have covered the solar panels and the camera lens itself.

4.3.2 Mask creation

As a result of masts slightly shaking in the wind throughout the season, of viewing geometry changes due to tightening the ropes of the mast and of the wheat growing, the plot masks needed to be adjusted for images throughout the season. This step was very time-consuming, as it had to be done for four cameras in each year. An approach using scale-invariant feature transformation (SIFT; Lowe, 2004) in OpenCV (Bradski and Kaehler, 2000) failed due to incremental error propagation, resulting in inaccurate plot shapes after about ten images.

4.3.3 Index dynamics from different sensors and image formats

27 RGB VIs were calculated from PhenoCam and drone-based data. For drone-based data, 12 additional multispectral VIs were calculated. Example data of two RGB VIs, VARI and $\mathrm{ExGR}_{\mathrm{Zhang}}$, is shown in Fig. 4.5, based on JPEG DNs and JPEG-based reflectance. For PhenoCam data, the VI based on JPEG DNs (*e.g.* Figs. 4.5a & 4.5b) showed a higher variability compared to reflectance-based VIs and the latter appeared smoothed, with this effect

being stronger for ExGR $_{Zhang}$ than for VARI. However, in general, the temporal dynamics appeared very similar between JPEG DN and JPEG reflectance data. The same VIs based on DNG raw DNs instead of JPEG DNs could look different with more variability and less pronounced temporal dynamics (Fig. S3.4).

VIs based on drone data appeared to be smoother than the ones derived from JPEG PhenoCam data, but to have similar temporal dynamics (*e.g.* Figs. 4.5c & 4.5d).

However, the patterns described before were not found for all VIs. For example, IKAW VI dynamics were more similar between PhenoCam DNG raw data and drone data than between PhenoCam JPEG and drone data. In addition, IKAW showed less pronounced temporal dynamics, especially for JPEG-based VIs (Fig. S3.5).

The maintenance of PhenoCams resulted in small changes in PhenoCam orientation. Consequently, the VI time series of plots at image borders could be interrupted at maintenance. (*e.g.* Plot_102 Fig. 4.5).

4.3.4 Temporal feature count overview

From the VI data, 758 to 2217 different temporal features could be derived depending on the different sensors, image formats and data treatments (Table 4.6). Most temporal features could be found for the of observations of the "Drone Multispectral" data, followed by "Drone RGB", which represents a subset of the "Drone Multispectral" set. This was followed by the PhenoCam-based JPEG and finally the DNG methods. For the latter two, more temporal features were found for the Reflectance option than for the DN option.

Only features that were available for a large proportion of observations were included in PLSR. Automated feature extraction was more effective on JPEG data than on DNG data, which led to a decrease in features from JPEG to DNG.

For drone-based methods, 810 observations were available for PLSR modeling as in each of the three years, 270 plots were observed. In contrast, plots could appear on multiple PhenoCams in the same season, and, on average, each plot was recorded by 2.6 cameras, although plots on the edges of the field were just recorded from one camera. Camera-plot combinations with 1500 or more missing temporal features were excluded. 2101 observations were available for PhenoCam JPEG data and 2092 for DNG.

Table 4.6: Number of temporal features and observations for different sensors, image formats and data treatments

Method	No. of temporal features	No. of observations
JPEG DN	1005	2101
JPEG Reflectance	1226	2101
DNG DN	758	2092
DNG Reflectance	837	2092
Drone RGB	1452	810
Drone Multispectral	2217	810

Using PLSR In a first feature selection round, 200 temporal features were selected for each field reference type and reference level. If using less than 200 features, correlations between PLSR prediction and field reference levels started to decrease (Fig. S3.11).

4.3.5 Comparison of PLSR prediction performance of different methods

Eight methods to predict plant development, depending on the different sensors, image formats and data treatments, were compared individually by Pearson's correlation for the different reference rating types. The mean correlations for all cross-validation data with reference rating values across 100 iterations for each reference level were summarized in boxplots (Fig. 4.6).

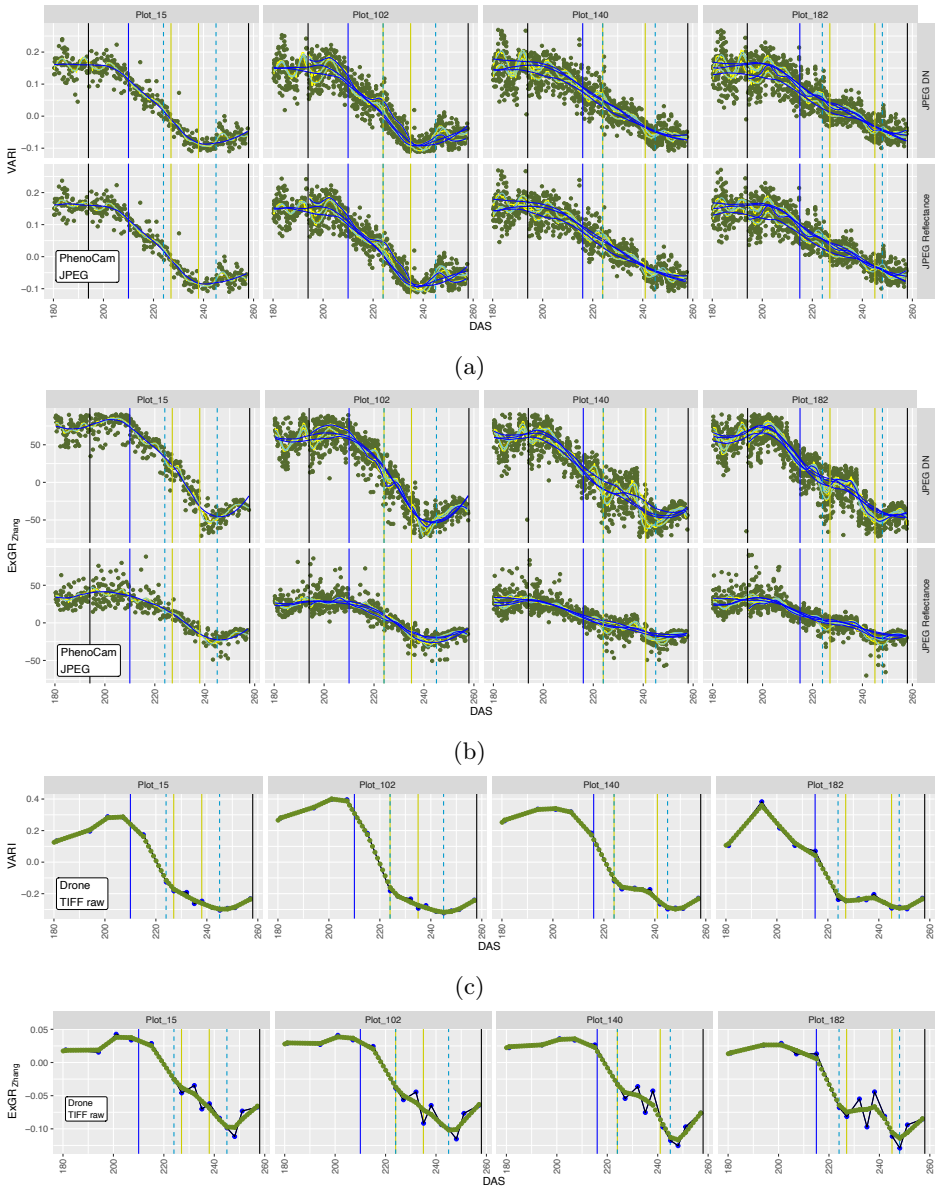
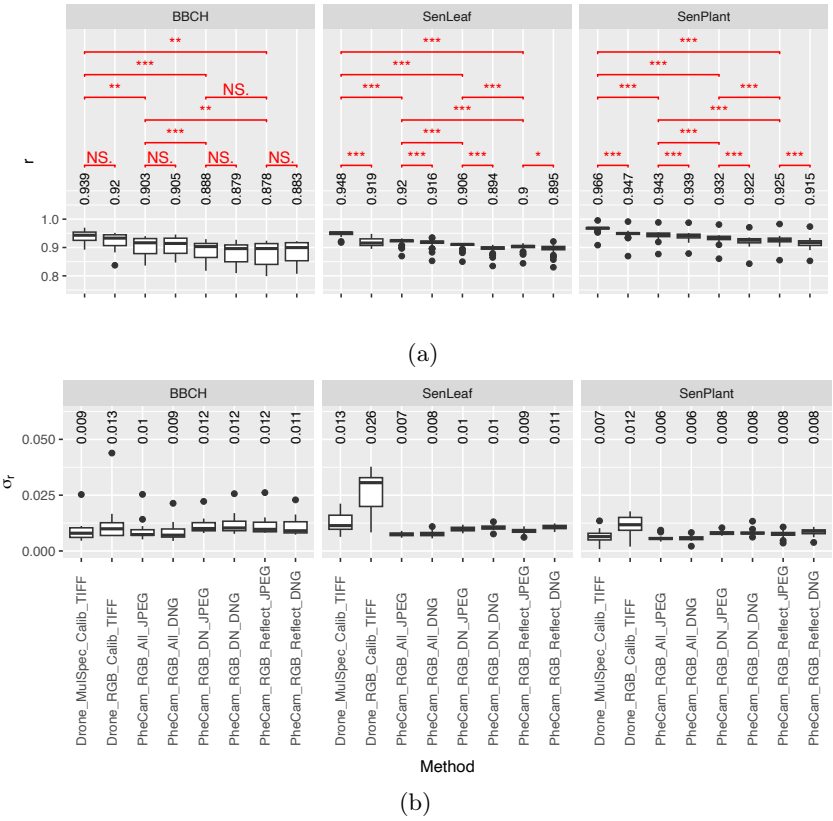


FIGURE 4.5: Example of VI data derived from PhenoCams (a & b) and a drone-based camera (c & d) for two VIs (VARI and $ExGR_{Zhang}$) and four plots during the seasons 2022. The temporal axis is in days after sowing (DAS). For the PhenoCam data, greenish points in the PhenoCam image are initial VI values and lines represent smoothed data of different smoothing methods (dark yellow: rolling mean; dark blue: loess smoothing; yellow: Savitzky–Golay; light blue: spline smoothing). In plots with multiple lines of the same color, multiple cameras observed the same plot. Data is shown for unprocessed data ("DN") and for calculated reflectance values . For the drone data, the initial VI values are blue dots. Greenish lines represent a smoothed spline interpolated for a daily temporal resolution. The colored vertical lines indicate specific levels of visual field reference ratings as observed on the respective plots: Solid blue line indicates the heading date (BBCH 59), the dashed blue lines indicate plant senescence levels of 10 % and 90 % respectively. The yellow lines correspond to flag leaf senescence at 10 % and 90 %. The black lines toward the end mark the harvest date. The first vertical black line for the PhenoCam data shows the date of PhenoCam maintenance.

Six methods were based on PhenoCam data, three each for JPEG and DNG data formats. For both formats, a method was based on calibrated reflectance data, one on DN and one on both data types. Two methods were based on calibrated drone data in the multispectral and RGB color space.



(a)

(b)

FIGURE 4.6: Overview on Pearson's correlation of PhenoCam- and drone-based predictions of timing of phenology (BBCH), flag leaf senescence (SenLeaf) and plant senescence (SenPlant) with field reference measurements. Values above boxplots indicate mean values. (a) Mean correlations for all cross validation data (only 75 %/25 % train/test data split in cross-validation) with reference rating values across 100 iterations for each reference level. The mean standard deviations of these correlations across 100 iterations for each reference level σ_r is shown in (b). In the method names of the x-axis labels, "PheCam" and "Drone" indicate the platform of image acquisition, "RGB" and "MulSpec" the color space of the features. "Reflect" indicates that only calibrated reflectance data (reference panels) was used as opposed to "DN" for the use of digital numbers. "All" means that "Reflectance" and "DN" data was used. "JPEG" and "TIFF" indicate the image data format used. Pairwise t-tests were applied to examine whether the different methods produced significantly different results. Pairing was by reference levels of the three reference types. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$.

Models based on multispectral drone data were best correlated with reference levels for all reference rating types (Fig. 4.6a), followed by RGB drone data. PhenoCam methods showed slightly weaker correlations compared to drone-based VIs, when "All' data types (DN and Reflect) were used, but almost no difference between JPEG and DNG. The JPEG DN method was slightly inferior to using both data types but superior to the remaining three methods "Reflect_JPEG", "DNG_DN" and "Reflect_DNG".

The standard deviation of these correlations σ_r (Fig. 4.6b) revealed that higher correlations were also more consistent as σ_r was lower for higher correlations, except for the correlation between SenLeaf reference type and drone-based RGB methods.

4.3.6 Detailed comparison of selected methods

As the JPEG based PhenoCam methods seemed to perform slightly better than the DNG based models, they were compared to drone-based methods in more detail. When correlating the PLSR predictions for all reference levels of the different reference types (Fig. 4.7a), the correlations were very strong ($r > 0.8$) and even stronger than 0.9 for the later levels of the BBCH scale and the intermediate levels of SenLeaf and SenPlant in 100-times repeated cross validation. Early BBCH stages and early as well as late senescence stages showed weaker correlations in general. The correlations were consistently higher for drone-based methods, except for later stages for SenLeaf, where drone-based predictions in the RGB color space showed high variability. When PLSR models trained on the training data were applied to predict all observations, the correlations were stronger and more consistent than when only predicting and correlating the test data set in cross-validation. When using the 75 %/25 % train/test data split in cross-validation, correlations were slightly higher (0.02 on average) compared to 19 genotypes/11 genotypes train/test cross validation with standard deviation increasing by only 0.01 between the two. Therefore, the remaining analysis was just conducted on cross-validation data based on a 75 %/25 % data-split.

When correlating separately for the three years, the correlations were weaker in general, but the trends remained similar. Correlations were weak to very weak for early SenPlant and weak to strong for late stages of all reference types. For SenLeaf and early stages of SenPlant, correlations were weaker in 2023 than in the other two years. SenPlant showed weaker overall correlations in 2022. Correlation of BBCH did not show a distinct year-wise pattern except for weak correlations for the latest BBCH levels in 2022.

The root mean square error (RMSE) was similarly low for both senescene rating types in 2021 and 2023 with slightly higher RMSE for earlier stages in 2022 (Fig. S3.7). As for correlations, no distinct year-wise pattern was found for RMSE of BBCH.

To better understand the reason behind varying correlations in dependence of the different years, the temporal density of the reference measurements was examined (Fig. 4.8). Later stages of BBCH occurred in a short period in 2022 compared to the other tow years. Stages of senescence and especially SenLeaf occurred in a shorter interval for most reference levels in 2023.

4.3.7 VI and feature type importance

Normalized relative PLSR coefficient sums $norm.\beta_{rel,sum}$ were analyzed to determine the importance, *i.e.* predictiveness of VIs and feature types. For PhenoCam "JPEG DN" format (Fig. 4.9a), the VIs of the ExGR type and especially $\text{ExGR}_{\text{Zhang}}$ were the dominant features. For BBCH and early SenLeaf, VARI was also important. For BBCH, GCC played a crucial role too and the G_R_Ratio for SenLef. In SenPlant, GLI was an important feature. In contrast, for the PhenoCam "DNG raw DN" format (Fig. 4.9c), $\text{ExGR}_{\text{Zhang}}$, although still important, was just the dominant feature for late BBCH and intermediate to late SenPlant. Otherwise, GCC and RGBVI were important for BBCH. ExR_{g}, MGRVI, MNVI, and RGBVI were predictive for SenLeaf and ExR_{g} and MNVI also for SenPlant. The predictiveness of VIs varied from early to late reference classes.

The local maximum of the first derivative of the VIs $\text{D1}_{\text{LocMax\_1}}$ was an important feature type in the prediction of all reference types and values for "JPEG DN" and "DNG raw DN" data (Fig. 4.9b & 4.9d), while the first local minimum of the second derivative $\text{D2}_{\text{LocMin\_1}}$ and the first local maximum of the third derivative $\text{D3}_{\text{LocMax\_1}}$ were increasingly important from early to late reference stages. For other feature types, a similar but less pronounced trend from early to late could be observed. For the different data aggregation methods no

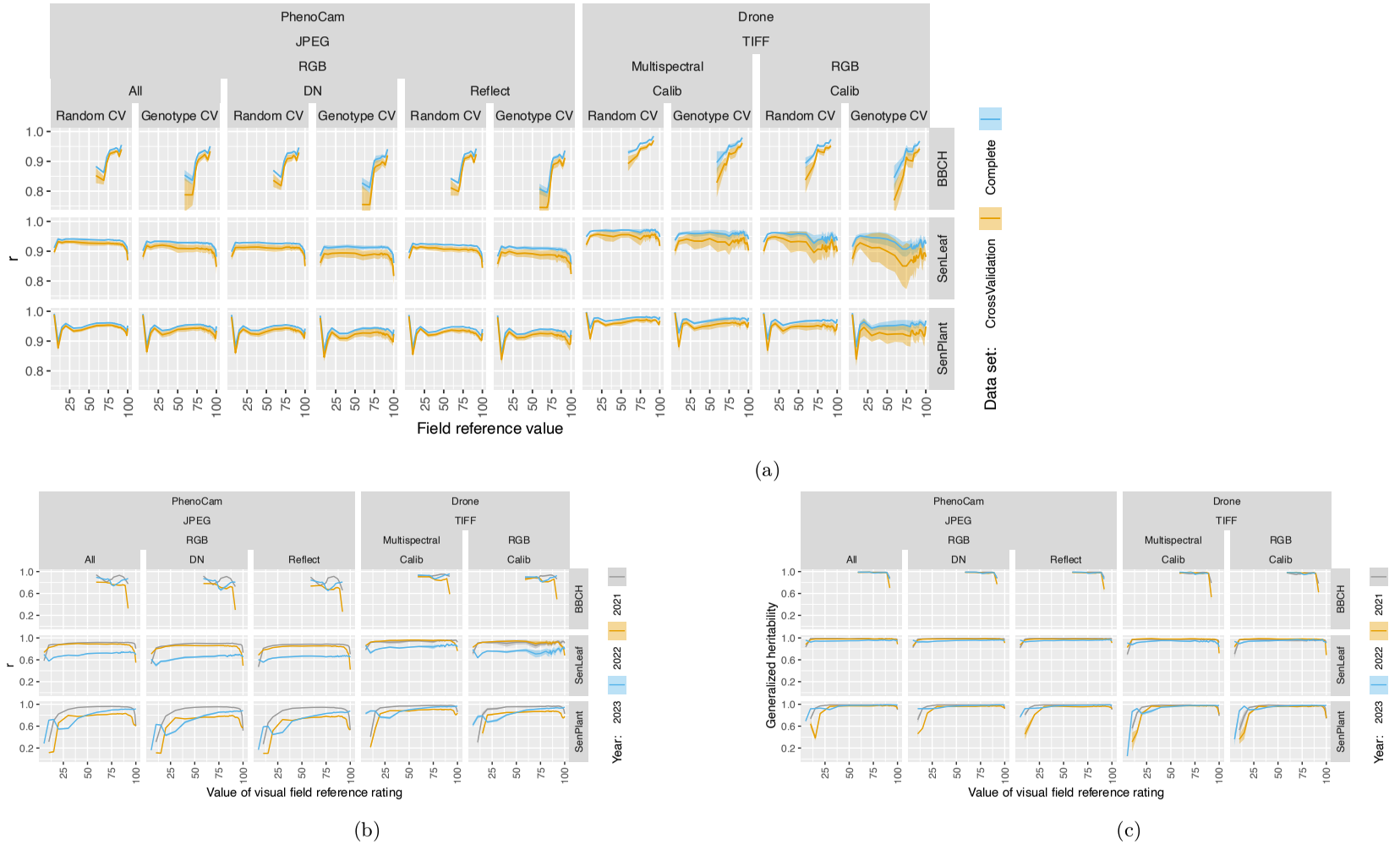


FIGURE 4.7: Quality metrics of PLSR models. (a) The Pearsons's correlation between PLSR-based predictions and field reference ratings. Data of the different sensors, processing methods and cross-validation (CV) types are arranged in columns. Random CV refers to a 75 %/25 % train/test data split. Genotype CV refers to a 19/11 genotype train-test data split. The rows represent the three types of field reference measurements. Yellow indicates the correlations only for test data from the test-training split, and blue indicates the correlation for all data (train + test). The shaded areas indicated the mean ± standard deviation in 100 repetitions of cross-validation while the lines are the means. (b) Correlations based on all data but analyzed for individual years. (c) Heritability of PLSR-based predictions for 10 repetitions per reference level based on PLSR predictions for all data as input in the SpATS models. (b & c) are only shown for 75 %/25 % train/test data split in cross-validation.

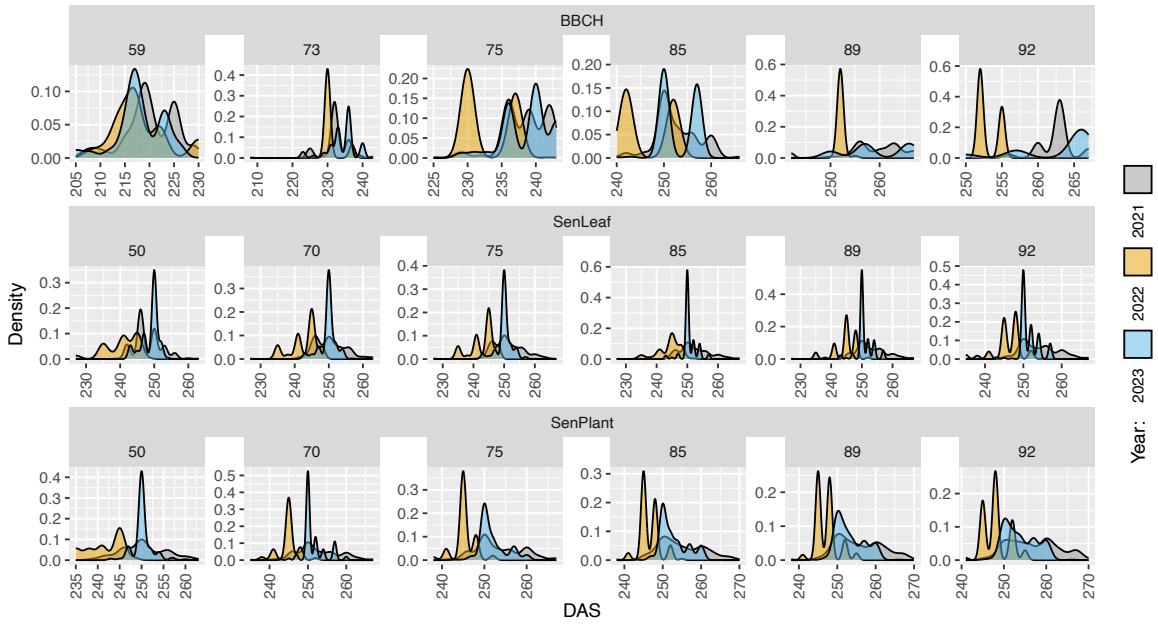


FIGURE 4.8: Temporal density of interpolated reference observations of all three reference rating types by year for generically selected reference levels. The temporal axis is in days after sowing (DAS).

clear trend could be found but the 50^{th} percentile and/or the mean were generally among the most important aggregation methods (Fig. S3.10).

For drone-based data from the RGB colorspace (Fig. S3.8a), BCC, GCC, IKAW, MGRVI, MNVI, RCC and RGBVI were especially predictive and the same feature types ($D1_{\text{LocMax\_1}}$, $D2_{\text{LocMin\_1}}$, $D3_{\text{LocMax\_1}}$) but also $D4_{\text{LocMax\_1}}$ and the nonparametric temporal features $\text{Spline}_{0.02}$ and $\text{Spline}_{0.1}$ were important (Fig. S3.8b).

In the multispectral color space (Fig. S3.9a), in addition to BCC and IKAW, mostly multispectral VIs became dominant, such as DVI, NDRE, NDVI, PSRI7_{717}, PSRI7_{740} and SAVI. Most predictive feature types were similar to drone-based RGB data (Fig. S3.9b).

4.3.8 Method cost comparison

The cost estimates for the different methods were highly dependent on the number of plots observed, the number of measurements and seasons carried out, and the distance of the experimental site from the research station.

The visual field reference rating, closely followed by the RGB drone method, had the lowest initial costs (Figs. 4.10a & 4.10b). The costs then increased almost linearly with the number of measurements. For the 700 plots scenario, visual rating was cheaper than the RGB drone for all driving distances and this difference increased with the number of measurements. In the 1400 plots scenario, for 0.5 h of driving, the visual rating was slightly cheaper than the RGB drone, but for 1.5 - 3 h of driving time, the RGB drone method was cheaper and this difference increased with the number of measurements. The method with the next-highest initial cost was multispectral drones, for which the costs increased almost linearly with the number of measurements. The marginal costs for additional measurements were higher for multispectral measurement, compared to the RGB drone method and the visual rating (Table. 4.7). The PhenoCam method had the highest initial cost. In contrast, additional measurements had only a slight effect on costs. New costs arose, above all, for the setup and dismantling of the camera systems. However, these seasonal costs were higher than with the other methods. For non-PhenoCam methods, the costs were relatively low at the beginning of a new season. Only

FIGURE 4.9: Importance of VIs and temporal feature types in PLSR models based on PhenoCam derived JPEG (a & b) and DNG raw (c & d) images. Data is arranged by reference types (Phenology: BBCH; Flag leaf senescence: SenLeaf; Plant senescence: SenPlant), and reference classes (Early, Intermediate, Late). (a & c) Importance of the different VIs as the normalized sum of the relative PLSR coefficients $norm.\beta_{rel,sum}$ over 100 repetitions of cross-validation. (b & b) Importance of the different feature types as $norm.\beta_{rel,sum}$ over the 100 repetitions of cross validation. D1 - D4 indicate the first four derivatives of the Gompertz function. LocMax and LocMin refer to local maxima and minima, and the number after LocMax and LocMin represents the order along increasing DAS, when there were multiple local maxima/minima of the same type. The remaining features correspond to nonparametric temporal features of the smoothing types loess, rolling mean, Savitzky-Golay and spline.

new flight plans had to be created and new plot masks had to be drawn for the analysis of the images of drone-based methods. The different travel times had a relatively little effect on the PhenoCam method, while they led to significantly different costs for drone flights and visual ratings. In the 700 plots scenario, at around 30 measurements or 3 seasons, the PhenoCams became cheaper compared to multispectral drones, and around 90 flights or 6 seasons, they became cheaper compared to RGB drones except for the 0.5 h driving distance. These general patterns were similar for the 1400 plots scenario, but it took more measurements until the PhenoCam method became cheaper than drone-based approaches.

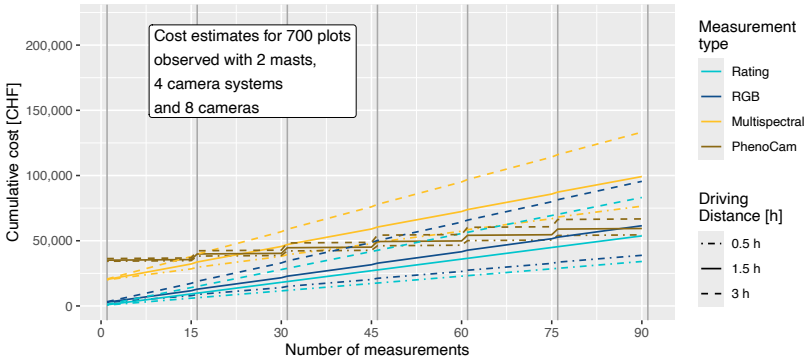
Looking at cost structures, transport was an important cost factor from 1.5 h driving time onward and the most significant cost driver for visual field reference ratings, especially in the 1400 plot scenario, as the field needed to be visited twice to complete the rating of all plots. Another scenario would be to stay overnight in a hotel, which would also lead to higher costs, but this scenario was not included here. The sum of measurements cost (*i.e.* drone piloting) and the image processing cost for the RGB drone was similar compared to the measurement cost of visual ratings in the 700 plots scenario. For 1400 plots, the RGB drone was cheaper for 1.5 h or more driving time, as it could cover more plots in a shorter period without the need of an overnight stay or a double visit to complete the measurements.

The multispectral drone method came with higher processing costs, and as large data volumes were produced in multispectral imaging, the storage of the images became an important cost driver in addition to higher initial material costs. It was the most expensive method after 6 seasons in all scenarios.

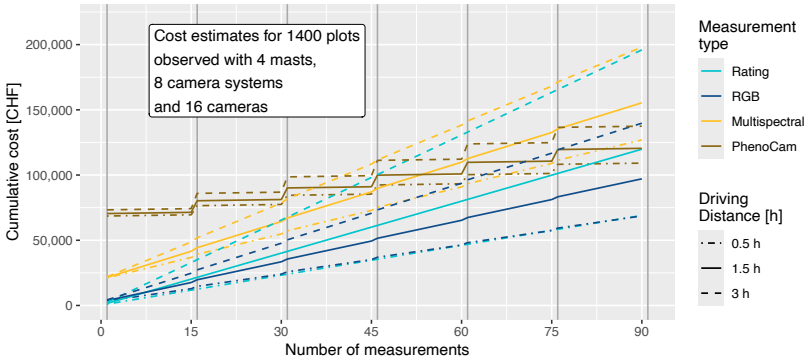
Table 4.7: Marginal cost of different methods after 90 measurements for the 700 plot and the 1400 plot scenario.

Method	**Driving Distance (h/one-way)**	**Marginal costs (CHF)**	
		700 Plots Scenario	1400 Plots Scenario
Rating	0.5	376	764
	1.5	594	1328
	3	921	2174
PhenoCam	0.5	227	456
	1.5	273	562
	3	342	721
Drone - RGB	0.5	407	732
	1.5	659	1048
	3	1037	1522
Multispectral	0.5	635	1188
	1.5	887	1504
	3	1265	1978

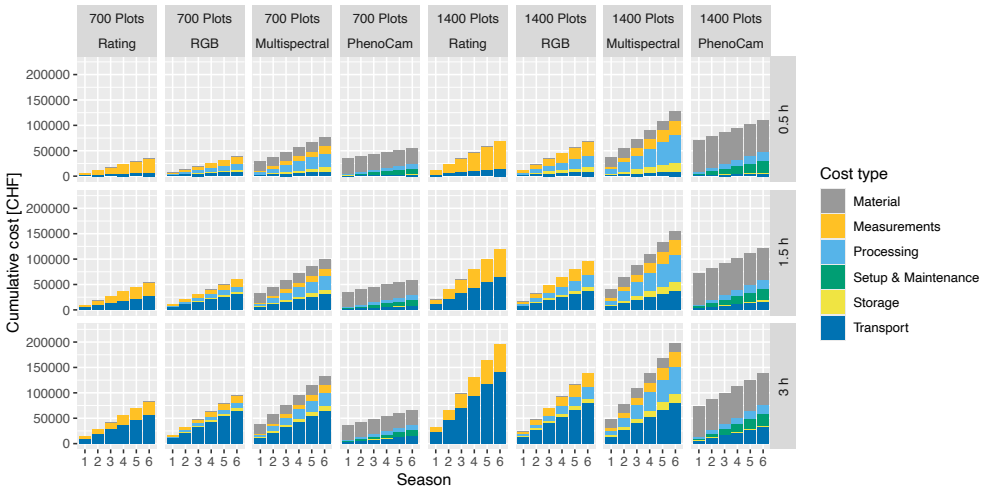
PhenoCams had by far the highest initial material costs. On the other hand, the transportation costs were low, as only two field visits were necessary for set up and dismounting, assuming no technical incidents occurred. Although each PhenoCam could shoot many images per day, the total amount of data was very manageable compared to drone measurements, especially multispectral. Even if the initial processing costs were relatively high, as plot masks had to be corrected for perspective and adapted to shaking cameras and growing vegetation, the overall processing costs remained relatively low. Thus, once the material was acquired, the cost for additional measurements was relatively cheap, making this the most economical method for scenarios of 3 h driving time after 6 seasons. In the 1.5 h driving time scenario, PhenoCam costs were comparable to RGB drones and to visual field ratings after 6 seasons. Just for the 1400 plots scenairo, PhenoCams remained more expensive compared to RGB drones. Yet, due to the low marginal costs of additional measurements (Table. 4.7), PhenoCams would



(a)



(b)



(c)

FIGURE 4.10: Overview of cost development and coast types in dependence of number of measurements and seasons. (a) & (b) show the estimated total costs with repsect to number of measurements. 15 measurements were assumed to correspond to one year, and years are marked with vertical gray lines. Line types indicate different distances of experimental site from research station in hours (one-way). Line colors indicate different methods that were compared to each other in this study. (c) summarizes these costs by cost types for one to six seasons of 15 measurements each.

become the cheapest method with additional seasons, which would finally also be the case for 1 h driving time at a high number of seasons to measure.

4.4 Discussion

4.4.1 Ability of PhenoCams to track phenology and senescence in field conditions

The PhenoCams allowed the tracking of phenology and senescence over three seasons with high reliability and genotype specificity under field conditions in real a variety testing experiment. Field phenotyping is known to be one of the most challenging phenotyping settings due to confounding effects, such as spatial and temporal variability of traits due to *e.g.*, heterogeneous field conditions, changing weather throughout a measurement campaign, or disruptive weather events, which can make data acquisition but also interpretation very challenging (*e.g.* Araus, Kefauver, et al., 2018; Aasen, Kirchgessner, et al., 2020; M. P. Reynolds, S. C. Chapman, et al., 2020). For example, in this study, shaking cameras, dust and dirt, changing illumination, non-continuous VI dynamics due to drought/rain interplay, memory restrictions of autonomous cameras, trade-offs in the experimental setup due to the needs of field operations and animal interference made it complicated to operate the cameras and analyze the images. Nevertheless, strong correlations with visual field reference ratings and high heritabilities for PhenoCam-based traits were attained in field conditions.

The quality of the predictions derived from PhenoCams was slightly inferior to drone-based predictions, and especially to multispectral predictions. Nevertheless, PhenoCams are a promising tool for the field phenotyping of dynamic traits. They allow to increase the temporal resolution of image acquisition considerably, even at remote experimental sites.

4.4.2 Quality of predictions from PhenoCam and drone setups

Comparison of PhenoCam data with drone-based multispectral data was conducted to estimate the benefit of bands from near-infrared and red-edge regions to track plant development. Comparison with drone-based RGB data allowed estimation of effects related to viewing geometries.

The better performance of the drone-based VIs compared to PhenoCams might be largely due to the near-optimal conditions under which the drone measurements took place. Plots were observed at a close to nadir view, with rather homogeneous viewing geometries, while in PhenoCams, plots were seen from extremely different viewing geometries and distances. In plots close to the PhenoCam, single plants, even plant organs, were distinguishable in a rather nadir oriented view. For the most distant plots, a single pixel corresponded to several plants and only plot-wise mean color properties could be tracked from a very lateral view on the upper part of the canopy. The timing of flying was, whenever possible, close to noon, also allowing for relatively homogeneous illumination of images between flights. This increased the signal-to-noise ratio compared to PhenoCams, where images were taken at a much higher frequency but with a large variability of illumination and viewing geometries.

PhenoCams covered many plots with two or even three cameras. Although the same plot was observed, the viewing geometry and distance from the camera were often very different, especially with regard to the path of light from the sun via the plot to the camera. The plots were subjected to a bidirectional reflectance changes (Nicodemus, 1977; Schaepman-Strub et al., 2006) for the different cameras. Nonetheless, based on a visual comparison (Figs. 4.5 & S3.5) and high heritability (Fig. 4.7c), the different cameras tracked similar VI dynamics for the same plot, as VIs normalize and reduce the effects of bidirectional reflectance changes (Aasen, Kirchgessner, et al., 2020; Sonnentag et al., 2012).

4.4.3 Quality of predictions in different years

Although the prediction correlations were lower for SenLeaf in 2023 than for the other years (Fig. 4.7b), the RMSE of the predictions was lower or similar than in the other years (Fig. S3.7). At the same time, correlations for SenLeaf in 2022 were strong despite a relatively high RMSE. This might be explained by the extended duration during which specific SenLeaf levels occurred in 2022 and short duration in 2023, as shown by the temporal density of selected rating levels (Fig. 4.8). When levels occur in a relatively short period, strong correlations are more difficult to attain and weaker for the same RMSE compared to situations with a more temporally dispersed occurrence of the same rating level.

This highlights that the quality of the method is also affected by G×E interactions, as meteorological conditions that promote rapid plant development and senescence lead to weaker correlations. That is also the rationale why data from the three years was used to train the PLSR models. The meteorological conditions and therefore the development of the plants contrasted strongly for the three years. A low predictive performance would be expected when predicting a wet year from dry years, and also the two dryer seasons had a very different phenological development.

In addition, VI dynamics are directly affected by meteorological conditions. For example, 2022 was a hot and dry year, which caused the flag leaves to roll. After rain events, the leaves were able to recover slightly, which could lead to a short-term flattening of the temporal dynamics toward maturity. This might be a valid explanation of noncontinuous trends *e.g.* for the VIs VARI, $\text{ExGR}_{\text{Zhang}}$ and NDVI in 2022, where in June, the slope of declining VIs flattened out after significant rains or even increased again (Figs. S3.6).

4.4.4 Quality of predictions in different stages

The prediction accuracy of PLSR models for the early or late phenology and senescence stages were often low. This might be related to the small phenotypic changes that these early and late stages are associated with, which might be too small to be detected from cameras at distance. In addition, in this study, multiple raters conducted the ratings over years but also within years, inevitably leading to rater bias. Later stages of phenology are tedious to track, as they require the manual inspection of some grains in each plot and are limited in precision to track small changes between rating events (Anderegg, Yu, et al., 2020). This is also true, especially for the early and late stages of senescence. Visual scoring methods are subjective and can be affected by foliar diseases, abiotic foliar damage, and other confounding influences, leading to phenotypic heterogeneity within plots (*e.g.* E. A. Chapman et al., 2021; J. T. Christopher, M. J. Christopher, et al., 2016; Kipp et al., 2014). Thus, the precision limitation of visual scoring itself is likely to limit the precision of the approach (Anderegg, Yu, et al., 2020).

Finally, later stages of phenology do not address the external phenotypes of the plants but the state of the grains, which cannot be visually seen without opening the husk. The high predictiveness of the RGB but also multispectral VIs is thus rather surprising and most likely the result of a relationship between external visual features and grain-internal processes.

4.4.5 Comparing VIs from different image formats and data treatments

Even within data from the same sensor, VIs can show very distinct patterns depending on image format and data treatment (*e.g.* Fig. S3.5). JPEG DN method was shown to be superior to JPEG reflectance and both DNG raw methods in this study while inferior to the combined use of DN and reflectance in the same PLSR analysis, though the differences were rather minor (*e.g.* Fig. 4.6).

JPEG format images are derived from raw images after Bayer matrix decomposition by multiple transformations, such as white balance application, gamma correction, and dynamic range compression. These transformations increase the contrast in images and lead to a visually enhanced nonlinear representation of light intensities. The nonlinear nature of these transformations also leads to changes in VI dynamics. With gamma correction exponents <1, which make images appear brighter, changes for high values in the linear format represent smaller changes for nonlinear JPEG. For lower values, changes in the linear DNG raw format lead to larger changes in nonlinear JPEG. With ratios including high/low values and especially both, the ratios can look very different between linear and nonlinear formats. If VI formulas include sums/differences, the VIs can even change from positive to negative or vice versa. Indices like VARI, ExGR did not seem to be affected a lot, while indices like IKAW did. These transformations thus can amplify VI dynamics, which might be an explanation why automated feature extraction was more effective on JPEG data than on DNG raw data in this study.

4.4.6 Practical challengers of calibration panels in field conditions

The continuous use of reflectance calibration panels in a field setting is prone to disturbances. Even when regularly cleaning calibration panels, it was inevitable that their reflectance changed over time due to dirt, sun-bleaching, and a change of wet and dry conditions. In addition, they were sometimes shaded by surrounding plants or weeds that grew between them, which could not be immediately removed.

Despite these adverse influences and a reduced temporal resolution of reflectance-based VI values, smooth reflectance patterns were achieved in this study (Figs. 4.5 & S3.5). Yet, the calibration panel setup could be improved. The panels could be installed higher above the ground, but they would still need regular cleaning, which would undermine the main benefit of PhenoCams: *i.e.*, the reduced need for frequent field visits. Thus, using uncalibrated JPEG might be the sweet spot between quality and effort when tracking plant development in a lean-phenotyping approach with day-to-day applicability. Although JPEG-based VIs are not a representation of real physical property such as reflectance, they were shown to be as predictive as reflectance based on a linear DNG raw format. However, 8-bit JPEG produces far less data than linear 16-bit raw formats and is a widely used image standard that can be easily handled and visualized. In summary, the use of calibration panels, which are expensive and time-consuming in application, can be omitted without a major loss of predictive quality.

4.4.7 Comparing color spaces and RGB sensors

Cao et al. (2021) showed the superiority of multispectral VIs over RGB VIs and that they could reveal more detailed phenotype changes but were also more sensitive to rainfall than RGB VIs, which also seemed to be the case in our study (*e.g.* Fig. S3.6). In the study at hand, the drone-based PLSR prediction from RGB VIs was often almost as strongly correlated with visual reference ratings as those from multispectral measurements, which is in line with Cao et al., 2021, who showed that multispectral and RGB VIs together only slightly outperformed the pure RGB VIs. However, the robustness of the prediction (σ_r), seemed more affected, especially for later stages of SenLeaf.

This study only used RGB-bands from a narrowband multispectral camera for RGB-based VIs, but Cao et al. (2021) compared a low-cost integrated RGB camera of a DJI Phantom 4 drone with more expensive MicaSense multispectral narrowband sensors for their ability to track stay-green phenotypes in wheat. RGB VIs based on the cheaper sensor better classified senescence types than the RGB VIs from the more expensive narrowband sensor in their study. Thus, it is highly likely that the method presented herein would lead to results of similar or better quality as in this study, when applied to drone-based low-cost RGB camera data.

4.4.8 Combining multiple temporal features in a PLSR analysis

Differences in absolute values of spectral bands or VIs depend not only on the phenology of a plant but also on morphology and canopy structure (Anderegg, Yu, et al., 2020), leaf pigments and epicuticular waxes (Tafesse et al., 2022) and viewing geometry (Aasen, Kirchgessner, et al., 2020). However, relative changes over time, *i.e.* the dynamics of the VIs are stable and suitable for the extraction of temporal features (Aasen, Kirchgessner, et al., 2020; Anderegg, Yu, et al., 2020).

Pigments such as chlorophyll, anthocyanins, and carotenoids are formed and degraded at specific times during plant growth (A. Fischer and Feller, 1994; Hörtensteiner, 2006) and these changes are temporally correlated with dynamic changes in VIs. At different developmental stages, different VI-based temporal features of different VIs are best correlated with physiological processes of the plant. Thus, it is reasonable to base the analysis not on absolute VI values but their dynamics and not to use a single VI for all stages of phenology and senescence, but to combine multiple VIs in an analysis. With this rationale, using temporal features in a PLSR analysis is a promising new approach. PLSR can be used to handle datasets, where the number of predictor variables is higher than the number of observations, and where the predictor variables are strongly correlated (Carrascal et al., 2009), which can be expected for the different temporal features used as predictor variables in this study.

Thus, selecting the 200 most predictive features was not meant to avoid overfitting but to reduce computational effort in the 100 times repeated cross-validation. In PLSR analysis, overfitting can be avoided by choosing a number of PLSR components that is significantly smaller than the number of predictor variables. With a maximum number of 15 components in our PLSR models, the ratio predictors/components was ≥ 139 for PhenoCam data (2092 or 2101 observations and 15 or fewer components), and ≥ 54 for drone data (810 observations and 15 or fewer components). Thus, the number of observations was much larger than the number of components, and no overfitting would be expected.

While the 75 %/25 % train/test data split in cross-validation led to slightly better prediction accuracy, also 19 genotypes/11 genotypes train/test cross-validation allowed for high correlations between predictions and visual reference measurements. The development of 11 randomly chosen genotypes, and thus also 99 plots, unseen in training, were accurately predicted by PLSR models in 100 repetitions for each reference level, which demonstrates the generalizability of the method and argues against overfitting. An increased set of genotypes in training could further increase the generalizability of the PLSR prediction.

As the temporal features used as predictors were expected to be highly correlated, PLSR was preferred over other approaches such as random forest regression. Random forests are prone to overfitting when using highly correlated data (*e.g.* Gregorutti et al., 2016) and PLSR was shown to produce more generalizable predictions than random forests (*e.g.* Lee et al., 2017). Recursive feature elimination (RFE) can be applied to increase generalizability and the risk of overfitting (Gregorutti et al., 2016; Darst et al., 2018), however, PLSR does not need a lot of computational capacity, whereas repeated RFE can be very computationally intensive.

4.4.9 Cost and measurement frequency of different methods

PhenoCams were the cheapest method for tracking phenology and senescence after five seasons when considering an experimental site with a driving distance of 3 h (one-way). While the initial costs of PhenoCams for hardware and the efforts for setup were fairly high, they were allowing for an almost unlimited increase of the temporal resolution of image acquisition to hourly or even beyond without significantly increasing data acquisition costs. In contrast, for visual field ratings and drone-base approaches, every additional measurement came at considerable marginal costs.

The main driver of cost for additional measurements of the next cheapest approaches (visual ratings and RGB drones) were the round trips, necessary for each measurement.

In contrast, PhenoCams need - assuming no technical incidents occur - just two field visits for setup and dismantling and if images could be transmitted via mobile networks, SD cards would not need to be changed when full. DNG raw format used in this study was too data-heavy to be transmitted to a server via a mobile network. However, it was shown that the JPEG format-based VIs allowed the tracking of senescence and phenology even slightly better than those based on DNG raw format. 8-bit JPEG format is lighter and can be transmitted to servers; thus, no SD card change would be necessary in such a setup, and also no storage limitation would hinder frequent image acquisition. An image in JPEG could be transmitted every 10 min, maximizing the probability of capturing good quality images on many days. Such a JPEG based setup also would make it possible to follow the seamless operation of the cameras almost in real time, without the need to visit the PhenoCams in person. Using PhenoCams with JPEG format therefore offers many benefits without a major loss in quality.

Tschurr et al. (2024) argued, that at higher temporal resolution, RGB VIs can make up for lower spectral resolution. Such comparisons are difficult for multiple reasons. *E.g.* Tschurr et al. (2024) did not include the DVI, PSRI and SAVI multispectral VIs in their study, which all showed high predictiveness in this study and PSRI was approximating visual senescence ratings best also in Anderegg, Yu, et al. (2020). In addition, the study at hand confirmed the number of field visits as an important cost driver in the context of remote field experiments (Barreto et al., 2024; Montazeaud et al., 2016; Velumani et al., 2020) . The marginal costs for the RBG drone at 0.5 h, 1.5 h and 3 h driving distance were 407, 659 and 1037 CHF respectively in the 700 plots scenario. For the 1400 plot scenario, the cost was again significantly higher (Table 4.7). Thus, additional measurements came at a price, and if multispectral sensors need to be flown less often, this could lead to multispectral VIs being the economically more favorable option despite higher marginal costs, depending on the difference in the number of flights required compared to RGB VIs. Multispectral becomes particularly interesting if the sensor has already been procured to measure other plant traits.

In addition, drone flights must be organized along with other activities, and the logistics of a field season can be very demanding, as many tasks can only be completed in good weather conditions. Due to time constraints, it is often not possible to fly in optimal conditions or to fly at all, especially in rainy periods and for distant experiments. Multispectral drones therefore also have an advantage in these aspects, due to the lower number of flights required in dense field seasons, thereby allowing for more flexibility in planing and a lower workload. PhenoCams, on the contrary, might capture an image during that rare 20 min of a day under suitable conditions without the need of intervention. Further improvements in the PhenoCam setup and pipeline could lead to an additional shift in balance in favor of PhenoCams.

However, with the same argumentation, autonomous drone systems might shift the balance in favor of drones again. Although these systems were strongly restricted by regulation some years ago in many countries (Aasen, Kirchgessner, et al., 2020), the legislation changed in some countries and such systems can be operated, drastically facilitating the logistics of distant experiments and increasing the probability of a high frequency of flights in fair meteorological conditions.

The cost comparison did not include additional benefits of the methods. Especially multispectral VIs come with additional information on the plant state such as general health, nitrogen content, etc. It is challenging to put a price tag on this type of information, but the aspect should not be neglected in such considerations.

Finally, the cost considerations presented herein are meant to serve as a conceptual framework that is allowing to approximate real costs and to reason about most relevant cost

structures. They are not meant to be precise representation of true costs, which are even more complex.

4.4.10 VI and feature type importance in PLSR modeling without ex-ante knowledge on phenology

Normalized relative PLSR coefficient sums $norm.\beta_{rel,sum}$ were used to describe the importance of different VIs and feature types in respective PLSR models. They are an integrative measure as they also are impacted by the number of temporal features in the input data of the PLSR models. This depended a lot on how well the VIs could be smoothed with the different smoothing functions or fit with a Gompertz function. Many studies usually normalize dynamics with ex ante knowledge about phenology, *e.g.* by calculating the days after anthesis (*e.g.* Anderegg, Yu, et al., 2020; J. T. Christopher, Veyradier, et al., 2014; J. T. Christopher, M. J. Christopher, et al., 2016; Cao et al., 2021). This requires preceding tracking of phenology. The methods developed and examined in this study were meant to work without the need for supplement ratings. The method relied on a clear VI dynamic from an early minimum to a late maximum or vice versa. VIs that did not follow such a clear dynamic may have failed in the automated analysis procedure, may thus be underrepresented in the PLSR input data, and may never reach high $norm.\beta_{rel,sum}$ values. Yet, $norm.\beta_{rel,sum}$ still is a valid metric to describe the overall usefulness of a VI to the process presented.

As seen before, different sensors, image formats, and data treatments can lead to different VI dynamics. That is why between methods, different VIs and feature types were most predictive. In addition, the input data was highly correlated and with correlated data, small changes in the data can lead to the preference of one feature over the other.

Within one data type or image format, the temporal features of the input data in the PLSR models were the same for the PLSR model of the different reference rating types (BBCH, SenLeaf, Sen Plant) and classes (Early, Intermediate, Late). Changes in $norm.\beta_{rel,sum}$ for VIs and feature types between reference rating types and classes showed that for different phenotypic processes and stages, different temporal features were most useful, indicating that the tracking of such processes should not rely on one but multiple VIs.

4.4.11 Mask creation

Creating the plot masks was a tedious and time consuming task, as the cameras were shaking and manual adjustments for multiple plot mask files were necessary throughout the season. This was made even more difficult by the lens distortion of the cameras, which meant that the plot mask had to be not only rotated and shifted in QGIS, but sometimes also adjusted in size and shape.

Using SIFT to automatically adjust plot mask failed due to image-to-image error propagation. This was most likely caused by growing vegetation, which changed the appearance of the scenery and maybe most importantly the shadow patterns between the plots. Shadows are often strongly contrasted visual features and have a high importance in SIFT.

4.4.12 Recommendation for PheonCam setup and pipeline improvement

Based on the experience gained in this work, the following recommendations are given on how the PhenoCam setup and pipeline could be improved in the future.

- Establish an automated digital image stabilization to compensate for shaking cameras.
- Enable co-registration of images with mask and adjustment to a growing vegetation to decrease initial processing efforts.

- Implement reliable image-wise quality estimates in pre-processing, *e.g.* brightness, blurriness, visibility, direct sun, etc. to sort out inferior quality images.
- Adapt viewing geometry to field specific conditions, to avoid sky and direct sun in the images for more homogeneous illumination.
- Adapt viewing geometry of cameras to reduce "lost" image sections that do not cover the field but neighboring terrain.
- Maintain narrow optical opening angles towards the horizon, as wide opening angles make plot recognition at the far end of the experiment difficult.
- Use masts that are less prone to shaking in windy conditions.
- Position the masts on the wide side of the field, ideally at two opposite locations, to decreases the maximal distance between PhenoCams and plots.
- Use masts with a smaller footprint, that could even be installed inside the field without impeding field operation and cover the experiment at 360° and not just from one side.
- Use JPEG format, which can be transmitted with mobile networks and sent to a server.
- Adapt the rationale for temporal feature extraction specifically for different VIs and data types.

4.5 Conclusion

A mobile PhenoCam was installed in a wheat variety testing trial for three consecutive seasons. The aim was to track phenology from heading onward and senescence at plant and flag leaf levels. With a PLSR approach, multiple temporal features of different Vis were analyzed in one model. A high prediction accuracy for all phases was attained for all developmental stages, and the prediction accuracy of a drone-based multispectral sensor only slightly outperformed PhenoCams. Uncalibrated JPEG images were sufficient to track plant development, and in future setups, images could directly be transmitted via mobile networks, which allows for an almost real-time tracking of plant development even at remote experimental sites. A cost analysis showed that transfers between experimental sites became an important cost driver for visual ratings and drone-based methods. PhenoCams came at a higher initial material investment, but tracked plant development at a lower marginal cost and became the cheaper option over time, with an increasing number of measurements. Thus, the proposed PhenoCam setup, combined with a PLSR analysis based on temporal features, is a cost-effective lean-phenotyping method to replace visual ratings of phenology and senescence in multi-environment trials.

Authors' contribution

Simon Treier: conceptualization, methodology, software, formal analysis, visualization, writing – original draft. Juan M. Herrera: project administration, funding acquisition, conceptualization, supervision, methodology, acquisition, writing – review & editing. Lukas Roth: writing – review & editing. Nicolas Vuille-dit-Bille: Support in data acquisition and index curation, writing – review & editing. Margot Visse-Mansiaux: Support in data acquisition, review & editing. Helge Aasen: Practical advice, review & editing. Frank Liebisch: Data acquisition, review & editing. Achim Walter, Helge Aasen: review & editing.

Acknowledgments

We thank Johanna Antretter, Fernanda Arelmann Steinbrecher, Ulysse Schaller, Matthias Schmid and Julien Vaudroz for rating of phenology; Nicolas Widmer and his team as well as Yann Imhoff for field management.

Conflict of interest statement

The authors declare no conflict of interest.

Data availability

A detailed description of the mask creation and adjustment procedure suggested in this publication, including source code and example data, is provided on GitHub (`https://github.com/TreAgron/ShapeFromCSVHomographyTransform/tree/main`).

Funding

This study was in part supported by the two H2020 projects InnoVar and Invite.

5 Evaluating the potential of chlorophyll fluorescence to detect and rate *Fusarium* head blight on field experiments for winter wheat variety testing

Simon Treier[1,3], Romina Morisoli[2], Achim Walter[3], Fabio Mascher[4,5], Juan M. Herrera[1]

1 – Production Technology & Cropping Systems Group, Agroscope, Route de Duiller 60, 1260 Nyon, Switzerland

2 – Neobiota group, Agroscope, A Ramél 18, 6593 Cadenazzo, Switzerland

3 – ETH Zürich, Institute of Agricultural Sciences, Universitätstrasse 2, 8092 Zürich, Switzerland

4 – University of Applied Sciences, School of Agricultural, Food and Forest Sciences, Länggasse 85, 3052 Zollikofen, Switzerland

5 – Field-Crop Breeding and Genetic Resources Group, Agroscope, Route de Duillier 60, 1260 Nyon, Switzerland

Abstract

Fusarium head blight (FHB) is a fungal disease caused by diverse *Fusarium* species and affecting various cereals. It reduces yield and quality of cereal crops, leading to significant economic losses. A major issue with *Fusarium* spp. are the various mycotoxins produced by the fungi, such as the vomitoxin deoxynivalenol (DON), which is the toxin related to the highest economic losses in cereal production. Cropping resistant genotypes is one of the most promising means of controlling the disease and to reduce DON. It's often used in combination with agronomic and chemical control strategies. As testing for DON is expensive and destructive, visual ratings are usually conducted to identify resistant varieties in field trials, which is tedious, time-consuming and subjective. Chlorophyll fluorescence (CF), detected with handheld point-measurement devices or CF cameras was proposed to detect *Fusarium* in the field, yet, the applicability of such methods in day-to-day practice under field conditions remained limited. In this study, a hand-held CF device, the FluorPen, was used to track *Fusarium* infestations first in a greenhouse on a trial comprising four wheat varieties. The method was then transferred to a field trial with 16 wheat varieties and tested for two seasons, together with a CF imaging approach, using a FluorCam. While FluorPen and FluorCam allowed to detect high infestation levels and to tell resistant from susceptible varieties, the FluorPen failed in low-level infestations and the FluorCam could not be tested on low-level

infestation. In addition, FluorPen and FluorCam depended on dark adaptation to create optimal conditions for CF measurements. Dark adaptation is very time consuming, as the samples need to be shaded or - as in this study - detached from the plant and transferred to a dark place for measurements. Thus a rapid FluorPen protocol with full field applicability without dark adaptation was tested. The higher efficiency of the rapid FluorPen protocol came at the cost of more variable measurements. Rapid CF distinguished well between healthy and infested tissue in high-level infestations, but it is hypothesized, that all methods tested in the field would fail at low-level infestations due to a too low number of measurements.

5.1 Introduction

Fusarium head blight (FHB) is a fungal disease caused by diverse *Fusarium* species and affects various cereals. It reduces yield and quality of cereal crops, leading sometimes to significant economic losses for farmers (McMullen et al., 1997; Windels, 2000). A major issue with *Fusarium* spp. are the various mycotoxins produced by the fungi, which have adverse effects on health of humans and livestock (Ferrigo et al., 2016). The vomitoxin deoxynivalenol (DON) is the toxin related to the highest economic losses in cereal production (Munkvold, 2017). Harvest products contaminated with DON are subject to steep price discounts, as they are subject to regulatory limits (Wilson et al. 2018). Agronomic and chemical control of *Fusarium* spp. is only partially effective. Cropping resistant genotypes is one of the most promising means of controlling the disease (Peiris, Bockus, et al., 2016). For example, in Germany, breeding for disease resistant genotypes was an important driver of yield gain in recent years (Zetzsche et al., 2020). Resistant genotypes are often combined with chemical and agronomic control strategies (Wilson et al., 2018). While the aim of avoiding *Fusarium* spp. infestation in the fields is mainly to avoid contamination of the crop with toxins, direct detection of DON is not possible and resistance selection is usually based on visually observable symptoms. Detecting such symptoms is tedious and time consuming (Almoujahed et al., 2022; Bannihatti et al., 2022) and could be subjective due to rater bias (Elke Bauriegel and Herppich, 2014; Mahlein et al., 2019; Hong et al., 2022). In addition, correlation of *Fusarium* spp. symptoms and DON content is often rather weak (Schlang et al., 2008; Ajigboye et al., 2016), as several types of resistance contribute to the reduction of DON accumulation (Martin et al. 2017). At the same time, the detection of DON by the standard laboratory method of liquid chromatography coupled with mass-spectrometry is troublesome, destructive, and expensive (Peiris, Bockus, et al., 2016) as is the detection by enzyme-linked immunosorbent assay (ELISA) (Levasseur-Garcia, 2018; Wilson et al., 2018). Therefore, fast and cost-effective methods to detect genotypes that do not accumulate high levels of DON would facilitate the assessment of plant resistance to DON accumulation in breeding and variety testing programs. Attempts were made to detect FHB and related toxins in cereals by exploiting information of hyperspectral reflectance (E. Bauriegel et al., 2010; E. Alisaac et al., 2018; Almoujahed et al., 2022; Vincke et al., 2023), computer vision on RGB images (Qiu et al., 2019; Su et al., 2021; Hong et al., 2022), thermal imaging (Al Masri et al., 2017) or chlorophyll fluorescence (CF) (Ajigboye et al., 2016; Sunic et al., 2023) and different combinations of these methods (*e.g.* Mahlein et al., 2019; L. Huang et al., 2020; Mustafa et al., 2023). On field scale, drones were proposed to detect infested field areas (Francesconi et al., 2021; H. Zhang et al., 2022) *e.g.* for exclusion from harvest (Elke Bauriegel and Herppich, 2014) or early detection for a more successful chemical treatment (Francesconi et al., 2021; Elias Alisaac and Mahlein, 2023).

In kernels and flour, DON detection was proposed with near-infrared and mid-infrared spectroscopy (Peiris, Pumphrey, et al., 2010; Almoujahed et al., 2024) and hyperspectral imaging (Elias Alisaac, Behmann, et al., 2019).

FHB infections cause tissue colonization with hyphae and, consequently, the plugging of plant vessels and the degradation of plant tissue. This changes the physiology and photosynthesis of plant organs and affects the reflectance, transpiration, and fluorescence of the tissue. Such alterations are ultimately detectable by the aforementioned methods (Elke Bauriegel and Herppich, 2014; Al Masri et al., 2017). These methods are therefore indirect measurements as they do not directly detect the presence of fungal tissue or toxins but rather the effect of the pathogens on the plant. An exception is reflectance, as it might change due to a change in plant physiology, as well as due to the presence of fungal material (*e.g.* pinkish mycelium; Qiu et al., 2019). Nevertheless, these sensor-based methods might be more sensitive or more objective and thus informative than visual ratings.

CF is a technique for indirectly estimating the efficiency and status of the photosynthetic apparatus. It has been applied for a wide range of examinations of plants and their response to stressors and environments (Pask et al., 2012). Ajigboye et al. (2016) used a portable handheld fluorometer to detect various *Fusarium* species and to describe the progress of disease in a greenhouse experiment on a single wheat variety and Sunic et al. (2023) used a similar approach on six wheat genotypes, but also in a potted experiment. Elke Bauriegel, Giebel, et al. (2011) used CF imaging to measure the impact of *Fusarium culmorum* on the photosynthetic status of wheat ears. They analyzed CF image time series of a pot experiment on a single variety and conducted in-field measurements on three varieties. While a handheld portable fluorometer is easy to apply and relatively inexpensive, using CF imaging provides information on spatial patterns of CF values, and these patterns might be more informative than mere CF values of isolated spots.

Although these studies have shown the potential of CF to distinguish infected tissue from healthy tissue, its application in variety testing or breeding poses some specific challenges. A higher number of genotypes needs to be screened in field environments, which increases the workload and thereby logistical challenges, especially in the context of multilocation trials. Also adverse effects of outdoor conditions, like restrictions of accessibility due to weather or field heterogeneity must be taken into account. An important drawback of CF is that samples must be dark adapted for measurements. This can be done by sheltering samples before and during measurements or by measuring indoors, however, both methods come with a significant increase of measurement logistics efforts.

This study tested the suitability of CF to characterize *Fusarium* spp. infestations on wheat spikes at a plot level in field experiments. A hand-held FluorPen and a CF imaging device, the FluorCam, were tested. The different tools come with different benefits. A handheld CF is relatively cheap and easy to apply, and no special pre-processing of data is required to obtain results related to *Fusarium* spp. infestation. A CF imaging device provides spatial information on infections, but the measurement protocol is more complex to implement and pre-processing of images is necessary to obtain results related to *Fusarium* spp. infestation. Both methods require a dark adaptation of the samples for optimal results.

5.2 Methods

The study consists of four main parts. A greenhouse experiment was conducted to examine the potential of a FluorPen FP110 (Photon Systems Instruments, Drásov, Czech Republic) to track disease progression on four varieties. The FluorPen FP110 is a portable battery-powered Pulse-Amplitude-Modulation (PAM) fluorometer that provides multiple fluorescence measurement protocols and the OJIP protocol with dark adaptation was used for the greenhouse experiment (Fig. 5.1a). The approach was then transferred to a real wheat variety field experiment in a destructive approach, in which for each measurement event, the wheat tillers were cut off, placed in water and transferred to a dark environment for dark adaptation before running

the OJIP protocol with the FluorPen (Fig. 5.1b). The same spikes were also analyzed with the FluorCam CF imaging device. (Fig. 5.1c). Finally, a simple F_v/F_m protocol was applied to spikes directly in the field without dark adaptation to estimate the potential of CF under realistic field conditions (Fig. 5.1d).

5.2.1 Greenhouse experiment

Seeds of the four Swiss wheat varieties CADLIMO, CH-NARA, MONTALBANO and PIZNAIR were seeded in small pots on 2020-11-19 and transferred to a growth chamber on 2022-11-23 for initial plant establishment. The potting soil, contained 25 percent mineral soil, 60 percent white peat and 15 percent perlite (custom mixture, Ricoter Erdaufbereitung AG, Aarberg, Switzerland). On 2020-11-26, the plots were transferred to a cold chamber for vernalization. Twelve plants of each variety were replanted in larger pots with a volume of 3 L and transferred to the greenhouse on 2021-01-29. Inoculum of *Fusarium culmorum* was collected in the fields of Changins in previous years and most virulent strains were propagated on sterilized oat husks. Inoculation was conducted on six plants of each variety on the main tiller on 2021-04-09, when most of the plants were flowering with visible yellow anthers. For inoculation, a conidia solution with demineralized water was adjusted to 10^6conidia mL^{-1} using a hemocytometer (Ajigboye et al., 2016). Cotton balls about the size of a wheat grain were formed and dipped into the solution and then gently placed inside a lateral floret (F3 or F4 according to Wilhelm and G. S. McMaster, 1996) of a spikelet in the center of the spike (S5, S6 or S7 according to Wilhelm and G. S. McMaster, 1996). As a control, cotton balls just dipped in demineralized water were used for healthy plants. In total, the experiment comprised six healthy and six inoculated plants of four varieties, or 48 plants in total.

The plants were then placed inside a greenhouse and exposed to 100 % humidity for 72 h, which was maintained with a water nebulizer. Visual ratings of infestation severity on different spikelets along the spike were then performed on 15 occasions, from 3 days after inoculation (DAI) to 46 DAI, when all spikes were completely bleached and the disease symptoms were no longer distinguishable from the senescent spikes, similar to the work of Peiris, Bockus, et al., 2016, which examined the DON content of kernels in different spikelets with near-infrared spectroscopy. Five spikelets were labeled with red paint (Fig. 5.1a). The “Center” spike represented the site of inoculation. “Mid-Tip” and “Tip” represent the spikelets at two and four positions above the inoculation site, “Mid-Base” and “Base” the spikelets at two and four positions below. Rating was according to the rating scheme of Mahlein et al. (2019), where a perfectly healthy spikelet corresponds to a value of 0 % and a completely infected and bleached spikelet to a value of 100 %. As bleaching related to senescence is very similar to bleaching related to early states of *Fusarium* infestations, senescent spikes were rated according to the same scale. On nine plants, no visual signs of infection could be found at eleven DAI. A moistened plastic bag was placed over these spikes for 72 h to reinforce the development of the disease. After this treatment, all sites of inoculation developed visible *Fusarium* symptoms.

In parallel to visual ratings, CF measurements were performed with a handheld portable fluorometer (FluorPen FP110, Photon Systems Instruments, Drásov, Czech Republic) with the fast induction fluorescence rise protocol, named OJIP (PSI Photon Systems Instruments, 2021), which was proposed to track *Fusarium* development on wheat by Ajigboye et al., 2016. For a thorough description of the OJIP protocol and its parameters, see Strasser et al. (2000). The pots were transferred to a completely dark place prior to measurements and were adapted to dark for at least 20 min. Measurements were then conducted with a weak headlight and by avoiding directly pointing the light beam at the location of measurements. A leaf clip was positioned on the flat side of each spikelet and the FluorPen smoothly pressed onto the leaf clip for the OJIP measurements, which took approximately 10 s for each spikelet. The saturating

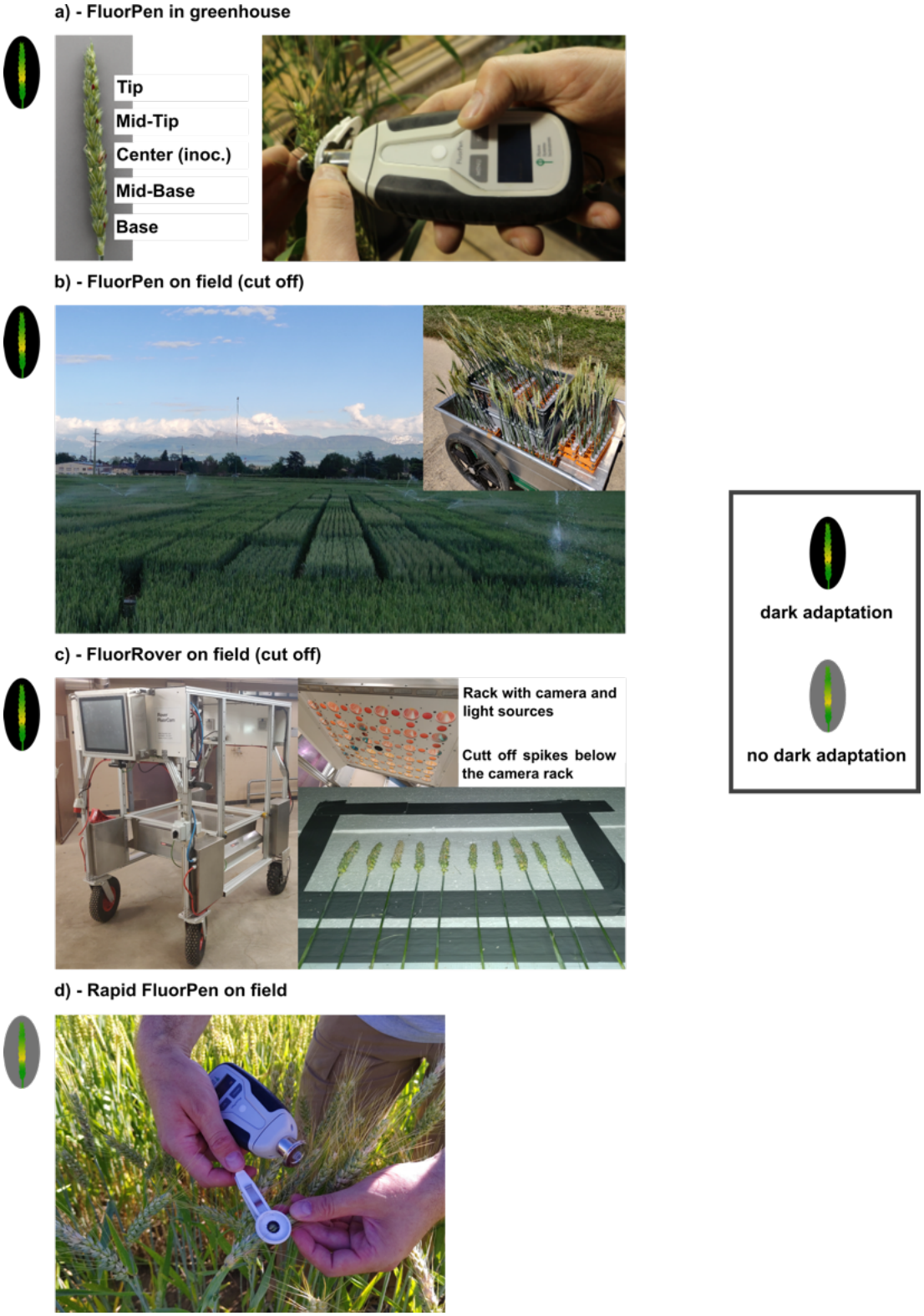


FIGURE 5.1: Overview on the experimental parts of the study: The OJIP protocol of the FluorPen was tested on five spike positions on potted wheat plants of four genotypes in a greenhouse experiment and with dark adaptation prior to measurements (a). The approach was then applied to a field experiment in a destructive approach, where for each measurement event, wheat tillers were cut off, put into water and transferred to a dark environment for dark adaptation before running the OJIP protocol with the FluorPen (b). The same spikes were also analyzed with the FluorCam CF imaging device (c). Finally, a simple F_v'/F_m' protocol was applied to spikes directly in the field without dark adaptation. The icon on the top-left indicates, whether dark adaptation was applied (d).

illumination was emitted with 2850 µmol m^{-2} s^{-1} at 470 nm, the actinic illumination pulse was at 300 µmol m^{-2} s^{-1} and fluorescence was detected between 667 nm and 750 nm. The suitability of the different OJIP parameters to distinguish infested from healthy tissues was tested by analysis of variance (ANOVA), with the model

$$Parameter_{OJIP} \sim Inoculation\ treatment + Spikelet\ position + Variety, \quad (5.1)$$

where effects were considered significantly different at $p < 0.01$. Visual ratings and OJIP parameters were compared by visual inspection of the data and by correlation. Statistical analysis was conducted in R (R Development Core Team, 2022).

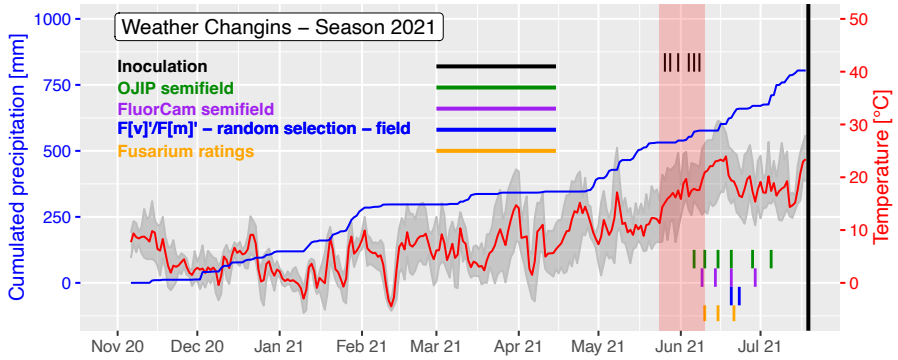
The FluorPen assigned a continuous number to each measurement. The Field Book app (Rife and Poland, 2014) app was used to record this number for the corresponding samples for a correct attribution of the data in later analysis.

At physiological maturity, the grains were harvested for individual spikelets and weighed. The rate of *Fusarium* damaged kernels (FDK) was calculated as the proportion of kernels with *Fusarium* symptoms out of the total number of kernels per spikelet. DON content was estimated by ELISA for individual spikelets of inoculated plants. As very small amounts of grains were harvested from individual spikelets, the grains were crushed in a mortar and then placed in 1.5 mL Eppendorf tubes together with a steel grinding ball. The Eppendorf tubes were then placed in a custom-made container and ground to fine flour in a ball grinder. After each sample, the material used was thoroughly cleaned to avoid cross-contamination. By weighing the Eppendorf tubes when they were empty and after the ball was removed, the flour available for ELISA was determined. The flour was then diluted to meet the concentration needed for ELISA analysis. With just very small amounts of dilution, filtration of the dilution would have caused too much loss. The solution was thus separated from the solid parts of the flour in a centrifuge by rotating for 3 min at 5'000 rotations min^{-1}. The concentration-adjusted solution was then pipetted on the ELISA kit (RIDASCREEN®FAST DON, r-biopharm AG, Pfungstadt, Germany), with a detection range between 0.2 and 6 mg DON kg^{-1} flour. From previous experience, it was known that the DON load for strongly infested samples was often above this range and heavily infested samples were diluted again by a factor of 10 to stay within the detection range of the ELISA kit.

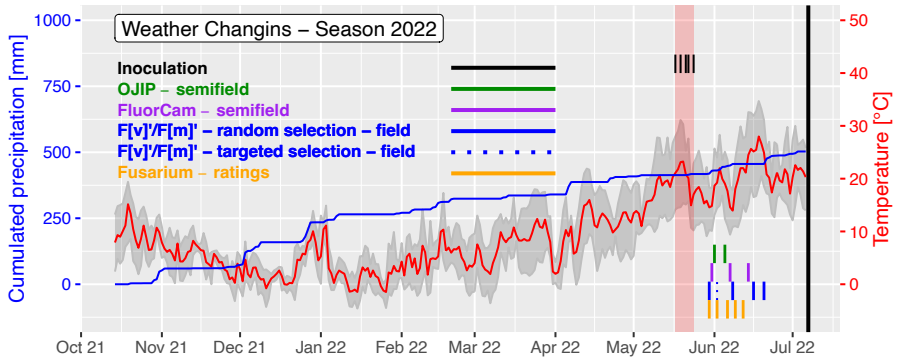
To relate the DON loads of the samples to the OJIP parameters, the concept of area under the disease-progress curve (AUDPC; Jeger and Viljanen-Rollinson, 2001) was used. The area under the curve (AUC) was normalized for each variety by dividing the AUC values by the mean of the AUC of the healthy samples of each variety. The normalized AUC_{norm} of the inoculated samples was then correlated with the DON loads.

5.2.2 Field experiments

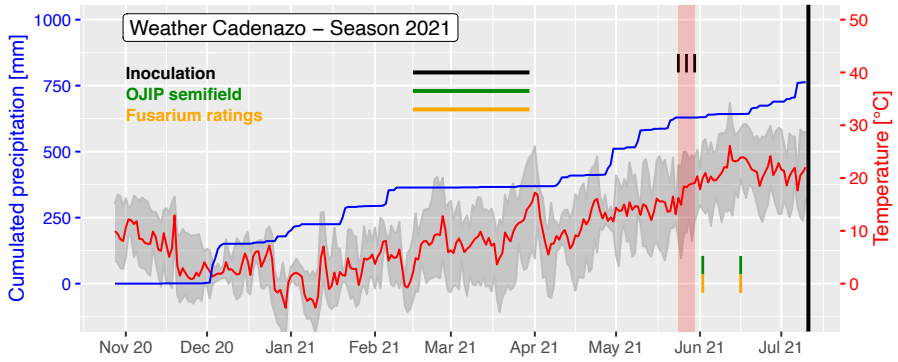
Field experiments (Fig. 5.1b) were sown at Agroscope agricultural research stations in two locations within Switzerland. At Changins [46°23′55.4″N 6°14′20.4″E, 425 m.a.s.l., the World Geodetic System (WGS) 84], experiments were sown in the seasons 2020-2021 and 2021-2022. The soil of the experimental site is a shallow Calcaric Cambisol (Baxter, 2007; Cárcer et al., 2019). At Cadenazzo [46°9′36.82″N 8°56′5.05″E, 203 m.a.s.l., the World Geodetic System (WGS) 84] there was an experiment only in the season 2020-2021 on a Eutric Fluvisol (Baxter, 2007; Gallet et al., 2003). The three experiments are further referred to as CHA21, CHA22 and CAD21. The trials comprised 16 wheat varieties currently or recently inscribed in the Swiss national variety catalog with different levels of tolerance towards *Fusarium* head blight (Table S4.1). 13 varieties were winter wheat (ARINA, AXEN, BARETTA, BODELI, CADLIMO, CH-NARA, DIAVEL, MONTALBANO, PIZNAIR, POSMEDA, ROSATCH, SCHILTHORN, VARAPPE) and three spring wheat (ARPILLE, FIORINA, QUARNA). Each



(a)



(b)



(c)

FIGURE 5.2: Overview on measurements and meteorological conditions in the two experimental sites in 2021 (a & c) and in 2022 at Changins only (b) from sowing until after harvest. Red shows the mean air temperature, and the shades indicate daily temperature minima and maxima. Cumulative precipitation is shown as a rising blue line. The vertical green lines indicate the dates of FluoPen measurements according to the OJIP protocol with dark adaptation and purple indicates FluorCam measurements. Blue lines indicated rapid F_v'/F_m' parameter measurements without dark adaptation where the dottet line in 2022 indicates an approach targeted on symptomatic areas of the spike. Orange lines mark dates when visual ratings of infestation were conducted. Short black lines indicate inoculation dates. During the period shaded in red, flouring (BBCH 61 - 69, Lancashire et al., 1991) was observed in the field. Harvest dates are marked by black lines at the end of the seasons.

variety was sown in plots in two treatments; "FUS" (with inoculation) and "CON" (control). Within single plots, a wheat variety was sown in eight rows, with a spacing of 15 cm between them. Between the plots, separation gaps were kept free of wheat. The gaps were 0.3 m between the long side of the plots and 1 m direction of sowing. This resulted in plots of about 1.25 m x 4.6 m each. Each variety-treatment combination was replicated on four plots. The replications were arranged in blocks of 4 × 4 plots and blocks separated with additional border plots. The four blocks of each treatment were arranged in a 4 × 2 pattern and the total of 128 plots of the experiment span 16 rows (which followed the tractor track direction) and 8 columns. The single plots were arranged in enhanced random designs, which were generated with the R package DiGGer (Coombes, 2009; `http://nswdpibiom.org/austatgen/software`). The design was identical for CHA21 and CAD21 (Fig. S4.1) but differed for CHA22 (Fig. S4.2). In total, the experiments were about 54 m long (in tractor track direction) and about 30 m wide. Fertilizers and herbicides were applied in three splits and at equal rates to all treatments according to the Proof of Ecological Performance (PEP) certification guidelines Swiss Federal Council, 2013, which represent a minimal standard for best practice for conventional agriculture in Switzerland. Tables S4.2 and S4.3 provide more detail of the different treatments.

The air temperature and daily precipitation were obtained from Meteoswiss weather stations (Federal Office of Meteorology and Climatology, `https://www.meteoswiss.admin.ch`), which were located about 800 m from the experimental site at Changins [46°24′3.7″N 6°13′39.6″E, 458 m.a.s.l., WGS 84], and in direct vicinity of the experiment at Cadenazzo.

The meteorological conditions for the different years and sites are presented in Fig. 5.2 for the period from sowing to harvest together with the timing of inoculation events and measurement campaigns.

5.2.3 Inoculation of field trials

To carry out artificial inoculations in the field, the same *Fusarium culmorum* inoculum was used as for the greenhouse trial. For CHA21, conidia (Karlsson et al., 2021; Pellan et al., 2021) were dissolved in demineralized water and adjusted to 10^{-6} conidia mL^{-1} with a hemocytometer (Ajigboye et al., 2016). The plots were individually inoculated with a backpack sprayer and a flat fan nozzle with 0.35 L of conidia solution per plot. The first inoculations were made when the first signs of flowering (yellow anthers) were observed for the first time and were repeated every two to four days. To consider differences in flowering time, the first inoculation occurred at different dates for different plots and, in total, inoculation was conducted at six different dates (2021-05-27, 2021-05-29, 2021-06-01, 2021-06-05, 2021-06-07, 2021-06-09, *cf.* Fig. 5.2a), where the same plot was just inoculated three times. To ensure humid conditions for the establishment of the disease, an irrigation system was installed in the field that covered the entire experiment with eight 360° impact sprinklers (Fig. 5.1b). Sprinklers were turned on for more than 15 min before inoculation and for several 15 min intervals in the morning after inoculation events on days without rain with a total irrigation rate of approximately 200 L min^{-1} for the entire field (or $\sim$ 0.06 L m^{-2} min^{-1}).

2022 was a hot and dry season (Fig. 5.2b), and the inoculation regimen was modified. Inoculation was carried out for five days in all plots (2022-05-18, 2022-05-20, 2022-05-22, 2022-05-23, 2022-05-25, *cf.* Fig. 5.2a) with a concentration of 0.5×10^6conidia mL^{-1}, starting on the date the first flowering was observed in the field. To create a humid field environment, the eight sprinklers were on for up to two hours before inoculation. Inoculation was carried out in the evening close to or after sunset to reduce exposure of conidia to radiation and to profit from more humid conditions throughout the night. In the morning after inoculation, the sprinklers were again on for several 15 min intervals to maintain humidity until noon.

At Cadenazzo, a motorized 12 L backpack sprayer equipped with a 5-nozzle bar was used to disperse a solution of 10^6conidia mL^{-1} for inoculation on three dates (2021-05-25, 2021-05-28, 2021-05-31, *cf.* Fig. 5.2c). Approximately 5-6 hours after inoculation, a pass with the same sprayer with only water was conducted to maintain humidity on the spikes.

5.2.4 Reference rating on field trials

Fusarium infestation was rated using the scales proposed in Moll et al. (2000), where the the wheat spikes were rated to be at 0 % when no symptoms were visible and at 100 % when the whole spike was infested. From each plot, 15 spikes were rated, resulting in 60 spikes per variety and treatment, and 1'920 rated spikes in total for each measurement campaign. For CHA21, there were three rating campaigns (2021-06-12, 2021-06-16, 2021-06-22, *cf.* Fig. 5.2a), while there were two each for CHA22 (2022-06-02, 2022-06-06, *cf.* Fig. 5.2c) and CAD21 (2021-06-03, 2021-06-13, *cf.* Fig. 5.2c).

5.2.5 OJIP on field experiment in a destructive approach

Dark adaptation is time-consuming and often not practical. For this experiment, a semifield approach was applied. Wheat tillers were detached from the plants in the field, below the top node. The detached internodes with the spikes were placed in plastic test tubes that were filled with water (Fig. 5.1b) and transferred to a relatively cool and shaded room for dark adaptation. The OJIP protocol was then applied to the detached spikes after dark adaptation.

As CF is related to the state of the photosynthetic apparatus, it was tested whether detaching the spikes from the tillers leads to a systematic difference between detached and non-detached spikes. To that end, three plants of each variety and treatment from the greenhouse trial were taken and two secondary tillers (those not used for the principle experiment) were labeled. One secondary tiller was detached below the top node and placed in water. The spikes still on the plant as well as the spikes on the detached internodes were then measured nine times throughout the next 26 h with the OJIP protocol.

There were six OJIP campaigns for CHA21 (2021-06-07, 2021-06-11, 2021-06-16, 2021-06-21, 2021-06-29, 2021-07-06, *cf.* Fig. 5.2) and two for CAD21 (2021-06-03, 2021-06-13). In both sites of 2021, three spikes per plot were measured in randomly chosen positions at the top of the spikelet (around spikelet S10 according to Wilhelm and G. S. McMaster, 1996) and in the central part of the spike (around spikelet S7 according to Wilhelm and G. S. McMaster, 1996). For each field-measurement campaign, new spikes were harvested. In CHA22, there were only two OJIP measurement campaigns (2022-06-02, 2022-06-06, *cf.* Fig. 5.2), but ten spikes per plot were measured instead of three for CHA21 and measured in the central part of the spike. Again, the suitability of the different OJIP parameters to distinguish infested from healthy tissues was tested by analysis of variance (ANOVA), with the model

$$Parameter_{OJIP} \sim Inoculation\ treatment + Variety, \tag{5.2}$$

where effects were considered significantly different at $p < 0.01$. This ANOVA model was also applied for both CF methods described in the following.

In CHA21, the OJIP measurements covered the period from inoculation to early senescence and AUC_{norm} could be calculated. AUC_{norm} of the inoculated samples was then correlated with the DON loads.

5.2.6 FluorCam measurements

Although the CF imaging device, a Rover FluorCam FC1000-R (Photon Systems Instruments, Drásov, Czech Republic) would allow for entering the field, the sequential dark adaptation

of 128 plots for at least 20 min was considered too time consuming and not practical. Thus, FluorCam measurements were conducted in a dark room.

The detached spikes were placed on a Styrofoam background at a distance of 37 cm from the rover sensor panel (Fig. 5.1c). The FluorCam could not run the OJIP protocol, which requires extremely high frame rates of the camera systems. Thus, the widely used ratio of variable fluorescence F_v and maximal fluorescence F_m (PSI Photon Systems Instruments, 2021) was used instead of the more complex OJIP protocol. The F_v/F_m protocol is fast ($\sim$ 10 s) and describes the maximum quantum yield of primary PSII photochemistry in dark adapted samples. Please, see Ajigboye et al. (*e.g.* 2016) and Strasser et al. (2000) and PSI Photon Systems Instruments (2021) for more details on the protocol. Shutter time was set to 5 µs, and the sensitivity was 35. The super pulse was set to 100 %, which corresponded to 2983 µmol m^{-2} s^{-1} at a distance of 25 cm from the slight source. For an efficient operation, two people carried out the measurements. One person operated the FluorCam, while the other handled the samples. The CF data were saved as TAR files.

For analysis, the TAR files were decompiled with a custom Python 3.8 (van Rossum, Guido and Drake, Fred L., 2009) script. CF images were then processed using the Python package "OpenCV" (Bradski and Kaehler, 2000) and "NumPy" (Harris et al., 2020). CF images taken throughout the different phases of the F_v/F_m protocol were offset against each other according to the FluorCam protocol. Variable fluorescence F_v was calculated as the difference between the fluorescence in the absence of photosynthetic light F_0 and the maximum fluorescence after the saturation pulse F_m and used to form the ratio F_v/F_m.

Raw F_v/F_m images were then filtered with an automatic Otsu threshold and cleaned from noise by morphological erosion. Connected components were detected with "OpenCV" to create masks for single spikes. Holes within the masks were filled with morphological dilation in one and two directions. The masks were then applied to the initial F_v/F_m images for a spike-wise analysis. With advanced infestations of the spikes and towards senescence, the spikes were not always recognized as whole on F_v/F_m images. In such situations, connected components with the highest overlap perpendicular to the direction of the rachis were considered as one spike.

With disease progression, the area of the spikes with photosynthetic activity became smaller and the intensity of F_v/F_m decreased. To capture these two effects in one metric, the median F_v/F_m per spike was multiplied by the area per spike expressed in numbers of pixels. The spike-wise index $Area \times med(F_v/F_m)$ was then normalized by the mean index value per variety and year to allow for a representation of very different value ranges on one scale.

5.2.7 FluorPen in the field without dark adaptation

With the experience gained through applying the FluorPen and FluorCam in the destructive experiment, it became very evident that the time-consuming logistics of cutting tillers and transferring them to a dark environment represent a major obstacle to the everyday use of this technology. Thus, a workflow without dark adaptation was tested with the FluorPen device (Fig. 5.1d). Spikes were chosen at random and measured with the "Qy" protocol of the FluorPen in the central part of the spike to derive the F_v'/F_m' parameter, which in contrast to F_v/F_m describes the maximum quantum yield of primary PSII photochemistry in light adapted samples. The spikes were shaded during the measurement by the body of the person taking the measurement, as illustrated in Fig. 5.1d. In addition, measurements were taken on the side facing away from the sun. In CHA21, the method was tested on just two dates (2021-06-21, 2021-06-24 10, *cf.* Fig. 5.2a). In CHA22, the method was applied more extensively on four dates (2022-05-31, 2022-06-09, 2022-06-17, 2022-06-21, *cf.* Fig. 5.2b). Ten spikes per plot were measured at random positions in the middle to upper part of the spikes ($\sim$ S6-S11 according

to Wilhelm and G. S. McMaster, 1996), resulting in 40 spikes per variety and inoculation treatment.

As a random selection of the measurement spot might miss the infestation of an otherwise infested spike, a targeted measurement was performed on 2022-06-03 with the FluorPen pointed at the symptomatic areas of randomly selected symptomatic spikes.

In CHA22, the F_v'/F_m' measurements covered the period from inoculation to early senescence and AUC_{norm} could be calculated. AUC_{norm} of the inoculated samples was then correlated with the DON loads.

5.2.8 DON and FDK estimation on field trials

In field trials, DON was estimated on flour samples for individual plots. All inoculated replication plots were sampled. It was known from previous trials that, as long as no symptoms were detected by visual ratings, the DON levels of the control plots were below the detection limit of 0.2 mg DON kg^{-1} flour. Thus, only one plot per variety was tested for DON among the control plots as a check measure. 5 g of ground wheat flour was soaked in 100 mL of demineralized water and the containers were shaken thoroughly and regularly for at least 5 min. The dispersion was then filtered through Whatman No. 1 filter and 50 µL were transferred to RIDASCREEN®FAST DON ELISA kits (r-biopharm AG, Pfungstadt, Germany). Again, strongly infested samples were diluted once more by a factor of 10 to remain within the detection range of the ELISA kit.

FDK proportion was visually determined by examining 100 randomly selected kernels from a well-mixed subsample of the harvested grains.

5.2.9 Experimental design and measurement sequence

When quantitative measurements span a long period, *i.e.* several hours, temporal trends might impact the measured values. It is therefore important not only to have an adequate replicated experimental design, but also to plan measurements accordingly. The sequence of measurements, from visual ratings to FluorPen and FluorCam destructive experiments to FluorPen field measurements, always followed a snake-pattern sequence through the plots within the field, ensuring that no more than eight plots of one inoculation treatment were measured in a row. This constant switch between inoculated and healthy spikes ensured to avoid artifacts by the confusion of a potential temporal trend with treatment effects.

5.3 Results

5.3.1 Greenhouse trial

Artificial inoculation of the four varieties used in the greenhouse trial caused *Fusarium* symptoms to start at the central spikelet (Fig. 5.1a) and spread, first to the top of the spike and later also toward the base, before senescence led to a complete bleaching of the spike (Fig. 5.3). The symptom succession of more susceptible genotypes (*e.g.* CH-NARA), was more rapid compared to more resistant genotypes (*e.g.* PIZNAIR).

When measuring these spikes with the FluorPen and the OJIP protocol, all OJIP parameters allowed the detection of significant effects ($p < 0.01$) between inoculation treatments ("Treatm."), varieties ("Variety."), and spikelet positions ("Spikel.") according to an ANOVA analysis (Eq. 5.1), for most DAI (Fig. 5.4). For the interaction "Treatm. × Spikel.", differences were significant for most OJIP parameters but only from 3 DAI to 32 DAI. "Treatm. ∼ Variety" was significantly different for most of the OJIP parameters from 35 DAI to 43 DAI. 'Spikel. ×

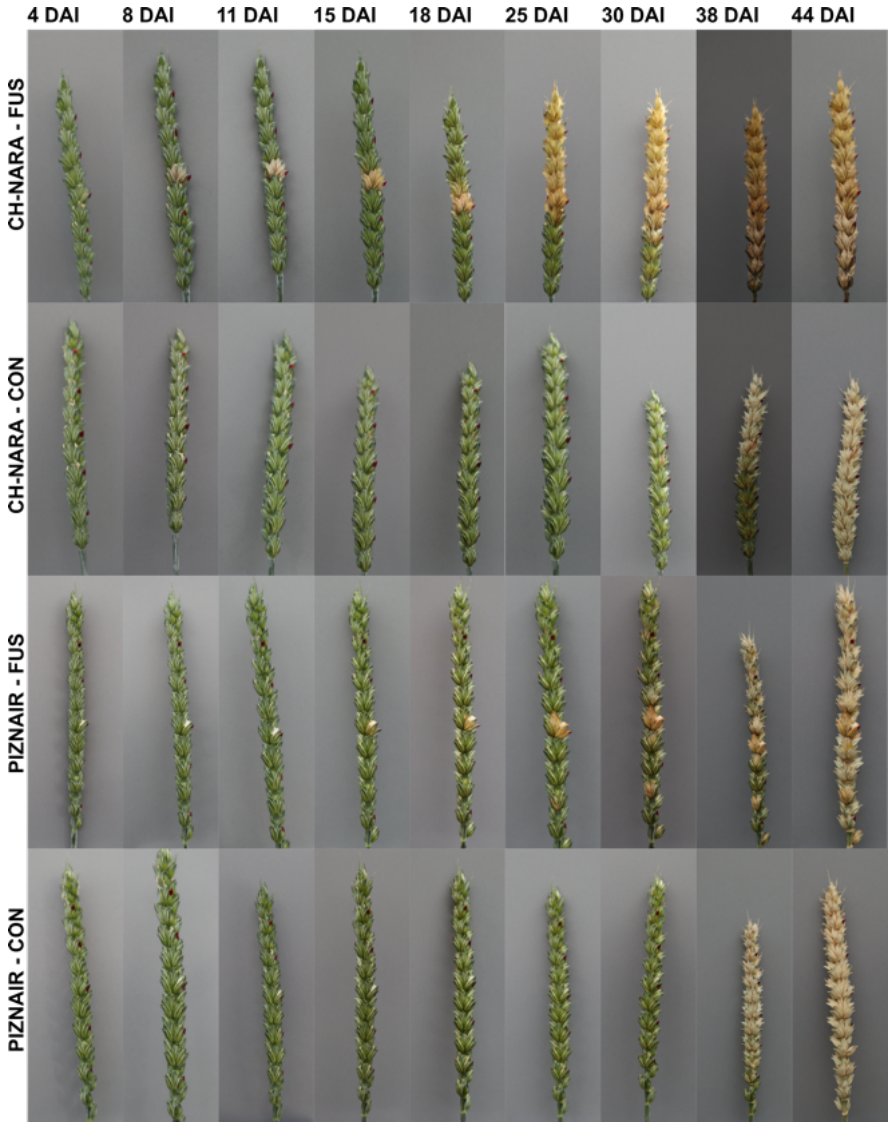


FIGURE 5.3: Succession fo *Fusarium culmorum* infestation in the greenhouse trial for two varieties (CH-NARA & PIZNAIR) and two inoculation treatments (FUS: Fusarium; CON: Control.) from 4 to 44 days after inoculation (DAI).

Variety" and "Treatm. × Spikel. × Variety" were just significantly different for relatively few parameters and DAI.

Especially the parameters Area, Fix Area, HACH Area, F_m, F_m/F_0, F_v, F_v/F_0, φ_{D_0}, φ_{E_0}, φ_{P_0} and φ_{Abs} seem promising based on a visual inspection. Although some parameters, *e.g.* φ_{E_0}, might slightly outperform F_v/F_m, the latter is a widely used CF parameter which can be obtained from more simple CF protocols than OJIP. It was thus chosen for further investigation.

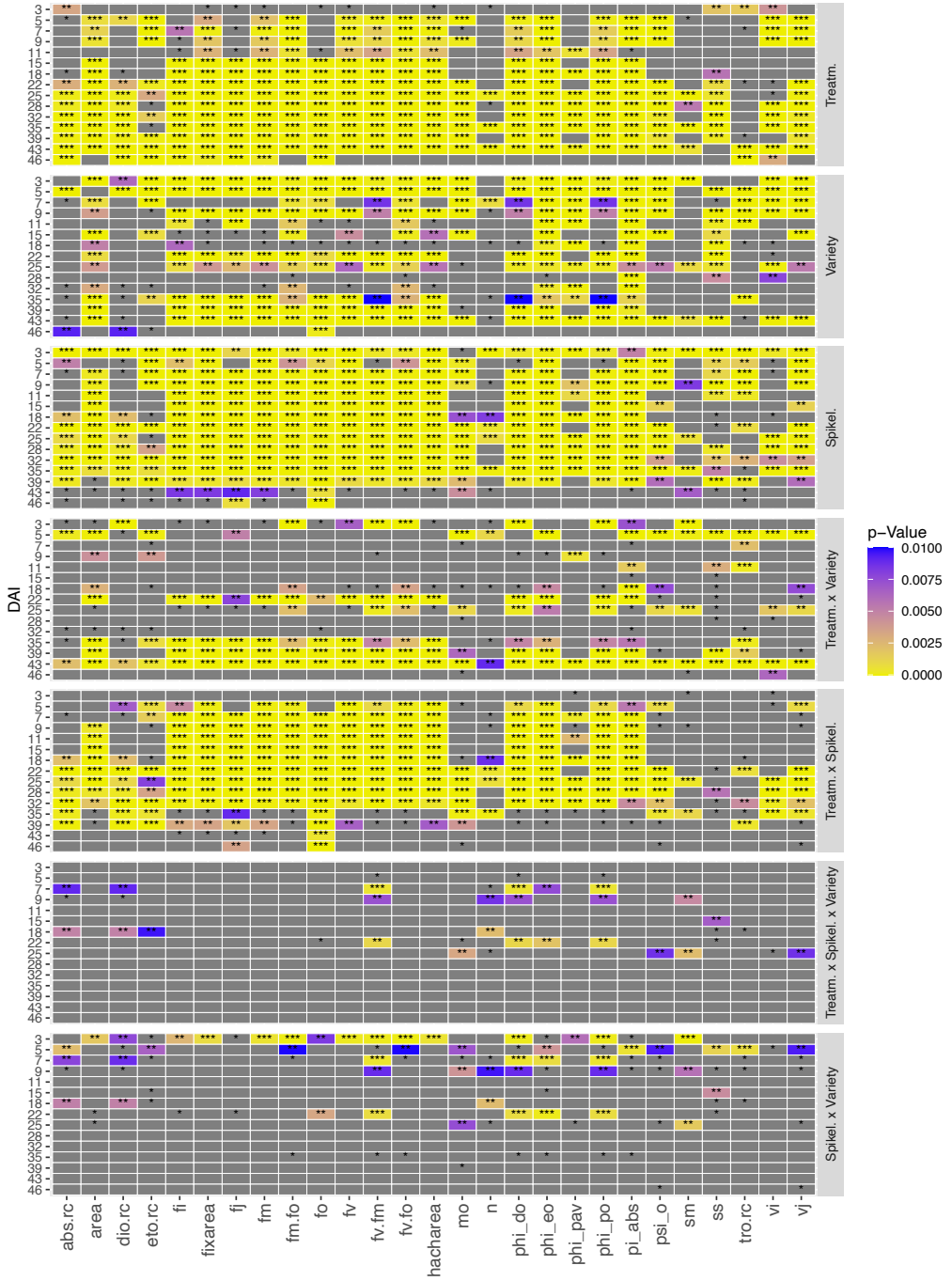


FIGURE 5.4: p-values of ANOVAs for 3 to 46 days after inoculation (DAI) on the greenhouse experiment for the factors "Treatm.", "Variety" & "Spikel." and their interactions. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. Colors indicate significance at $p < 0.01$.

F_v/F_m and visual ratings showed a very similar temporal development (Fig. 5.5) and well represent the disease progression based on visual inspection (*cf.* Fig. 5.3 & Fig. 5.5). The variability of F_v/F_m was mostly lower than that of visual ratings and more importantly showed a higher temporal continuity. On the central spike, where the inoculation took place, significant differences could be detected between inoculated and healthy plants after as early as 3 DAI. CH-NARA is known to be very susceptible to *Fusarium* (Strebel, Levy Häner,

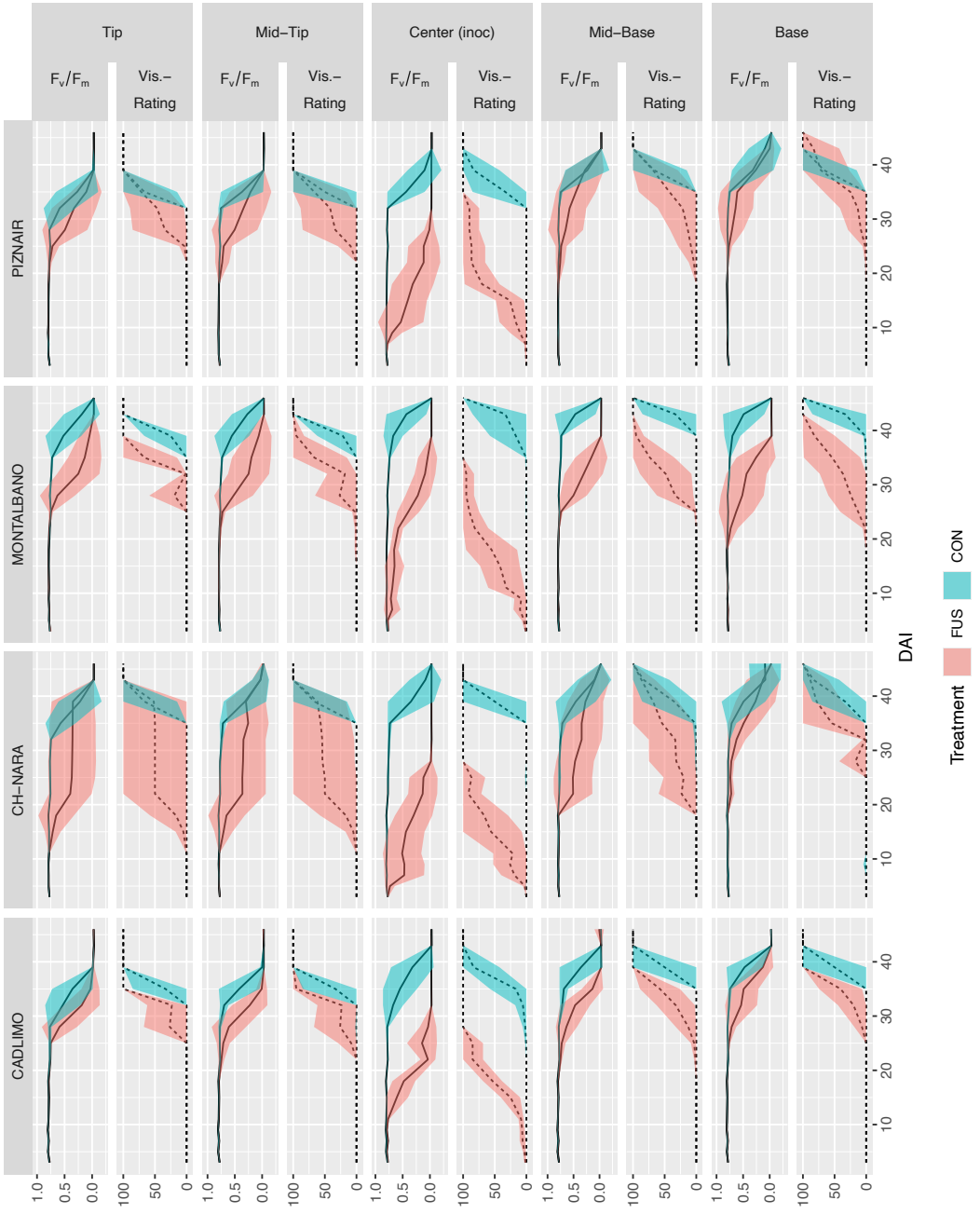


FIGURE 5.5: F_v/F_m and visual ratings for the greenhouse trial for four varieties and the five spikelet positions from 3 to 46 days after inoculation (DAI). Values are summarized by varieties and treatments (n = 6). Central lines indicate means and shaded areas mean ± standard deviation. Values were restricted for the range from 0 to 1 (F_v/F_m) and 0 to 100 (VisRating) respectively. FUS: Fusarium; CON: Control.

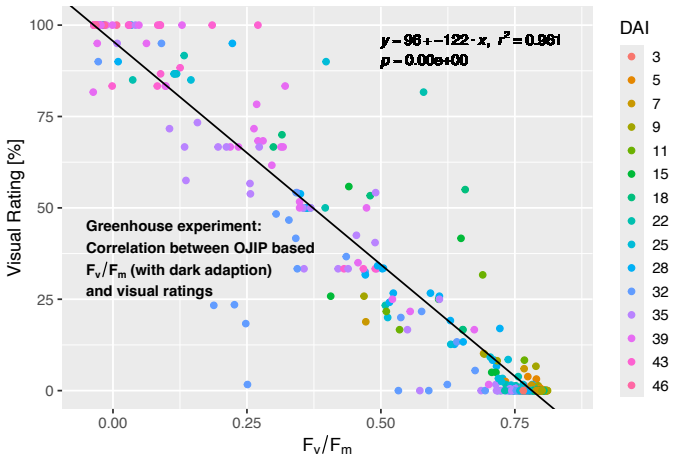


FIGURE 5.6: Correlation between visual ratings and F_v/F_m values for the greenhouse experiment. Values were grouped by variety, treatment, spikelet position and measurement event (DAI). Group-wise means were correlated with each other. Each dot represents the mean of the six replications of one genotype, in one treatment on one DAI and for one spikelet position. DAI: Days after inoculation.

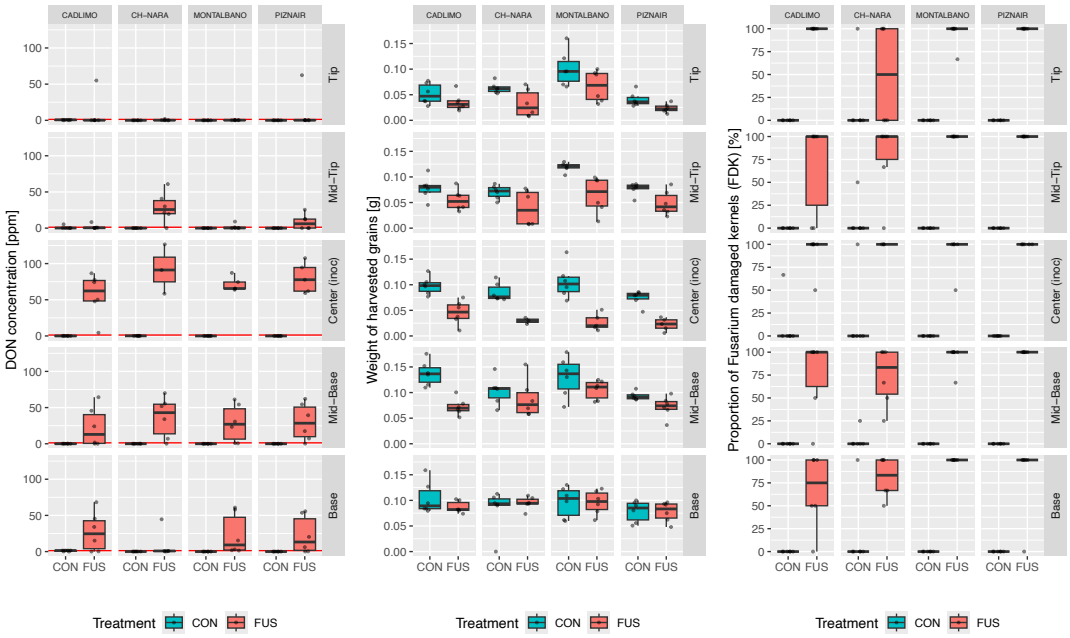
Watroba, et al., 2024), which is clearly visible in both, visual ratings and F_v/F_m values. The disease developed rapidly in the central spikelet and symptoms were detectable on the upper spikes (Mid-Tip, Tip) after approximately 12 DAI and on the lower spikes (Mid-Base, Base) after approximately 18 DAI. Infestations were stronger on top spikelets than for the lower spikes, which is characterized by a larger difference between the curves of the two inoculation treatments (FUS & CON). On the other three genotypes, trends were not as pronounced. MONTALBANO, which is supposed to be a resistant variety, showed a larger difference between inoculation treatments than the somewhat susceptible two varieties CADLIMO and PIZNAIR. Symptoms on upper and lower parts of the spike developed at the same time. CADLIMO and PIZNAIR maintained low *Fusarium* rating scores and a high F_v/F_m value long into the season and bleaching only started approximately 10 days before the onset of senescence of the spikes after 20 DAI. Visual rating and F_v/F_m started to increase/decrease at the same time for the individual varieties and treatments, thus, CF could not detect the *Fusarium* infestation presymptomatically.

Of the nine spikes that had not developed symptoms up to 11 DAI, four were PIZNAIR, three were CH-NARA, one was CADLIMO, and one was MONTALBANO. This could have contributed to the mild infestation of PIZNAIR, while the infestation of CH-NARA was, nevertheless, severe.

When visual ratings and F_v/F_m were grouped by treatment, variety, and measurement events, and group-wise means correlated with each other, the very strong correlation ($r^2 = 0.96$) confirmed the good correspondence between the two (Fig. 5.6). The strong correlation was also driven by clusters of data points with visual ratings of 0 and 100 respectively. But also after removing points with 0 or 100 ratings from the data, the correlation remained very strong ($r^2 = 0.82$, Fig. S4.3).

The DON content, as visual ratings and CF parameters, was also highest for the central spikelets of the inoculated treatment and in general for the variety CH-NARA (Fig. 5.7a). In contrast to visual ratings and F_v/F_m, the DON concentration of upper spikelets was very low, except for CH-NARA. For lower spikelets, the DON concentration was very similar, regardless of the differences for visual ratings and F_v/F_m. The "Base" spikelet of CH-NARA accumulated almost no DON.

Grain weight was generally lower for the inoculated spikelets. The difference was generally



(a) (b) (c)

FIGURE 5.7: Spikelet-wise DON content (a), grain weight (b) and FDK rate (c) for the greenhouse experiment. The red line in (a) shows the legal limit of 1250 µg DON kg^{-1} flour (Zorn et al., 2018).

most pronounced on the "Center" spikelet and the least pronounced on the "Base" spikelet (Fig. 5.7b). FDK rate was close to or at 100 % for almost all inoculated spikelets (Fig. 5.7c). AUC_{norm} was only correlated with DON for inoculated samples and the correlation was moderate ($r^2 = 0.35$).

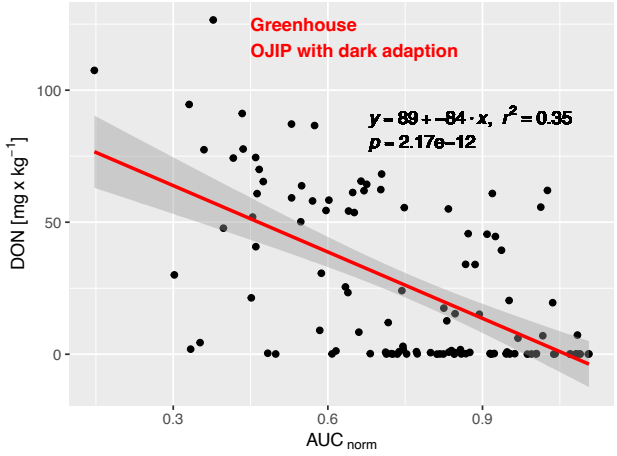


FIGURE 5.8: Correlation between DON content and AUC_{norm} of F_v/F_m values for single spikelets of the greenhouse trial. Only samples within the inoculated treatment were correlated.

5.3.2 Field trial - Infestation levels according to visual ratings

Inoculations worked well in all field experiments and led to very strong infestations in Changins in the rather wet year 2021 (Fig. 5.2a & S4.4) as well as in the hot and dry year 2022 (Fig. 5.2b,

S4.5 & S4.12). At Cadenazzo, infestation levels were low to intermediate, although the season was wet too (Fig. 5.2c & S4.6).

5.3.3 Field trial - FluorPen with dark adaptation

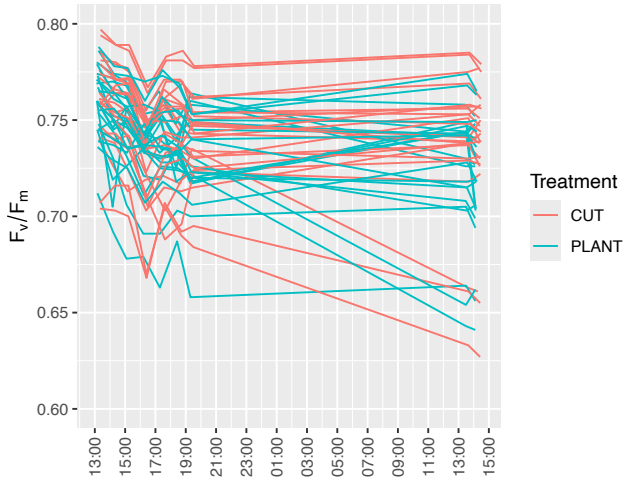


FIGURE 5.9: F_v/F_m values of the destructive experiment over time for spikes measured on plants (PLANT) and for detached spikes (CUT). Three spikes of four varieties were measured on two positions each, resulting in 24 sequences of measurements for detached and connected spikes respectively.

To measure whether detaching the spikes for measurements had a systematic effect on F_v/F_m, three tillers (upper internode with spike) per variety of the greenhouse experiment were detached from the plant and put in water. F_v/F_m values of detached tillers were not systematically different compared to spikes still attached to the plant (Fig. 5.9). There was considerable variability between the measurements, but the variability was not attributable to whether the spikes were detached or not for 26 h after being detachable from the plant, based on a visual inspection of the data.

ANOVA p-values of the OJIP parameters (Eq. 5.2) on the CHA21 field trial showed a similar pattern as for the greenhouse experiment with F_v/F_m being a promising parameter to distinguish inoculation treatments, varieties, and their interaction (Fig. 5.10). Thus, also for the field trial, the F_v/F_m parameter was further investigated.

F_v/F_m from the OJIP protocol was not significantly different for inoculated and non-inoculated plots for the first date of measurement at Changins in 2021 (2021-06-07, Fig. 5.11). Already at the second date (2021-06-07) the more susceptible varieties (*e.g* CH-NARA, Table S4.1) started to show significant differences between inoculation treatments, which were even more pronounced for later measurement dates up to senescence. More resistant genotypes such as *e.g.* ARINA or ROSATCH developed significant differences between treatments only later, before the differences between treatments narrowed again in the senescence process.

In Cadenazzo, all OJIP parameters were significantly different for the factor "Variety" at $p < 0.01$. For "Treatm." and "Treatm. × Variety", parameters were not different for most situations (Fig. S4.7). Only f_j and f_o were significantly different for "Treatm. × Variety" on 2021-06-17. F_v/F_m from the OJIP protocol was not significantly different for neither of the two measurement events (Fig. 5.12).

In 2022, OJIP with dark adaptation was only carried out at Changins and on two dates, but with 10 spikes per plot. The effects were again significant for most OJIP parameters for

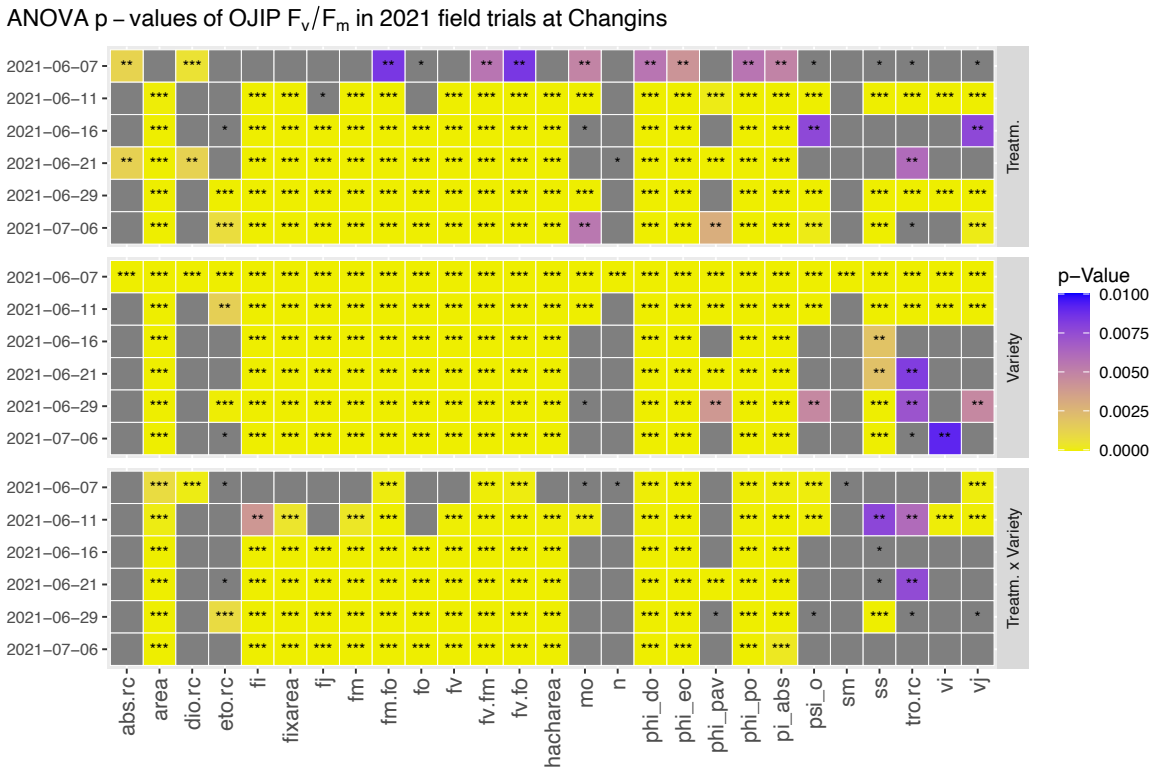


FIGURE 5.10: p-values of ANOVA on OJIP parameters of dark adapted measurements on field samples at Cadenazzo in 2022 for the factors "Treatm.", "Variety" and their interaction. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$, and colors indicate significance at $p < 0.01$.

inoculation treatments, varieties, and their interaction on the first date and for almost all parameters on the second date (Fig. S4.8 & Fig. S4.9).

5.3.4 Field trial - FluorPen without dark adaptation

The first field tests of the rapid in-field protocol to measure F_v'/F_m' parameter without dark adaptation at Changins in 2021 revealed that even suboptimal conditions without dark adaptation were allowing the derivation of CF parameters with significant differences between inoculation treatments.

F_v'/F_m' parameter was significantly affected by the inoculation treatments, varieties and their interaction at $p < 0.01$ for both dates, except for the inoculation treatment on 2021-06-24 (Fig. S4.10 & Fig. S4.11).

In 2022, the F_v'/F_m' parameter was significantly affected by the inoculation treatments, varieties and their interaction at $p < 0.01$ for all dates dates (Fig. 5.13 & Fig. 5.14).

When measurements were performed with a targeted rapid F_v'/F_m' protocol, the differences between inoculation treatments became even more evident (Fig. 5.15). The differences between treatments were more perceivable in the targeted approach on 2022-06-03 compared to a random approach on 2022-06-09 (Fig. 5.14) based on a visual comparison of boxplots, regardless of the fact that for the latter, symptoms were already six days more advanced. Inoculation treatments, varieties and their interction had a significant effect on F_v'/F_m' according to ANOVA at $p < 0.001$ (Eq. 5.2) .

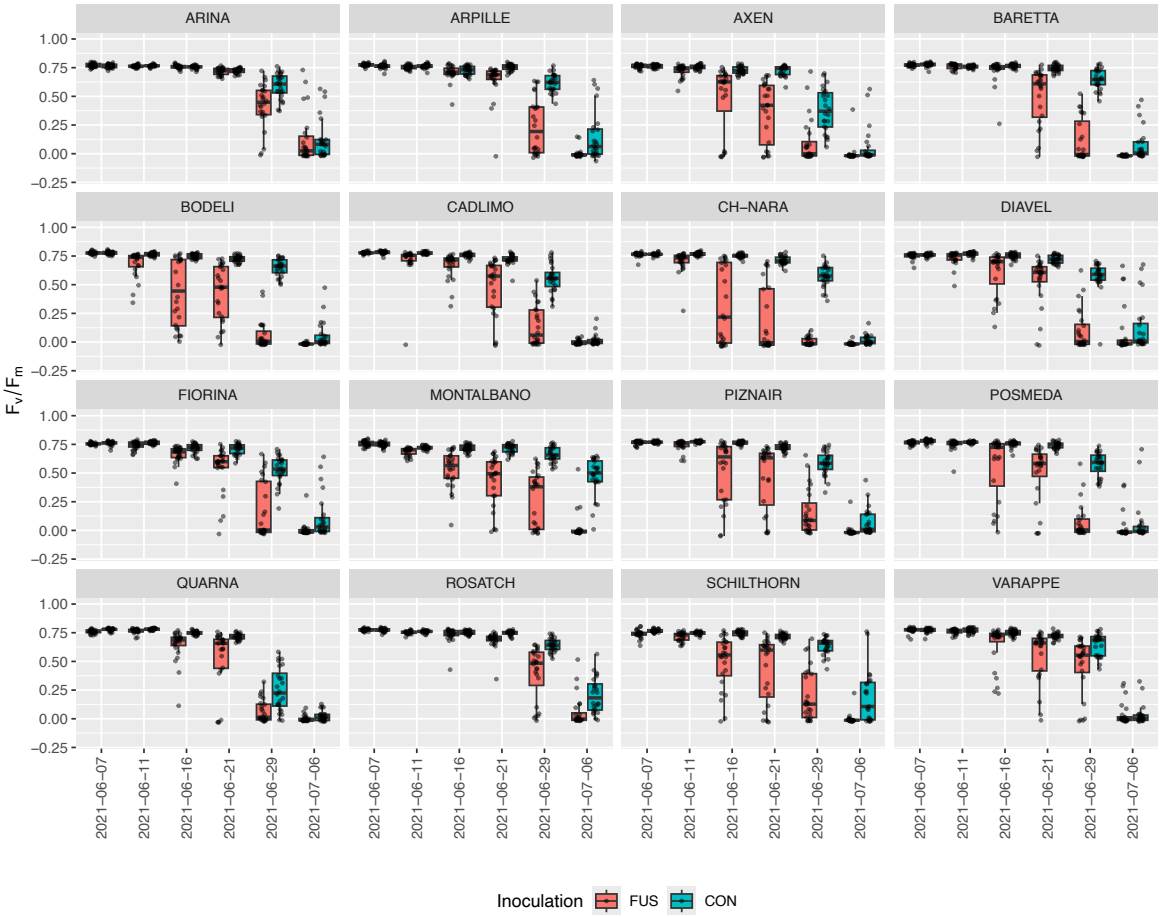


FIGURE 5.11: F_v/F_m parameter of OJIP data, Changins 2021. The 16 tiles represent the 16 wheat varieties tested over time. Dates are the individual measurement events. Inoculation treatments are indicated by color. 3 spikes were measured on spikelets toward the tip of the spike and on a central spikelet for each plot, resulting in n = 24 measurements from four replication per variety and inoculation treatment for each date of measurements (768 measurements for each date in total).

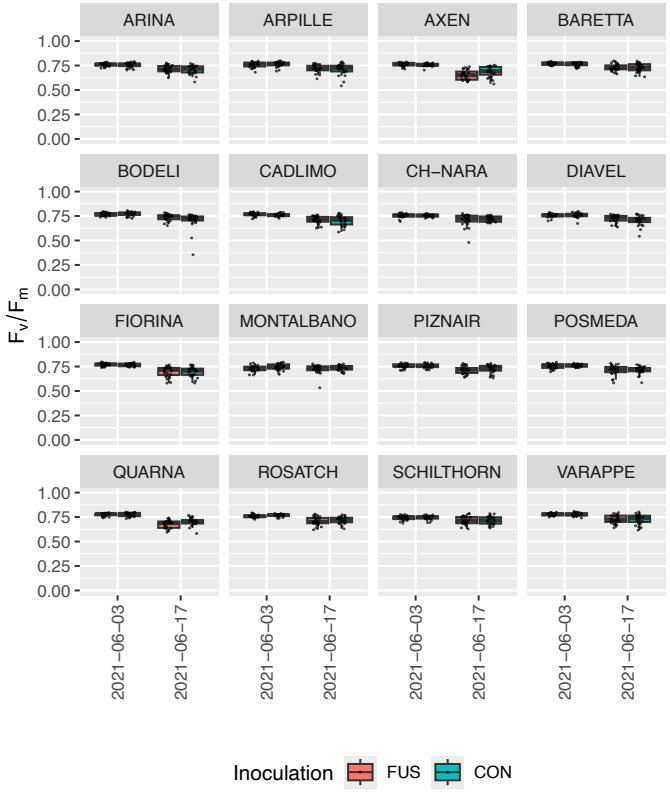


FIGURE 5.12: F_v/F_m parameter of OJIP data, Cadenazzo 2021. The 16 tiles represent the 16 wheat varieties tested over time. Dates are the individual measurement events. Inoculation treatments are indicated by color. 3 spikes were measured on spikelets toward the tip of the spike and on a central spikelet for each plot, resulting in n = 24 measurements from four replication per variety and inoculation treatment for each date of measurements (768 measurements for each date in total).

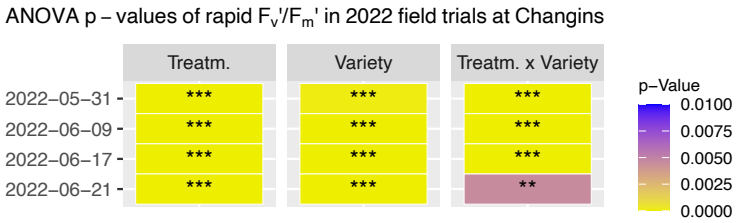


FIGURE 5.13: p-values of ANOVA on rapid F_v'/F_m' parameter without dark adaptation at Changins 2021 for the factors "Treatm.", "Variety" and their interaction. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$, and colors indicate significance at $p < 0.01$.

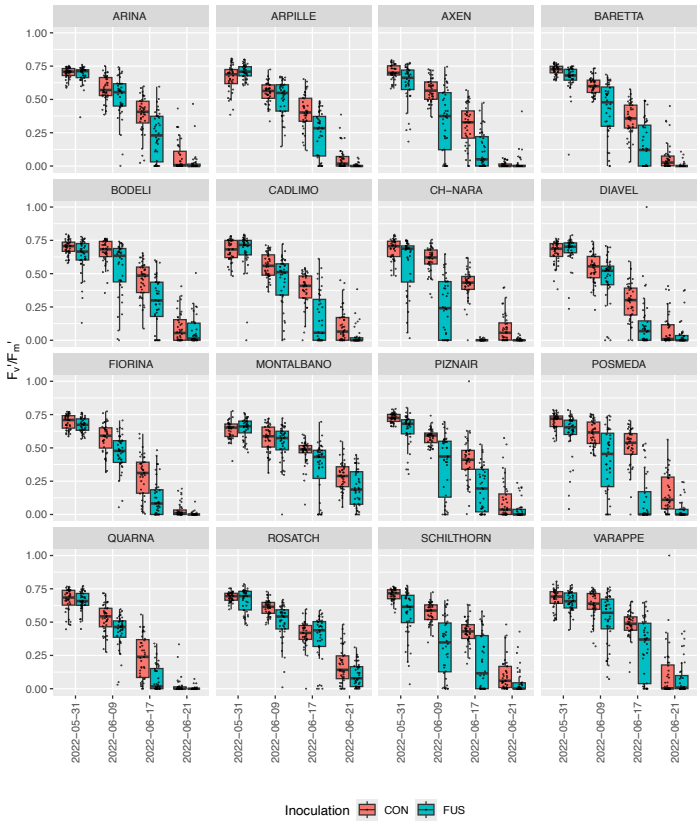


FIGURE 5.14: Rapid F_v'/F_m' parameter without dark adaptation, Changins 2022. The 16 tiles represent the 16 wheat varieties tested over time. Dates are the individual measurement events. Inoculation treatments are indicated by color. 10 spikes were measured on a spikelet from the central spike for each plot, resulting in n = 40 measurements from four replication per variety and inoculation treatment for each date of measurements (1'280 measurements for each date in total).

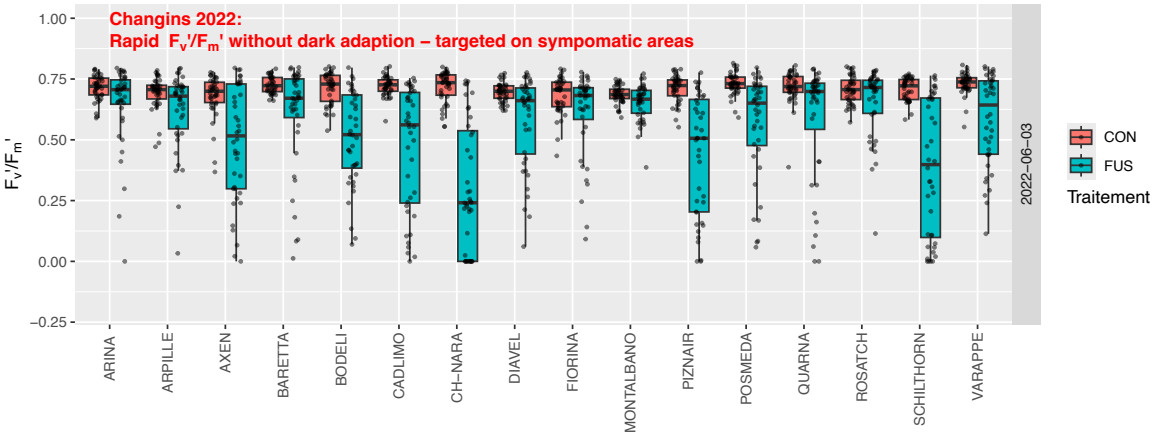
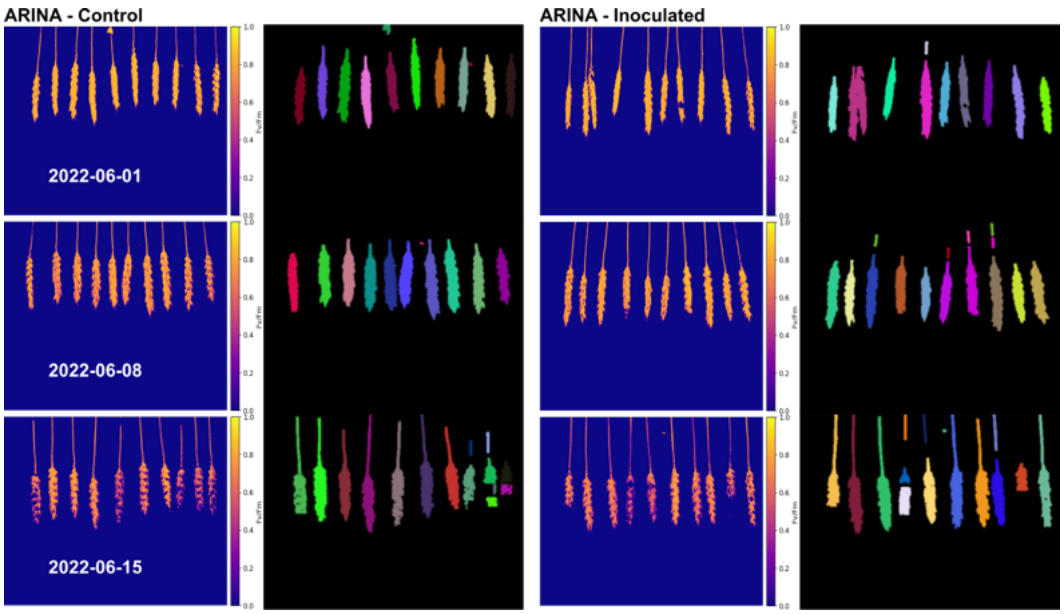
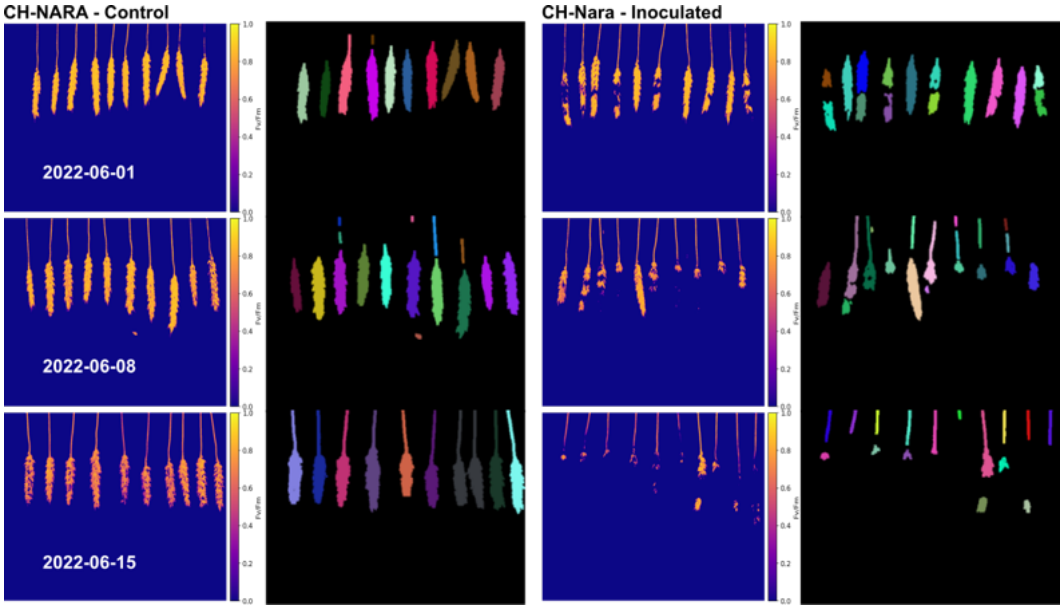


FIGURE 5.15: P-values of rapid F_v'/F_m' parameter in 2022 Changins field trials when using a targeted approach. The FluorPen was positioned on areas of the wheat spikes with visible *Fusarium* symptoms. 10 spikes were measured on a spikelet from the central spike for each plot, resulting in n = 40 measurements from four replication per variety and inoculation treatment (1'280 measurements in total).



(a)



(b)

FIGURE 5.16: Rover FluorCam data for the varieties ARINA (a) and CH-NARA (b). CF images were taken on three dates for control (non-inoculates) and inoculated spikes. The columns on the left show each treatment show raw values of F_v/F_m in the figure). Columns on the right show the masks for the individual connected components found on on the CF images and used for analysis.

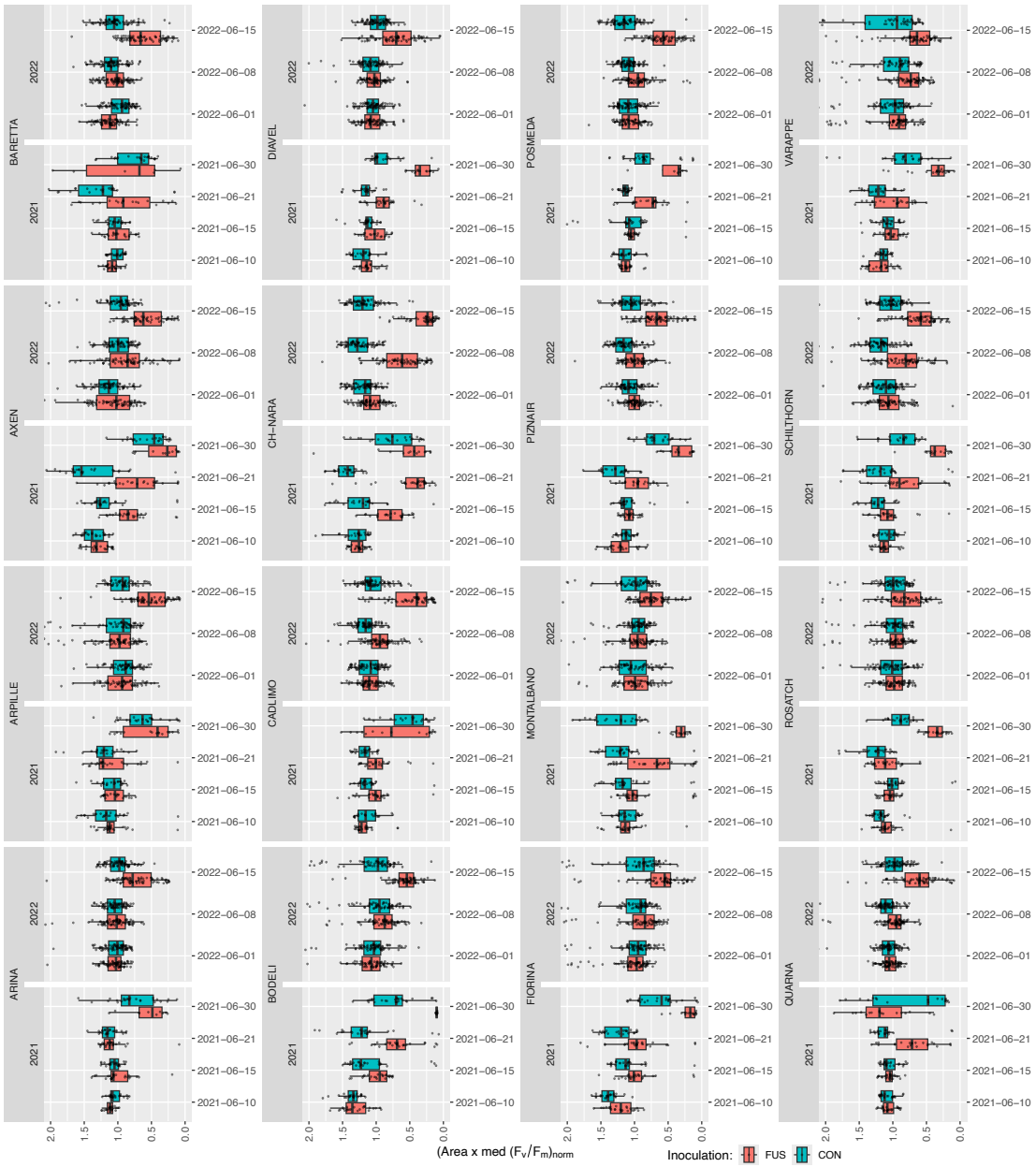


FIGURE 5.17: $Area \times med(F_v/F_m)_{norm}$ index from FluorCam images of spikes for Changins in seasons 2021 and 2022. The 16 tiles represent the 16 wheat varieties tested over time. Dates are the individual measurement events. Inoculation treatments are indicated by color. In 2021, 3 spikes were measured for each plot, resulting in n = 12 measurements from four replications per variety and inoculation treatment for each date of measurements (384 spikes for each date of 2021 in total). In 2022, 10 spikes were measured for each plot, resulting in n = 40 measurements from four replications per variety and inoculation treatment for each date of measurements (1'280 spikes for each date of 2022 in total).

5.3.5 Field trial - Rover data

CF imaging with the stationary rover FluorCam data allowed to well track the development of *Fusarium* symptoms and to distinguish heavily infested from healthy spikes. In CF images (*e.g.* Fig. 5.16), non-inoculated spikes maintained higher F_v/F_m values on the whole spikes and the F_v/F_m intensity decreased only toward maturity, but still, the whole spikes showed photosynthetic activity. For the very resistant variety ARINA, the F_v/F_m intensity decrease was very similar between non-inoculated and inoculated spikes. For the very susceptible variety CH-NARA, the decrease was much more pronounced for the inoculated spikes. Already on 2022-06-01, inoculated CH-NARA spikes featured patches without photosynthetic activity. On 2022-06-08, already large proportions of the spikes showed little to no photosynthetic activity. Finally, on 2022-06-15, photosynthetic activity was mainly limited to the peduncles and culms.

$Area \times med(F_v/F_m)_{norm}$ summarized the photosynthetic activity and the status of PSII in one metric, which was used to analyze the results statistically (Fig. 5.17). When comparing the two varieties ARINA and CH-NARA just examined on CF images, (Fig. 5.16), $Area \times med(F_v/F_m)_{norm}$ values remained similar for ARINA up until the last measurement date in both years and even on the last date, the quartiles of boxplots of the two inoculation treatments still overlapped. In contrast, the quartiles of CH-NARA did not overlap on the second date already.

The ANOVA analysis (Eq. 5.2) showed, that the inoculation treatment significantly impacted the $Area \times med(F_v/F_m)_{norm}$ values except for the first measurement dates in each year. The interaction of inoculation and variety always had a significant effect, while the variety had a significant effect on the values on most but not all dates at $p < 0.001$ (Fig. 5.18).

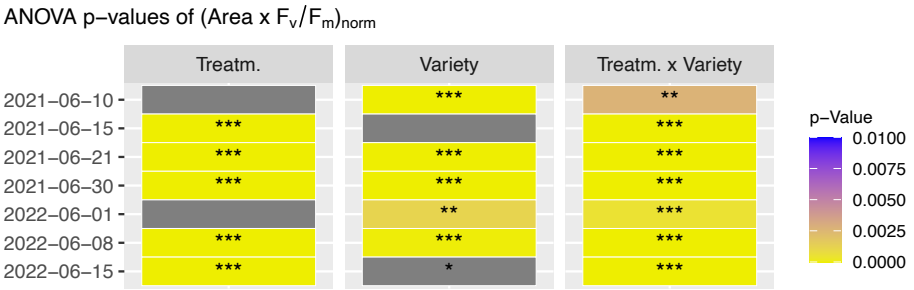
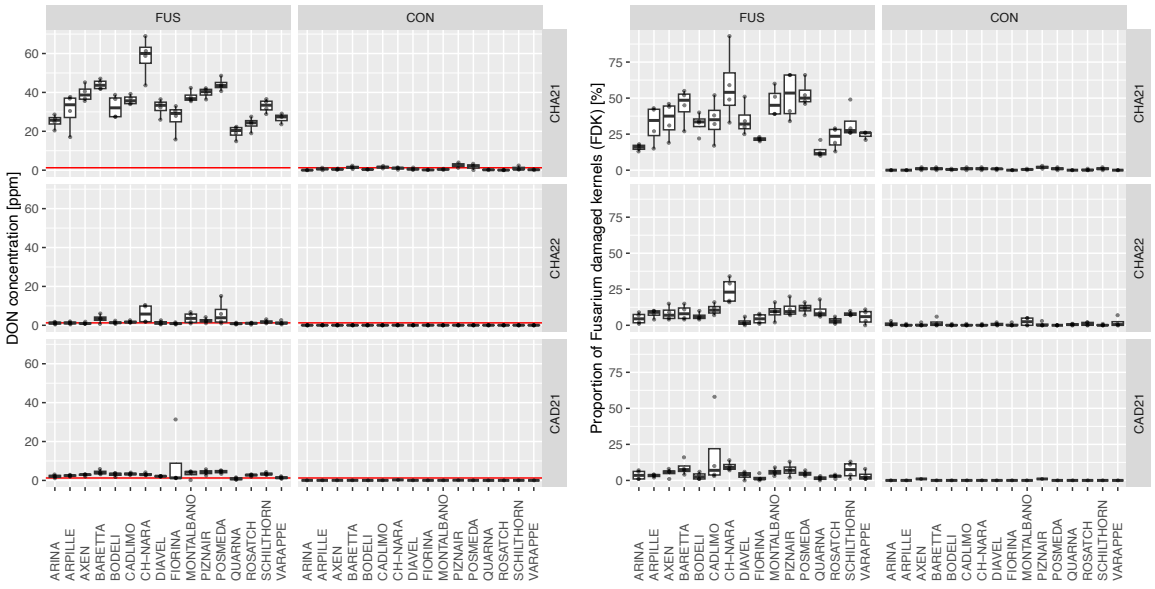


FIGURE 5.18: p-values of ANOVA on $Area \times med(F_v/F_m)_{norm}$ index for the factors "Treatm.", "Variety" and their interaction. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$, and colors indicate significance at $p < 0.01$.

5.3.6 Field trial - DON content

DON content was above the legal threshold of 1250 µg DON kg^{-1} flour (Zorn et al., 2018) for all varieties in CHA21 and CAD21 except for the variety QUARNA in CAD21 (Fig. 5.19a). In CHA21, even the non-inoculated control plots were close to or above the threshold, though the visual rating just revealed a mild infestation of some spikes in the control treatment (Fig. S4.4). BARETTA, CH-NARA, MONTALBANO and POSMEDA were above the threshold in CHA22, while the other varieties were around the threshold.

In CHA21, CH-NARA showed the highest concentration of DON, but even QUARNA, which showed the lowest DON concentration, was about 16 times above the threshold. Also in CHA22, CH-NARA showed the highest DON concentration, which was in line with the visual ratings, which were also highest for CH-NARA inf CHA21 and CHA22 (Figs. S4.4 & S4.5). Also BARETTA, MONTALBANO and POSMEDA were among the most contaminated in



(a) (b)

FIGURE 5.19: Deoxynivalenol (DON) (a) and proportion of *Fusarium* damaged kernels (FDK) (b) on field trials. Columns indicate different treatments, rows different years and locations. The red line in (a) shows the legal Swiss limit of DON concentration of 1250 µg DON kg^{-1} flour (Zorn et al., 2018).

both years at Changins, although they showed rather intermediate visual ratings. For CAD21, FIORINA featured the highest DON concentration, although it showed very low visual ratings.

FDK proportion corresponded well to the DON content for CHA21, but for CHA22 and CAD21, the correspondence was low. Only CH-NARA in CH22 showed an exceptionally high DON concentration and FDK proportion at the same time.

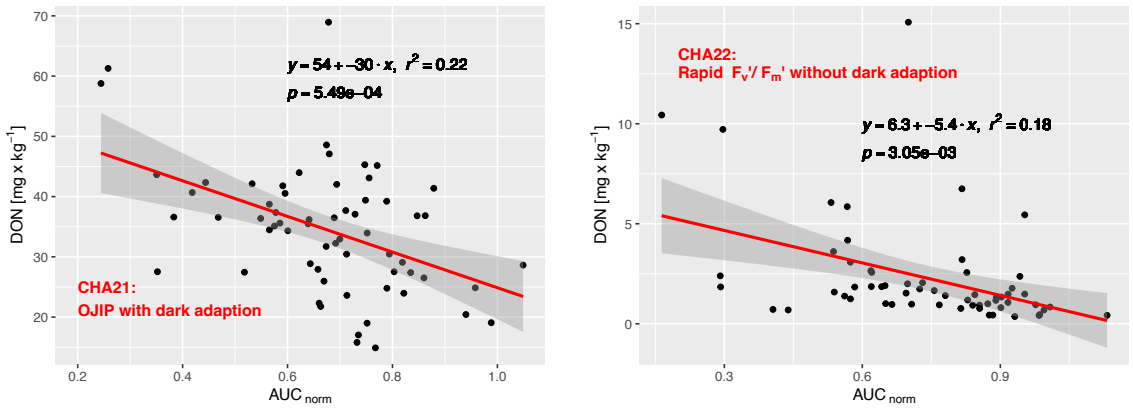
5.3.7 Field trial - DON content and AUC

If measurements covered the infestation period from inoculation to senescence, the plot-wise AUC_{norm} could be calculated and correlated with the DON content. This was the case for F_v/F_m measured with the OJIP protocol on dark adapted samples on CHA21 (Fig. 5.20a) as well as for the F_v'/F_m' parameter derived from the rapid in-field approach without dark adaptation on CHA22 (Fig. 5.20b). The correlation was significant ($p < 0.01$), negative and moderate in both cases, with $r^2 = 0.22$ for F_v/F_m on CHA21 and $r^2 = 0.\ 18$ for F_v'/F_m' on CHA22. Only samples within the inoculated treatment were correlated.

5.4 Discussion

In this study, the different CF methods were tested for their suitability to track *Fusarium culmorum* infestations on wheat. An initial greenhouse trial with measurements with the FluorPen at high temporal resolution, revealed the close relationship between visual ratings and CF parameters such as F_v/F_m. Due to the need for dark adaptation to derive good-quality CF measurements, transfer to field conditions was and remains challenging.

All CF devices and protocols used in this study were able to track infestations with *Fusarium culmorum* in the greenhouse, but also in field conditions, when infestation levels were high. This was the case for the greenhouse experiment and for both years in Changins



(a) (b)

FIGURE 5.20: Correlation between AUC_{norm} and DON content for OJIP based F_v/F_m parameter from dark adapted measurements of CHA21 (a), and for rapid F_v'/F_m' without dark adaptation of CHA22 (b). Only samples within the inoculated treatment were correlated.

(Figs. S4.4 & S4.5). However, when infestation levels were lower, as in Cadenazzo (Figs. S4.6), the FluorPen failed to find significant differences between treatments.

This was most likely due to the relatively low number of spikelets per plot measured with the FluorPen (n = 3). However, also the visual rating, which was done more quickly and thus allowed more spikes per plot to be rated (n = 15), was limited in capturing low levels of infestation. Nevertheless, even the lower infestation levels at Cadenazzo led to DON concentrations around the legal threshold or above, while they were far above the threshold in CHA21. DON concentration in CHA22 was very similar to CAD21, although visual symptoms were much more severe in CHA22.

Thus, the number of spikes measured per plot is the limiting factor, rather than the precision of the approach, as infected spikes might escape detection when not all spikes are infested. This was also confirmed with the targeted rapid F_v'/F_m measurements, where the difference between the inoculation treatments was increased by selecting an infested area of the spike for measurements. For visual field ratings, alternative nonlinear rating scales are known, where a score from 0 to 9 does not represent infestation levels from 0 % to 100 % but 9 rather distinct classes of infestation, similar to what was used on leaf level in V. Michel (2001) and on spike level in Mustafa et al. (2023). However, the human eye is capable of detecting the one single symptomatic spike within a plot. Digital phenotyping approaches would need to scan all spikes with a perfect accuracy to achieve the same result, which would take an unrealistic amount of time for all known approaches.

There are new approaches for the phenotyping of *Fusarium* spp. under field conditions that are not spike-by-spike, but image-based and do not depend on dark adaptation. Hong et al. (2022) analyzed RGB images with deep learning and attained a classification accuracy of 93.69 %, which might be sufficient for an accurate tracking of the disease at higher infestation levels, but not at lower levels. However, for a correct characterization of genotype resistance toward *Fusarium* spp., low levels of infestation must also be monitored correctly. Still, RGB sensors have major advantages over other sensors. They are relatively inexpensive and can capture images quickly, which enables fast measurements and thus the coverage of large areas in a short time (Grishina et al., 2024).

In the context of CF, Lauterberg et al., 2024 proposed a new CF imaging approach, based on theoretical non-photochemical quenching, which does not need dark adaptation but can be applied in daylight with fluctuating light conditions. Although the respective

study was conducted on drought stress, it can be hypothesized that it would also allow to detect *Fusarium* infestations. However, while drastically decreasing the time required for measurements, individual measurements still take more than 5 min, which is still long, considering that measuring a field experiment as used in the study at hand, consisting of 128 plots, would take approximately 11 h.

The use of hyperspectral imaging, in contrast, comes with the benefit of additional spectral features that can be exploited, but with the disadvantage of large volumes of data that have to be processed, which is more time-consuming, as demonstrated by Almoujahed et al. (2022). In their study, which included eight varieties, they also showed that the classification accuracy for different varieties ranged from 74.1 % to 99 %. A variety-dependent accuracy might bias the resistance classification and thus make a sensible application of such an approach problematic.

Drone-based approaches have a high performance with regard to field surface covered per time; however, for genotype characterization, their precision might be very limiting. For example, in Francesconi et al., 2021, the minimal infestation level detected was 20 %. At this level, a DON content above the legal threshold might be expected, as seen in the CAD21 trial, where the DON threshold was mostly reached, while the average infestation level was below 20 %. In addition, the drone supported hyperspectral imaging approach of H. Zhang et al. (2022) reached an accuracy of 85 % when classifying regions into the three classes "mild", "moderate" and "severe", without a differentiation of "mild" infections and healthy. Thus, while such approaches might have the potential to inform field management decisions (*e.g.* Elke Bauriegel and Herppich, 2014; Francesconi et al., 2021; Elias Alisaac and Mahlein, 2023), they not yet proved the sensitivity needed to characterize resistance levels in wheat variety testing.

Nevertheless, the different methods tested in this study had different advantages and disadvantages, which are discussed below.

5.4.1 FluorPen with dark adaptation in the greenhouse

In the greenhouse experiment, infestations could be observed with high temporal resolution. With a known infestation site, the disease progression along the spike could be tracked. Like that, experience with the FluorPen in optimal dark adapted conditions could be gained before transferring the method to the field.

The CF parameters corresponded well to the visual ratings. However, even though CF has the potential to detect changes in plant status that are usually hidden to the human eye (Valcke, 2021) no presymptomatic detection was possible, which would matter in the context of early detection for timely fungicide treatment (Francesconi et al., 2021; Elias Alisaac and Mahlein, 2023). CF parameters became increasingly similar between inoculation treatments toward the late stages of senescence. It therefore can be hypothesized, that CF based approaches have little specificity for *Fusarium* spp., and cannot distinguish between *Fusarium* and senescence or other diseases that affect the wheat spike in a similar way, such as *Microdochium nivale* (Vincke et al., 2023).

Multiple OJIP CF parameters were well suited to distinguish between inoculated and healthy spikes as early as 5 DAI. Some OJIP parameters were also distinguishing between genotypes, especially for dates shortly after inoculation. Thus, CF values not only depend on the level of infestation, but must always be interpreted with respect to genotypes, as the CF parameters are genotype specific (Keller et al., 2019).

While AUC of F_v/F_m was correlated with DON content, the relationship between DON content and symptoms is not straightforward. For example, spikelets above the inoculation site showed similar visual rankings and F_v/F_m values compared to spikelets below the inoculation site (Fig. 5.5), but rather low DON concentrations on infected spikelets (Fig. 5.7a). Furthermore,

PIZNAIR had relatively low visual *Fusarium* ratings but a very similar DON content as CHNARA, which had high *Fusarium* ratings. At the same time, the reduction in grain weight (Fig. 5.7b) and the FDK proportion (Fig. 5.7c) were not directly related to the rating severity. Furthermore, when comparing AUC_{norm} with DON, very high DON content was associated with very high AUC_{norm} values and vice versa (Figs. 5.8 & 5.20), which significantly weakened the correlation between the two. This confirms other studies which concluded that visual *Fusarium* spp. symptoms are not a reliable predictor of DON content (Schlang et al., 2008; Ajigboye et al., 2016; Mesterhazy, 2020).

The use of plastic bags to reinforce infestations on asymptomatic spikes would have been a severe bias if the greenhouse experiment had been about resistance characterization. However, the goal of the experiment was to test the FluorPen and the correspondence between visual observations and CF parameters, for which symptomatic spikes were absolutely necessary.

Applying the FluorCam in addition to the FluorPen to the spikes of potted plants would have allowed for a more quantitative comparison between the two devices. However, the FluorCam required the horizontal alignment of the samples. Aligning spikes of plants in pots horizontally carries the risk of damaging them and having to end the measurement series prematurely, which is why we only tested FluorPen in the greenhouse experiment.

5.4.2 FluorPen with dark adaptation in the field

The transfer of the method to the field came with multiple challenges. The larger, but still relatively small number of varieties tested needed to be replicated in a well-designed experiment, to avoid artifacts of spatial heterogeneity or temporal trends of CF parameters throughout the day to impact the results. This led to a large number of spikes being measured. Shading of hundreds of samples from multiple genotypes and replications prior to measurements is impractical in a day-to-day variety testing routine. Thus, the destructive approach was applied as a compromise where samples from a field were measured under controlled conditions.

While CF values are highly variable over time and sensitive to physiological changes of the plant, CF parameters with relatively little variation within genotypes and inoculation treatments were derived from detached spikes.

When AUCs were calculated from F_v/F_m values on CHA21, moderate correlations with DON were achieved, confirming the basic functioning of the method. But even though the destructive approach worked, and F_v/F_m was shown to well track visual ratings, the approach failed to find a difference between inoculation treatments in low infestations. This is most likely due to the limited number of CF measurements taken within each plot. With several dozens, no to say hundreds of measurements per plot, the sensitivity of the approach could possibly be increased, but again, this would to be too labor-intensive for a day-to-day variety testing routine.

5.4.3 FluorCam with dark adaptation in the field

These considerations also apply to the FluorCam approach. However, the FluorCam images provide a spatially integrated analysis of the spikes instead of the point measurements of the FluorPen. This reduces the risk of missing the infection on a spike. The relatively simple index $Area \times med(F_v/F_m)_{norm}$ allowed one to characterize infestation levels. With a more bespoke image analysis pipeline and the inclusion of deep learning, the segmentation and analysis of CF parameter intensities could be improved (*e.g.* Almoujahed et al., 2022; Bannihatti et al., 2022; Hong et al., 2022).

Yet, the measurements were similarly time-consuming as the FluorPen measurements with dark adaptation. Equipping an autonomous field phenotyping mobile (Qiu et al., 2019; Xu and C. Li, 2022) that could operate at night in dark adapted conditions with a FluorCam

(Lorence and Medina Jimenez, 2022) would allow efficient application of CF imaging. Also, fusion of CF data with other data types such as RGB or hyperspectral images could increase the performance (C. Zhang et al., 2022; Mustafa et al., 2023), efficiency and sensitivity of the approach, as *e.g.* segmentation of objects in a single-channel image like CF images is more challenging than on three-channel RGB images.

5.4.4 FluorPen without dark adaptation as a rapid field protocol

The rapid F_v'/F_m' protocol without dark adaptation was tested because both methods with dark adaptation, FluorPen and FluorCam, were utterly time-consuming. Compared to F_v/F_m with dark adaptation, F_v'/F_m' showed a higher variability of the data within varieties and inoculation treatments. Nevertheless, it was significantly different between inoculation treatments, varieties, and treatment × variety interactions in high level infestations. When AUC were calculated from F_v'/F_m' values on CHA22, also moderate correlations with DON were achieved. The correlation was weaker than the for F_v/F_m values in CHA21, but this might have largely been due to much lower DON concentrations found in CHA22 compared to CHA21.

When using a targeted approach with the FluorPen pointing at the symptomatic parts of the spike, the difference between inoculated and healthy spikes was even more pronounced. This confirmed the findings of the FluorPen and FluorCam procedures with dark adaptation, where CF well tracked the visual symptoms of *Fusarium* spp. symptoms, but the problem lied in finding the symptoms on field plots, where hundreds of spikes would need to be measured at low infestation levels, to find symptomatic ears and statistically significant differences between infested and noninfested plots.

5.4.5 Inoculum used

FHB can be caused by different *Fusarium* species and is very often associated with *Fusarium graminearum* and *Fusarium culmorum* (H. Buerstmayr et al., 2009; M. Buerstmayr et al., 2020; Miedaner et al., 2008; Oldenburg and Ellner, 2015; Trail, 2009). However, resistance to *Fusarium graminearum* is highly correlated to resistance to *Fusarium culmorum* (E. Alisaac et al., 2018; Mesterhazy, 2020). In addition, both produce very similar symptoms and a microscopic examination is needed for a distinction between the two. Thus, if a phenotyping approach is capable of identifying wheat genotypes resistant to one *Fusarium* species, the genotype is highly likely to also be resistant to the other species. This study only used *Fusarium culmorum* as this *Fusarium* species is abundant on farms around the Agroscope research station in seasons with conducive conditions and as *Fusarium culmorum* is easily cultivated to produce inoculum for artificial inoculations.

5.5 Conclusion

Many studies are concerned with improving the precision and throughput of *Fusarium* detection, but few approaches are truly designed for day-to-day practice under field conditions in the context of disease resistance in variety testing. In this study, a simple FluorPen for tracking *Fusarium culmorum* infestations was first tested extensively in the greenhouse and the findings were then transferred to a field trial, where field grown spikes were harvested and measured indoors. In addition, these harvested ears were also measured with a FluorCam CF imaging device. Both methods were able to detect high infestation levels, but both were also time consuming to apply, and it became apparent that a much higher number of measurements would be necessary to correctly characterize low infestation levels. Consequently, a rapid CF

approach with the FluorPen and a full-field protocol without dark adaptation was tested, which increased the efficiency but also variability of measurements. Thus, the trade-off between precision and throughput persisted. However, this trade-off is not unique to the CF methods presented but also applies to all known methods. Here, RGB images, together with light and fast but precise computer models, are a promising option to increase throughput. But such models still need to be developed first. For a high precision of *Fusarium* characterization, the combination of CF sensors, *e.g.* with RGB cameras could be optimized and implemented on autonomous phenotyping mobile for an increased throughput.

Authors' contribution

Simon Treier: trials, conceptualization, methodology, software, formal analysis, visualization, writing – original draft. Fabio Mascher: trials, methodology, review & editing. Juan M. Herrera: project administration, funding acquisition, conceptualization, supervision, methodology, acquisition, writing – review & editing. Romina Morisoli: trials, writing – review & editing. Achim Walter: writing – review & editing.

Acknowledgments

We thank Johanna Antretter, Fernanda Arelmann Steinbrecher, Matthias Schmid and Julien Vaudroz for rating of phenology, *Fusarium* ratings and support of chlorophyll fluorescence measurements; Stefan Kellenberger, Alain Handley-Cornillet and Amandine Fasel for preparing the inoculum and the support in DON measurements; Nicolas Widmer and his team as well as Yann Imhoff for field management.

Funding

This study was in part supported by the two H2020 projects InnoVar and Invite.

6 General discussion and conclusion

As stated in the Introduction, translational research such as the development of lean phenotyping approaches is time-consuming and bound to ignore many shortcuts and simplifications that would be perfectly acceptable for basic research. One of the biggest challenges is the collection of high-quality reference ground-truth data over multiple seasons, which is very time-consuming. Nevertheless, in this thesis, three methodologies of optical lean phenotyping were introduced, improved, or tested in the context of lean phenotyping (Fig. 1.2) for wheat variety testing. In this chapter first, the contribution of individual methods to the field of lean phenotyping is debated and how they could be further pursued. Then, some general challenges of lean phenotyping will be discussed. It will be shown how the methodologies are interlinked and how they could be parts of a larger phenotyping concept for the future of variety testing.

6.1 Contribution to the field of lean phenotyping and possible future pathways

The starting situation for all the approaches presented here was similar. Although publications had already demonstrated some potential for the individual approaches, many challenges remained, and it was unclear how the approaches could be utilized in daily practice of variety testing.

6.1.1 Drone-based lean phenotyping of canopy temperature

Airborne thermal imaging has been proposed for plant phenotyping for many years. Relatively heavy radiometrically calibrated thermal cameras must be carried by helicopters or larger drones, but they do not need calibration panels in the field (Deery, Rebetzke, Jimenez-Berni, James, et al., 2016). When lightweight sensors are used, they are more prone to disturbances, and radiometric calibration panels in the field are needed. Both complicate the day-to-day operation of airborne thermography. A helicopter is expensive to use, while distributing and measuring calibration panels during flight comes with additional work in the field as well as in post-processing. On top of that, field management, is disrupted if such panels are installed in the field.

The approach used in this thesis can be applied with an off the shelf drone with an already well-integrated thermal camera system. Thus, initial material and integration costs are relatively low and no specific technical knowledge is needed other than to pilot the drone and to use the correct camera settings to acquire the thermal data. With this equipment, relative canopy temperature (CT) differences can be reliably measured even without the use of field reference panels. This significantly increases the TRL of airborne thermography and reduces the hurdles for technology adoption by variety testing organizations, since the method proposed in this thesis itself could be largely automated.

Overcoming the need of reference panels would also allow for an automatization of the whole process, since no staff is needed in the field during measurements. DJI is providing autonomous drone docking stations (*e.g.* DJI Dock 3, SZ DJI Technology Co. Ltd., China) that can be operated remotely and allow full automatization of flying the drone, recharging it,

and downloading and transferring data to a server. These stocks are compatible with drones carrying a thermal camera (*e.g.* DJI Matrice 4TD, SZ DJI Technology Co. Ltd., China), which would allow for much higher flexibility in choosing the time of flying and possibly could increase temporal resolution of the measurements. These drones use real-time kinematic positioning (RTK) systems which allow for a precise flight, reducing also the need for ground control points for georeferencing. In addition, they carry an RGB camera and thermal and RGB information could be recorded at the same time, enabling an integrated analysis of CT together with structural traits such as fractional canopy cover or plant height. The use of such drone docks has become legal recently in Switzerland, as long as certain minimum requirements for the surroundings, such as distance from inhabited areas, are met.

In situations where absolute CT values are needed, a single stationary thermal infrared sensor pointing at a relatively small fraction of the canopy could allow for the referencing of relative CT measurements from the unreferenced multi-view approach. Together with an air temperature sensor, which is integrated in the DJI dock, this would also allow the calculation of indices related to crop water status, such as the crop water stress index (CWSI; Idso et al., 1981) or the standardized canopy temperature index (SCTI; Das, J. Christopher, Apan, Choudhury, et al., 2021). Such indices might be more closely related to corp water status than simple relative CT differences.

In conclusion, the automated multi-view thermography approach in conjunction with such drone docks offers a very interesting option for variety testing organizations for adopting a new trait.

6.1.2 Comprehensively understanding canopy temperature as a trait is crucial for its application in lean phenotyping

While canopy temperature is most commonly understood as a proxy measurement of stomatal conductance, it is a very complex secondary trait. There are many different sources of variance to be considered in analysis, especially when using a uncalibrated thermal camera.

A thorough understanding of the multiple sources of variance is important for a meaningful application of airborne thermography. These sources were described in the scientific literature, but it was difficult to gain an overview and a feeling for the importance of the different variance sources, as the different phenomena were described in separated, unrelated works and thus a tangible data example, allowing comparison of different effects, was lacking.

The aim of Chapter 3 was to connect the method of Chapter 2 with agronomic trials through a demonstration of its application in multiple trials with distinct treatments, and to relate CT with other primary and secondary wheat traits for a comprehensive study on determinants of CT in agricultural experiments. The multi-view approach presented in Chapter 2 made it possible to analyze and visualize confounding sources of variance, but Chapter 3 also integrated yield and multiple secondary traits into the analysis. As an applied review on sources of variance of CT, backed by rich experimental data, Chapter 3 enables potential users of airborne thermography to grasp many relevant aspects of the method in a condensed form. This will hopefully prevent practitioners from making conceptual errors in the execution and analysis of CT measurements.

It should be noted that while canopy temperature is often recommended in research on drought and heat tolerance, the phenotypic correlation between canopy temperature and yield within treatments was higher in the rather humid season 2021 than in the dry season 2022. This is in line with Bustos-Korts, Boer, Malosetti, et al., 2019, where in an early drought scenario, the phenotypic correlation of green canopy and biomass with yield was reduced compared to a scenario with only mild or no water stress. This highlights the complexity of G×E interactions and of the relation between primary and secondary traits. To be useful

in the identification of drought and heat tolerant genotypes, this complexity must be better understood for the context of temperate but warming climate of Switzerland and Central Europe. The trials in Chapter 3 only comprised a limited number of genotypes measured in only two seasons. With an increased number of seasons and genotypes covered, the relationship between CT, other secondary traits, and primary traits, especially in hot and dry conditions, might be further examined, *e.g.* with mixed models similar to Rebetzke et al., 2013.

That said, it remains the question of the value of drone-based thermography for wheat variety testing beyond the identification of drought- and heat-tolerant wheat genotypes. Within this thesis, the significant correlations between CT and yield were always stronger compared to correlations between multispectral indices and yield. Thus, CT is a good indicator of within-season plant performance, which is interesting per se, but also in the context of events that lead to partial or complete trial losses toward the end of the season (*e.g.* thunderstorms, hail or errors in the operation of machinery). With high-quality in-season measurements, such measurements could partially compensate for the loss of complete trials and enhance multi-environment trial (MET) statistics.

6.1.3 Lean phenotyping of phenology and senescence

In contrast to drone-based thermography, where the adoption of a relatively new trait was proposed for wheat variety testing, Chapter 4 was concerned with the development of a field-applicable lean phenotyping approach to measure classical variety testing traits in an automatic manner. A big advantage of such an approach would be the elimination of frequent, recurring field visits for the assessment of phenology and senescence. In contrast to previous studies, the PhenoCam approach presented here covered a whole variety testing trial at once, and not only single genotypes. The mast used allowed for a semimobile setup, which could be installed in a new location for each of three consecutive seasons and not a fixed setup as used *e.g.* in Aasen, Kirchgessner, et al., 2020. While still being at the prototype level, the full concept of PhenoCams was demonstrated under real variety testing conditions, which increased the TRL of the approach considerably.

The data quality of the JPEG (Joint Photographic Experts Group) image format was also shown to be sufficient to track phenology and senescence. This makes data transfer with mobile networks possible and the handling of SD (Secure Digital) cards obsolete, further improving the TRL. In addition, this allows for the continuous observation of the fields in almost real time. Trial problems can be detected in a timely manner and field visits planned according to the phenological progression of the trials, as observed by PhenoCams.

Most of the proposed improvements on the PhenoCam methodology, such as digital image stabilization and image quality assessment, are easily implemented by people with the appropriate knowledge. However, it might be more difficult to find a mast setup with a lower footprint that is still high enough, stable, and affordable. Here, a possible alternative would be the installation of a fixed phenotyping tower with a height of, for example 40 m. With such a mast in the center of a research field, plots could be observed at a less oblique angle in a circular area with a diameter of about 400 m (depending on plot size). Such a setup would lead to less flexibility and to higher initial costs, but it would also allow for more reliable recording of continuous measurements.

The DJI drone docking stations that were already mentioned for airborne thermography could also be an interesting option to derive RGB images of the experiment at a high temporal resolution. One option would be to create orthomosaics at a high temporal resolution, without the need for frequent field visits. But already having maybe three to four drone flights per day, weather permitting, taking only single overview images of individual experiments with a near-optimal viewing geometry could be an interesting alternative to PhenoCams.

6.1.4 Chlorophyll fluorescence as a tool for disease detection in lean phenotyping

The aim of translational research for lean phenotyping is to examine and develop phenotyping methodologies towards a higher TRL. However, sometimes its role is also to highlight shortcomings of proposed approaches, which have to be overcome before an approach can be further pursued.

Chlorophyll fluorescent (CF) can be used to well track *Fusarium* symptoms on single spikes. But to detect and quantify the disease at relevant infestation levels in the field, high precision but also high throughput is necessary. To our knowledge, no such method is currently available. In the near future, the use of neural networks to analyze RGB images seems most promising, as RGB images can be recorded inexpensively and processed with a reasonable effort. Including technologies like hyperspectral imaging or chlorophyll fluorescence imaging might help to increase the precision of detection, but this always comes at the cost of significantly higher initial investments and lower throughput.

6.2 Multidisciplinarity of lean phenotyping and limited resources

What all of the methods had in common was their multidisciplinary nature. Some manufacturing skills were required to set up the PhenoCams, although mostly consumer grade hardware was used. Software engineering skills were required for the image analysis pipelines. The operation of the drones and the processing of the photogrammetry data contained remote sensing user aspects. Geographical information systems (GIS) were used to organize and analyze the data. The analysis of the extensive measurement data required advanced statistical methods. Finally, these data had to be interpreted meaningfully and critically in the agronomic context of wheat variety testing.

Co-operation with many specialists and experts was therefore essential for the success of the individual methods. Many aspects of the different methods could have been refined or automated with more resources and time. Experts in the disciplines mentioned above would have had much to contribute to the quality and efficiency of the methods. However, this was not possible due to the limited time and resources available for this thesis. Yet, this reflects the reality of variety testing. Resources must be used in a targeted and needs-based manner. Variety testing requires methods that work and create added value. The high initial investment costs, even if they would pay off later, can be seen as a major obstacle to the implementation of lean phenotype in variety testing.

6.3 System integration and new technologies

This thesis addressed three lean typing methods in four separate chapters. An important next step would be the combination of these and other approaches in an integrated phenotyping setup. For example, PhenoCam could be used to determine optimal timing for thermal camera flights and will also enhance the interpretation of thermal data through a phenological characterization. The same holds for the screening of *Fusarium* or other diseases and other traits, where knowing the timing of the phenological development but also, for example, the exact date when a damage occurs in the field, would help to make more sense of the collected data.

One downside of digital phenotyping is the huge amount of data produced, as analyzing and organizing such large amounts of data is challenging (Coppens et al., 2017). Nevertheless, an integrated phenotyping setup offers the possibility of collecting good quality data at different

locations in a standardized way. Here, too, digitalization offers new possibilities for structuring and evaluating the data.

Analyzing multiple features simultaneously is challenging, but can reveal emergent information. For example, Roth, Binder, et al., 2024 showed how the dynamics of trait values are more valuable for breeding than the mere trait values at individual dates. The same applies to contextualization with weather data (F. Yang et al., 2023), which is now available for many regions of Central Europe with high temporal and spatial resolution. In particular, the differentiation between avoidance and tolerance to biotic and abiotic stresses could be improved in this way. Since this dissertation began in June 2020, great progress has been made in many technologies but especially deep learning. This has opened up new possibilities in the analysis of images, but also of data in general. Although new challenges arise in the computer-aided search for emergent dynamics (M. Yang et al., 2024), these new approaches offer new opportunities to analyze high-dimensional data from breeding and variety testing.

Based on such computer models, intelligent decision support systems could be developed to define suitable genotypes for appropriate environments (F. Yang et al., 2023). Such support systems might also be helpful, since the period between harvest and publication of lists of recommended varieties for the next season or for breeding, the period before selection decision, is short (Roth, 2021).

On the side of trait assessment, high spatial resolution satellite images offer new opportunities to observe field trials (Pinto et al., 2023; Hu et al., 2024), although they depend on cloud-free conditions during relevant crops stages, which can be a limitation in Central Europe. An alternative approach are autonomous field robots that are less dependent on cloud-free conditions (Xu and C. Li, 2022).

6.4 Use of new variety testing traits to inform breeding

While breeding and variety testing go hand in hand to enable farmers to grow high-performing and locally adapted varieties, it is very important that these two processes remain separate. Breeding is very often done by private companies, while variety testing is often public. If variety testing was carried out by breeding companies, they could promote their own varieties, but objective judgment is of great importance. Nevertheless, data generated in variety testing is very valuable for breeding. In METs, different sets of genotypes are sown over the years in different locations under different cropping systems. This generates many G×E×M combinations, which can be used for a better understanding of G×E×M interactions. If secondary traits such as CT, fractional canopy cover or an improved assessment of plant development were included in such variety testing data, it could further improve the understanding of yield formation for breeding purposes (Bustos-Korts, Boer, Malosetti, et al., 2019).

6.5 Data dissemination

The best data is of no use, if it is not shared with the relevant stakeholders. The Grains Research and Development Corporation (GRDC) of the Australian government is exemplary in this respect. It not only compiles lists of recommended varieties from the national variety trials (NVT), but also makes the results of the trials directly available online in a user-friendly format (`https://nvt.grdc.com.au/trial-results`). In addition, it operates an app with which the data can be searched according to criteria such as sowing time, trial region or variety (`https://app.nvt.grdc.com.au/lty/table`). With additional new traits, as described in the previous section, such data could be further enriched.

6.6 Transferability to other crops

Airborne multi-view thermography and PhenoCams were developed and tested on wheat in this thesis, but the methods should be transferrable to other crops. For other cereals, multi-view thermography would suffer from the same drawback of soil visible between rows and confounding CT estimates. For crops with almost full canopy closure, such as soybeans, the application of this method should even be more easy. When applying PhenoCams to other crops, the plant development traits would differ, but the hardware, and mostly even the analysis pipeline, would stay the same. Only the temporal features selected from the RGB space would change.

6.7 Conclusion

This work provides methodologies and insight for three optical lean phenotyping methods in the context of wheat variety testing. It would not be realistic to go from TRL one to nine in one step, but we pushed the TRL of both drone-based thermography and PhenoCams towards a higher level. We always did so with a strong focus on future applicability, without shying away from addressing or highlighting the challenges of workflows and methods, which is crucial for translational research. For drone-based thermography, a multi-view analysis workflow was proposed, with which the large temporal variability of thermal imaging can be estimated and corrected for, thereby improving the consistency and genotype specificity of CT estimates. This method was then applied to better understand the manifold sources of variance in CT estimates in wheat plot experiments. A strong link between CT and yield was found in conditions without water limitation and phenotypic correlations with other traits such as FCC and plant height were consistently significant and often strong. By applying multi-view thermography on more genotypes in more environments, a better understanding of the interaction of the different traits and their impact on the correlation between CT and yield could be obtained in the future. With PhenoCams, we improved a mobile setup and workflow to observe the temporal development of phenology and senescence in wheat, which are important for selecting varieties that are well adapted to local conditions. In addition, PhenoCams can be used to better contextualize other measurements, *e.g.* to determine whether a variety performs well due to stress avoidance or stress tolerance. Finally, chlorophyll fluorescence with a hand-held device but also with a chlorophyll camera was shown to be applicable to detect strong levels of *Fusarium* infestations on field experiments and to discriminate resistant genotypes from susceptible ones. However, the method lacks throughput, which must be increased to detect lower yet relevant levels of *Fusarium* infestations. We hope to have contributed our little part to the future of variety testing by increasing the quality and efficiency of phenotyping methods. This would finally lead to better information available to producers and enable them to close the on-farm yield gap by sowing varieties that are well aligned with local environments and production targets.

Bibliography

Aasen, Helge and Andreas Bolten (2018). "Multi-temporal high-resolution imaging spectroscopy with hyperspectral 2D imagers – From theory to application". In: *Remote Sensing of Environment* 205.2018, pp. 374–389. DOI: 10.1016/j.rse.2017.10.043.

Aasen, Helge, Eija Honkavaara, Arko Lucieer, and Pablo J. Zarco-Tejada (2018). "Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: A review of sensor technology, measurement procedures, and data correctionworkflows". In: *Remote Sensing* 10.7, pp. 1–42. DOI: 10.3390/rs10071091.

Aasen, Helge, Norbert Kirchgessner, Achim Walter, and Frank Liebisch (2020). "PhenoCams for Field Phenotyping: Using Very High Temporal Resolution Digital Repeated Photography to Investigate Interactions of Growth, Phenology, and Harvest Traits". In: *Frontiers in Plant Science* 11.June, pp. 1–16. DOI: 10.3389/fpls.2020.00593.

Adamsen, F. J., Paul J. Pinter, Edward M. Barnes, Robert L. LaMorte, Gerard W. Wall, Steven W. Leavitt, and Bruce A. Kimball (1999). "Measuring wheat senescence with a digital camera". In: *Crop Science* 39.3, pp. 719–724. DOI: 10.2135/cropsci1999.0011183X003900030019x.

Agroscope (2004). *Sommerweizen 2004 - Resultate der Sortenveruche*. URL: https://www.agroscope.admin.ch/agroscope/de/home/themen/pflanzenbau/ackerbau/kulturarten/strohgetreide/publikationen-sortenlisten-sortenpruefung/sortenpruefung-resultate-getreide.html (visited on 02/26/2025).

— (2017). *Winterweizen 2017 - Resultate der Sortenveruche*. URL: https://www.agroscope.admin.ch/agroscope/de/home/themen/pflanzenbau/ackerbau/kulturarten/strohgetreide/publikationen-sortenlisten-sortenpruefung/sortenpruefung-resultate-getreide.html (visited on 02/26/2025).

— (2020a). *Sommerweizen 2020 - Resultate der Sortenveruche*. URL: https://www.agroscope.admin.ch/agroscope/de/home/themen/pflanzenbau/ackerbau/kulturarten/strohgetreide/publikationen-sortenlisten-sortenpruefung/sortenpruefung-resultate-getreide.html (visited on 02/26/2025).

— (2020b). *Winterweizen 2020 - Resultate der Sortenveruche*. URL: https://www.agroscope.admin.ch/agroscope/de/home/themen/pflanzenbau/ackerbau/kulturarten/strohgetreide/publikationen-sortenlisten-sortenpruefung/sortenpruefung-resultate-getreide.html (visited on 02/26/2025).

Ahrends, H. E., W. Eugster, T. Gaiser, V. Rueda-Ayala, H. Hüging, F. Ewert, and S. Siebert (2018). "Genetic yield gains of winter wheat in Germany over more than 100 years (1895-2007) under contrasting fertilizer applications". In: *Environmental Research Letters* 13.10. DOI: 10.1088/1748-9326/aade12.

Ahrends, Hella Ellen, Sophie Etzold, Werner L Kutsch, Reto Stöckli, Robert Brügger, François Jeanneret, Heinz Wanner, Nina Buchmann, and Werner Eugster (2009). "Tree phenology and carbon dioxide fluxes: use of digital photography for process-based interpretation at the ecosystem scale". In: *Climate Research* 39.3, pp. 261–274.

Ajigboye, Olubukola O., Louise Bousquet, Erik H. Murchie, and Rumiana V. Ray (2016). "Chlorophyll fluorescence parameters allow the rapid detection and differentiation of plant

responses in three different wheat pathosystems". In: *Functional Plant Biology* 43.4, pp. 356–369. DOI: 10.1071/FP15280.

Al Masri, A., B. Hau, H. W. Dehne, A. K. Mahlein, and E. C. Oerke (2017). "Impact of primary infection site of Fusarium species on head blight development in wheat ears evaluated by IR-thermography". In: *European Journal of Plant Pathology* 147.4, pp. 855–868. DOI: 10.1007/s10658-016-1051-2.

Alisaac, E., J. Behmann, M. T. Kuska, H. W. Dehne, and A. K. Mahlein (2018). "Hyperspectral quantification of wheat resistance to Fusarium head blight: comparison of two Fusarium species". In: *European Journal of Plant Pathology* 152.4, pp. 869–884. DOI: 10.1007/s10658-018-1505-9.

Alisaac, Elias, Jan Behmann, Anna Rathgeb, Petr Karlovsky, Heinz-Wilhelm Dehne, and Anne-Katrin Mahlein (2019). "Assessment of Fusarium Infection and Mycotoxin Contamination of Wheat Kernels and Flour Using Hyperspectral Imaging". In: *Toxins* 11.10, p. 556. DOI: 10.3390/toxins11100556.

Alisaac, Elias and Anne-Katrin Mahlein (2023). "Fusarium Head Blight on Wheat: Biology, Modern Detection and Diagnosis and Integrated Disease Management". In: *Toxins* 15.3, p. 192. DOI: 10.3390/toxins15030192.

Almawazreh, Albara, Andreas Buerkert, Prem Jose Vazhacharickal, and Stephan Peth (2025). "Assessing canopy temperature responses to nitrogen fertilization in South Indian crops using UAV-based thermal sensing". In: *International Journal of Remote Sensing* 46.6, pp. 2389–2417. DOI: 10.1080/01431161.2025.2452312.

Almoujahed, Muhammad Baraa, Aravind Krishnaswamy Rangarajan, Rebecca L. Whetton, Damien Vincke, Damien Eylenbosch, Philippe Vermeulen, and Abdul M. Mouazen (2022). "Detection of fusarium head blight in wheat under field conditions using a hyperspectral camera and machine learning". In: *Computers and Electronics in Agriculture* 203, p. 107456. DOI: 10.1016/j.compag.2022.107456.

— (2024). "Non-destructive detection of fusarium head blight in wheat kernels and flour using visible near-infrared and mid-infrared spectroscopy". In: *Chemometrics and Intelligent Laboratory Systems* 245, p. 105050. DOI: 10.1016/j.chemolab.2023.105050.

Anderegg, Jonas, Helge Aasen, Gregor Perich, Lukas Roth, Achim Walter, and Andreas Hund (2021). "Temporal trends in canopy temperature and greenness are potential indicators of late-season drought avoidance and functional stay-green in wheat". In: *Field Crops Research* 274, p. 108311. DOI: 10.1016/j.fcr.2021.108311.

Anderegg, Jonas, Norbert Kirchgessner, Helge Aasen, Olivia Zumsteg, Beat Keller, Radek Zenkl, Achim Walter, and Andreas Hund (2024). "Thermal imaging can reveal variation in stay-green functionality of wheat canopies under temperate conditions". In: *Frontiers in Plant Science* 15, p. 1335037. DOI: 10.3389/fpls.2024.1335037.

Anderegg, Jonas, Flavian Tschurr, Norbert Kirchgessner, Simon Treier, Manuel Schmucki, Bernhard Streit, and Achim Walter (2023). "On-farm evaluation of UAV-based aerial imagery for season-long weed monitoring under contrasting management and pedoclimatic conditions in wheat". In: *Computers and Electronics in Agriculture* 204, p. 107558. DOI: 10.1016/j.compag.2022.107558.

Anderegg, Jonas, Kang Yu, Helge Aasen, Achim Walter, Frank Liebisch, and Andreas Hund (2020). "Spectral vegetation indices to track senescence dynamics in diverse wheat germplasm". In: *Frontiers in Plant Science* 10.January, pp. 1–20. DOI: 10.3389/fpls.2019.01749.

Aragon, Bruno, Kasper Johansen, Stephen Parkes, Yoann Malbeteau, Samir Al-Mashharawi, Talal Al-Amoudi, Cristhian F. Andrade, Darren Turner, Arko Lucieer, and Matthew F. McCabe (2020). "A calibration procedure for field and UAV-based uncooled thermal infrared instruments". In: *Sensors* 20.11, p. 3316. DOI: 10.3390/s20113316.

Araus, José Luis and Jill E. Cairns (2014). "Field high-throughput phenotyping: the new crop breeding frontier". In: *Trends in Plant Science* 19.1, pp. 52–61. DOI: 10.1016/j.tplants.2013.09.008.

Araus, José Luis, Shawn C. Kefauver, Mainassara Zaman-Allah, Mike S. Olsen, and Jill E. Cairns (2018). "Translating High-Throughput Phenotyping into Genetic Gain". In: *Trends in Plant Science* 23.5, pp. 451–466. DOI: 10.1016/j.tplants.2018.02.001.

Araus, José Luis, Gustavo A. Slafer, Conxita Royo, and M. Dolores Serret (2008). "Breeding for Yield Potential and Stress Adaptation in Cereals". In: *Critical Reviews in Plant Sciences* 27.6, pp. 377–412. DOI: 10.1080/07352680802467736.

Asseng, S., F. Ewert, P. Martre, et al. (2015). "Rising temperatures reduce global wheat production". In: *Nature Climate Change* 5.2, pp. 143–147. DOI: 10.1038/nclimate2470.

Asseng, S., F. Ewert, C. Rosenzweig, et al. (2013). "Uncertainty in simulating wheat yields under climate change". In: *Nature Climate Change* 3.9, pp. 827–832. DOI: 10.1038/nclimate1916.

Bai, Geng, Yufeng Ge, Bryan Leavitt, John A. Gamon, and David Scoby (2023). "Goniometer in the air: Enabling BRDF measurement of crop canopies using a cable-suspended plant phenotyping platform". In: *Biosystems Engineering* 230, pp. 344–360. DOI: 10.1016/j.biosystemseng.2023.04.017.

Baluja, Javier, Maria P. Diago, Pedro Balda, Roberto Zorer, Franco Meggio, Fermin Morales, and Javier Tardaguila (2012). "Assessment of vineyard water status variability by thermal and multispectral imagery using an unmanned aerial vehicle (UAV)". In: *Irrigation Science* 30.6, pp. 511–522. DOI: 10.1007/s00271-012-0382-9.

Bannihatti, Rudrappa K., Parimal Sinha, Dhandapani Raju, Shubhajyoti Das, S. N. Mandal, R. S. Raje, C. Viswanathan, Sudhir Kumar, K. Gaikwad, and R. Aggarwal (2022). "Image Based High throughput Phenotyping for Fusarium Wilt Resistance in Pigeon Pea (Cajanus cajan)". In: *Phytoparasitica* 50.5, pp. 1075–1090. DOI: 10.1007/s12600-022-00993-5.

Barnes, E. M., T. R. Clarke, S. E. Richards, P. D. Colaizzi, J. Haberland, M. Kostrzewski, P. Waller, C. Choi, E. Riley, and T. Thompson (2000). "Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data". In: *Proceedings of the Fifth International Conference on Precision Agriculture.* Vol. 1619, p. 1356.

Barreto, Cynthia Aparecida Valiati, Kaio Olimpio Das Graças Dias, Ithalo Coelho De Sousa, Camila Ferreira Azevedo, Ana Carolina Campana Nascimento, Lauro José Moreira Guimarães, Claudia Teixeira Guimarães, Maria Marta Pastina, and Moysés Nascimento (2024). "Genomic prediction in multi-environment trials in maize using statistical and machine learning methods". In: *Scientific Reports* 14.1, p. 1062. DOI: 10.1038/s41598-024-51792-3.

Bauriegel, E., A. Giebel, M. Geyer, U. Schmidt, and W. B. Herppich (2010). "Early detection of Fusarium infection in wheat using hyper-spectral imaging". In: *Computers and Electronics in Agriculture* 75.2. Publisher: Elsevier B.V., pp. 304–312. DOI: 10.1016/j.compag.2010.12.006.

Bauriegel, Elke, Antje Giebel, and Werner B. Herppich (2011). "Hyperspectral and chlorophyll fluorescence imaging to analyse the impact of fusarium culmorum on the photosynthetic integrity of infected wheat ears". In: *Sensors* 11.4, pp. 3765–3779. DOI: 10.3390/s110403765.

Bauriegel, Elke and Werner Herppich (2014). "Hyperspectral and Chlorophyll Fluorescence Imaging for Early Detection of Plant Diseases, with Special Reference to Fusarium spec. Infections on Wheat". In: *Agriculture* 4.1, pp. 32–57. DOI: 10.3390/agriculture4010032.

Baxter, Sam (2007). "World reference base for soil resources. World soil resources report 103. Rome: Food and Agriculture Organization of the United Nations (2006), pp. 132". In: *Experimental Agriculture* 43.2, pp. 264–264. DOI: 10.1017/S0014479706394902.

Becker, Heiko (2011). *Pflanzenzüchtung*. Vol. 1744. Series Title: UTB Agrarwissenschaften. Stuttgart: Ulmer. ISBN: 978-3-8252-3558-1.

Benassi, Francesco, Elisa Dall'Asta, Fabrizio Diotri, Gianfranco Forlani, Umberto Morra di Cella, Riccardo Roncella, and Marina Santise (2017). "Testing accuracy and repeatability of UAV blocks oriented with GNSS-supported aerial triangulation". In: *Remote Sensing* 9.2, p. 172. DOI: 10.3390/rs9020172.

Bendig, Juliane, Kang Yu, Helge Aasen, Andreas Bolten, Simon Bennertz, Janis Broscheit, Martin L. Gnyp, and Georg Bareth (2015). "Combining UAV-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley". In: *International Journal of Applied Earth Observation and Geoinformation* 39, pp. 79–87. DOI: 10.1016/j.jag.2015.02.012.

Berg, Stuart, Dominik Kutra, Thorben Kroeger, Christoph N. Straehle, Bernhard X. Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, Kemal Eren, Jaime I. Cervantes, Buote Xu, Fynn Beuttenmueller, Adrian Wolny, Chong Zhang, Ullrich Koethe, Fred A. Hamprecht, and Anna Kreshuk (2019). "ilastik: interactive machine learning for (bio)image analysis". In: *Nature Methods*. DOI: 10.1038/s41592-019-0582-9.

Berners-Lee, M., C. Kennelly, R. Watson, and C. N. Hewitt (2018). "Current global food production is sufficient to meet human nutritional needs in 2050 provided there is radical societal adaptation". In: *Elementa: Science of the Anthropocene* 6. Ed. by Anne R. Kapuscinski, Kim A. Locke, and Christian J. Peters, p. 52. DOI: 10.1525/elementa.310.

Bhatti, Muhammad Tousif, Hammad Gilani, Muhammad Ashraf, Muhammad Shahid Iqbal, and Sarfraz Munir (2024). "Field validation of NDVI to identify crop phenological signatures". In: *Precision Agriculture*. DOI: 10.1007/s11119-024-10165-6.

Birth, Gerald S. and George R. McVey (1968). "Measuring the Color of Growing Turf with a Reflectance Spectrophotometer". In: *Agronomy Journal* 60.6, pp. 640–643. DOI: 10.2134/agronj1968.00021962006000060016x.

Blackshaw, R. E., L. J. Molnar, and J. R. Moyer (2010). "Suitability of legume cover crop-winter wheat intercrops on the semi-arid Canadian Prairies". In: *Canadian Journal of Plant Science* 90.4, pp. 479–488. DOI: 10.4141/CJPS10006.

Blum, A., J. Mayer, and G. Gozlan (1982). "Infrared thermal sensing of plant canopies as a screening technique for dehydration avoidance in wheat". In: *Field Crops Research* 5, pp. 137–146. DOI: 10.1016/0378-4290(82)90014-4.

Boesch, R. (2017). "Thermal remote sensing with UAV-based workflows". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W6, pp. 41–46. DOI: 10.5194/isprs-archives-XLII-2-W6-41-2017.

Bradski, Gary and Adrian Kaehler (2000). "OpenCV". In: *Dr. Dobb's journal of software tools* 3.2.

Brennan, J. P., A. G. Condon, M. Van Ginkel, and Matthew P. Reynolds (2007). "Paper presented at international workshop on increasing wheat yield potential, CIMMYT, Obregon, Mexico, 20–24 March 2006 An economic assessment of the use of physiological selection for stomatal aperture-related traits in the CIMMYT wheat breeding programme". In: *The Journal of Agricultural Science* 145.3, pp. 187–194. DOI: 10.1017/S0021859607007009.

Brisson, Nadine, Philippe Gate, David Gouache, Gilles Charmet, François-Xavier Oury, and Frédéric Huard (2010). "Why are wheat yields stagnating in Europe? A comprehensive data analysis for France". In: *Field Crops Research* 119.1, pp. 201–212. DOI: 10.1016/j.fcr.2010.07.012.

Brocks, Sebastian and Georg Bareth (2018). "Estimating Barley Biomass with Crop Surface Models from Oblique RGB Imagery". In: *Remote Sensing* 10.2, p. 268. DOI: 10.3390/rs10020268.

Brocks, Sebastian, Juliane Bendig, and Georg Bareth (2016). "Toward an automated low-cost three-dimensional crop surface monitoring system using oblique stereo imagery from consumer-grade smart cameras". In: *Journal of Applied Remote Sensing* 10.4, p. 046021. DOI: 10.1117/1.JRS.10.046021.

Browning, Dawn M., Jason W. Karl, David Morin, Andrew D. Richardson, and Craig E. Tweedie (2017). "Phenocams bridge the gap between field and satellite observations in an arid grassland ecosystem". In: *Remote Sensing* 9.10. DOI: 10.3390/rs9101071.

Bruschi, Martina, Matteo Bozzoli, Claudio Ratti, Giuseppe Sciara, Ellen Goudemand, Pierre Devaux, Danara Ormanbekova, Cristian Forestan, Simona Corneti, Sandra Stefanelli, Sara Castelletti, Elena Fusari, Jad B Novi, Elisabetta Frascaroli, Silvio Salvi, Dragan Perovic, Agata Gadaleta, Concepcion Rubies-Autonell, Maria Corinna Sanguineti, Roberto Tuberosa, and Marco Maccaferri (2024). "Dissecting the genetic basis of resistance to Soil-borne cereal mosaic virus (SBCMV) in durum wheat by bi-parental mapping and GWAS". In: *Theoretical and Applied Genetics* 137.9, p. 213. DOI: 10.1007/s00122-024-04709-7.

Budzier, H. and G. Gerlach (2015). "Calibration of uncooled thermal infrared cameras". In: *Journal of Sensors and Sensor Systems* 4.1, pp. 187–197. DOI: 10.5194/jsss-4-187-2015.

Buerstmayr, H., T. Ban, and J. A. Anderson (2009). "QTL mapping and marker-assisted selection for *Fusarium* head blight resistance in wheat: a review". In: *Plant Breeding* 128.1, pp. 1–26. DOI: 10.1111/j.1439-0523.2008.01550.x.

Buerstmayr, Maria, Barbara Steiner, and Hermann Buerstmayr (2020). "Breeding for Fusarium head blight resistance in wheat—Progress and challenges". In: *Plant Breeding* 139.3. Ed. by Jens Léon, pp. 429–454. DOI: 10.1111/pbr.12797.

Burkart, A., V. L. Hecht, T. Kraska, and U. Rascher (2018). "Phenological analysis of unmanned aerial vehicle based time series of barley imagery with high temporal resolution". In: *Precision Agriculture* 19.1. Publisher: Springer US, pp. 134–146. DOI: 10.1007/s11119-017-9504-y.

Bustos-Korts, Daniela, Martin P. Boer, Jamie Layton, Anke Gehringer, Tom Tang, Ron Wehrens, Charlie Messina, Abelardo J. De La Vega, and Fred A. Van Eeuwijk (2022). "Identification of environment types and adaptation zones with self-organizing maps; applications to sunflower multi-environment data in Europe". In: *Theoretical and Applied Genetics* 135.6, pp. 2059–2082. DOI: 10.1007/s00122-022-04098-9.

Bustos-Korts, Daniela, Martin P. Boer, Marcos Malosetti, Scott C. Chapman, Karine Chenu, Bangyou Zheng, and Fred A. Van Eeuwijk (2019). "Combining Crop Growth Modeling and Statistical Genetic Modeling to Evaluate Phenotyping Strategies". In: *Frontiers in Plant Science* 10, p. 1491. DOI: 10.3389/fpls.2019.01491.

Butler, David (2019). "asreml: Fits the Linear Mixed Model. R package version 4.1. 0.110". In: *VSN International Ltd.: Hemel Hempstead, UK*.

Campbell, J.B. and R.H. Wynne (2011). *Introduction to Remote Sensing, Fifth Edition*. Guilford Publications. ISBN: 978-1-60918-177-2.

Cao, Xiaofeng, Yulin Liu, Rui Yu, Dejun Han, and Baofeng Su (2021). "A Comparison of UAV RGB and Multispectral Imaging in Phenotyping for Stay Green of Wheat Population". In: *Remote Sensing* 13.24, p. 5173. DOI: 10.3390/rs13245173.

Cárcer, Paula Sanginés de, Sokrat Sinaj, Mathieu Santonja, Dario Fossati, and Bernard Jeangros (2019). "Long-term effects of crop succession, soil tillage and climate on wheat yield and soil properties". In: *Soil and Tillage Research* 190, pp. 209–219. DOI: 10.1016/j.still.2019.01.012.

Carrascal, Luis M., Ismael Galván, and Oscar Gordo (2009). "Partial least squares regression as an alternative to current regression methods used in ecology". In: *Oikos* 118.5, pp. 681–690. DOI: 10.1111/j.1600-0706.2008.16881.x.

Carvalho, Humberto Fanelli, Simon Rio, Julian García-Abadillo, and Julio Isidro Y Sánchez (2024). "Revisiting superiority and stability metrics of cultivar performances using genomic data: derivations of new estimators". In: *Plant Methods* 20.1, p. 85. DOI: 10.1186/s13007-024-01207-1.

Chandel, Narendra S., Yogesh A. Rajwade, Kumkum Dubey, Abhilash K. Chandel, A. Subeesh, and Mukesh K. Tiwari (2022). "Water Stress Identification of Winter Wheat Crop with State-of-the-Art AI Techniques and High-Resolution Thermal-RGB Imagery". In: *Plants* 11.23, p. 3344. DOI: 10.3390/plants11233344.

Chapman, Elizabeth A., Simon Orford, Jacob Lage, and Simon Griffiths (2021). "Capturing and Selecting Senescence Variation in Wheat". In: *Frontiers in Plant Science* 12, p. 638738. DOI: 10.3389/fpls.2021.638738.

Cheng, Jie and Shengyue Dong (2024). "A New Canopy Emissivity Model for Sparsely Vegetated Surfaces Incorporating Soil Directional Emissivity and Topography". In: *IEEE Transactions on Geoscience and Remote Sensing* 62, pp. 1–11. DOI: 10.1109/TGRS.2024.3401840.

Christopher, John T., Mandy J. Christopher, Andrew K. Borrell, Susan Fletcher, and Karine Chenu (2016). "Stay-green traits to improve wheat adaptation in well-watered and water-limited environments". In: *Journal of Experimental Botany* 67.17, pp. 5159–5172. DOI: 10.1093/jxb/erw276.

Christopher, John T., Mathieu Veyradier, Andrew K. Borrell, Greg Harvey, Susan Fletcher, and Karine Chenu (2014). "Phenotyping novel stay-green traits to capture genetic variation in senescence dynamics". In: *Functional Plant Biology* 41.11, pp. 1035–1048. DOI: 10.1071/FP14052.

CIMMYT (2019). *Bottlenecks between basic and applied plant science jeopardize life-saving crop improvements*. en-US. URL: https://www.cimmyt.org/news/bottlenecks-between-basic-and-applied-plant-science-jeopardize-life-saving-crop-improvements/ (visited on 01/20/2025).

Cobb, Joshua N., Genevieve DeClerck, Anthony Greenberg, Randy Clark, and Susan McCouch (2013). "Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement". In: *Theoretical and Applied Genetics* 126.4, pp. 867–887. DOI: 10.1007/s00122-013-2066-0.

Cooke, Robert J. and James C. Reeves (2003). "Plant genetic resources and molecular markers: variety registration in a new era". In: *Plant Genetic Resources* 1.2-3, pp. 81–87. DOI: 10.1079/PGR200312.

Cooper, Mark, Kai P. Voss-Fels, Carlos D. Messina, Tom Tang, and Graeme L. Hammer (2021). "Tackling G × E × M interactions to close on-farm yield-gaps: creating novel pathways for crop improvement by predicting contributions of genetics and management to crop productivity". In: *Theoretical and Applied Genetics* 134.6, pp. 1625–1644. DOI: 10.1007/s00122-021-03812-3.

Coppens, Frederik, Nathalie Wuyts, Dirk Inzé, and Stijn Dhondt (2017). "Unlocking the potential of plant phenotyping data through integration and data-driven approaches". In: *Current Opinion in Systems Biology* 4, pp. 58–63. DOI: 10.1016/j.coisb.2017.07.002.

Costa-Neto, Germano, Leonardo Crespo-Herrera, Nick Fradgley, Keith Gardner, Alison R Bentley, Susanne Dreisigacker, Roberto Fritsche-Neto, Osval A Montesinos-López, and Jose Crossa (2023). "Envirome-wide associations enhance multi-year genome-based prediction of historical wheat breeding data". In: *G3* 13.2, jkac313. DOI: 10.1093/g3journal/jkac313.

Crain, Jared, Suchismita Mondal, Jessica Rutkoski, Ravi P. Singh, and Jesse Poland (2018). "Combining High-Throughput Phenotyping and Genomic Information to Increase Prediction and Selection Accuracy in Wheat Breeding". In: *The Plant Genome* 11.1, p. 170043. DOI: 10.3835/plantgenome2017.05.0043.

Crespo-Herrera, Leonardo A., Jose Crossa, Julio Huerta-Espino, Enrique Autrique, Suchismita Mondal, Govindan Velu, Mateo Vargas, Hans J. Braun, and Ravi P. Singh (2017). "Genetic Yield Gains In CIMMYT's International Elite Spring Wheat Yield Trials By Modeling The Genotype × Environment Interaction". In: *Crop Science* 57.2, pp. 789–801. DOI: 10.2135/cropsci2016.06.0553.

Cullis, B. R., A. Smith, C. Hunt, and A. Gilmour (2000). "An examination of the efficiency of Australian crop variety evaluation programmes". In: *Journal of Agricultural Science* 135.3, pp. 213–222. DOI: 10.1017/S0021859699008163.

Cullis, B. R., A. B. Smith, and N. E. Coombes (2006). "On the design of early generation variety trials with correlated data". In: *Journal of Agricultural, Biological, and Environmental Statistics* 11.4, pp. 381–393. DOI: 10.1198/108571106X154443.

Damm, A., S. Cogliati, R. Colombo, L. Fritsche, A. Genangeli, L. Genesio, J. Hanus, A. Peressotti, P. Rademske, U. Rascher, D. Schuettemeyer, B. Siegmann, J. Sturm, and F. Miglietta (2022). "Response times of remote sensing measured sun-induced chlorophyll fluorescence, surface temperature and vegetation indices to evolving soil water limitation in a crop canopy". In: *Remote Sensing of Environment* 273, p. 112957. DOI: 10.1016/j.rse.2022.112957.

Darst, Burcu F., Kristen C. Malecki, and Corinne D. Engelman (2018). "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data". In: *BMC Genetics* 19.S1, p. 65. DOI: 10.1186/s12863-018-0633-8.

Das, Sumanta, Scott C. Chapman, Jack Christopher, Malini Roy Choudhury, Neal W. Menzies, Armando Apan, and Yash P. Dang (2021). "UAV-thermal imaging: A technological breakthrough for monitoring and quantifying crop abiotic stress to help sustain productivity on sodic soils – A case review on wheat". In: *Remote Sensing Applications: Society and Environment* 23, p. 100583. DOI: 10.1016/j.rsase.2021.100583.

Das, Sumanta, Jack Christopher, Armando Apan, Malini Roy Choudhury, Scott C. Chapman, Neal W. Menzies, and Yash P. Dang (2021). "Evaluation of water status of wheat genotypes to aid prediction of yield on sodic soils using UAV-thermal imaging and machine learning". In: *Agricultural and Forest Meteorology* 307, p. 108477. DOI: 10.1016/j.agrformet.2021.108477.

Das, Sumanta, Jack Christopher, Armando Apan, Malini Roy Choudhury, Scott C. Chapman, Neal W. Menzies, and Yash P. Dang (2021). "UAV-thermal imaging and agglomerative hierarchical clustering techniques to evaluate and rank physiological performance of wheat genotypes on sodic soil". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 173, pp. 221–237. DOI: 10.1016/j.isprsjprs.2021.01.014.

Deery, David M., Greg J. Rebetzke, Jose A. Jimenez-Berni, William D. Bovill, Richard A. James, Anthony G. Condon, Robert T. Furbank, Scott C. Chapman, and Ralph A. Fischer (2019). "Evaluation of the phenotypic repeatability of canopy temperature in wheat using continuous-terrestrial and airborne measurements". In: *Frontiers in Plant Science* 10, p. 875. DOI: 10.3389/fpls.2019.00875.

Deery, David M., Greg J. Rebetzke, Jose A. Jimenez-Berni, Richard A. James, Anthony G. Condon, William D. Bovill, Paul Hutchinson, Jamie Scarrow, Robert Davy, and Robert T. Furbank (2016). "Methodology for high-throughput field phenotyping of canopy temperature using airborne thermography". In: *Frontiers in Plant Science* 7. DOI: 10.3389/fpls.2016.01808.

Diaz, Lucas Ribeiro, Daniel Caetano Santos, Pâmela Suélen Käfer, Nájila Souza Da Rocha, Savannah Tâmara Lemos Da Costa, Eduardo Andre Kaiser, and Silvia Beatriz Alves Rolim (2021). "Atmospheric Correction of Thermal Infrared Landsat Images Using High-Resolution Vertical Profiles Simulated by WRF Model". In: *The 4th International Electronic Conference on Atmospheric Sciences*. MDPI, p. 27. DOI: 10.3390/ecas2021-10351.

Dong, Baodi, Xin Zheng, Haipei Liu, Jason A. Able, Hong Yang, Huan Zhao, Mingming Zhang, Yunzhou Qiao, Yakai Wang, and Mengyu Liu (2017). "Effects of Drought Stress on Pollen Sterility, Grain Yield, Abscisic Acid and Protective Enzymes in Two Winter Wheat Cultivars". In: *Frontiers in Plant Science* 8, p. 1008. DOI: 10.3389/fpls.2017.01008.

Eichi, Vahid Rahimi, Mamoru Okamoto, Trevor Garnett, Paul Eckermann, Benoit Darrier, Matteo Riboni, and Peter Langridge (2020). "Strengths and weaknesses of national variety trial data for multi-environment analysis: A case study on grain yield and protein content". In: *Agronomy* 10.5. DOI: 10.3390/agronomy10050753.

Elmerich, Chloé, Michel-Pierre Faucon, Milagros Garcia, Patrice Jeanson, Guénolé Boulch, and Bastien Lange (2023). "Envirotyping to control genotype x environment interactions for efficient soybean breeding". In: *Field Crops Research* 303, p. 109113. DOI: 10.1016/j.fcr.2023.109113.

European Commission (2014). *Technology readiness levels (TRL). Horizon 2020 Work Programme 2015. Commission Decision C (2014) 4995*. Publisher: European Commission.

Fahlgren, Noah, Malia A. Gehan, and Ivan Baxter (2015). "Lights, camera, action: High-throughput plant phenotyping is ready for a close-up". In: *Current Opinion in Plant Biology* 24. Publisher: Elsevier Ltd, pp. 93–99. DOI: 10.1016/j.pbi.2015.02.006.

Fang, Zhou, Dewayne D. Deng, Johnie N. Jenkins, and Qian M. Zhou (2024). "An investigation of the impact of imbalance on the analysis of the US crop variety evaluation program data". In: *Crop Science* 64.4, pp. 2183–2194. DOI: 10.1002/csc2.21262.

FAO (2017). *The future of food and agriculture: trends and challenges*. ISBN: 978-92-5-109551-5.

FAOSTAT (2025a). *FAOSTAT Data - Crops and livestock products*. URL: https://www.fao.org/faostat/en/#data/QCL (visited on 01/21/2025).

— (2025b). *FAOSTAT Data - Food Balances*. URL: https://www.fao.org/faostat/en/#data/FBS (visited on 01/21/2025).

— (2025c). *FAOSTAT Data - Land Use Data*. URL: https://www.fao.org/faostat/en/#data/RL (visited on 01/20/2025).

Farooq, Muhammad, Mubshar Hussain, and Kadambot H. M. Siddique (2014). "Drought Stress in Wheat during Flowering and Grain-filling Periods". In: *Critical Reviews in Plant Sciences* 33.4, pp. 331–349. DOI: 10.1080/07352689.2014.875291.

Ferrigo, Davide, Alessandro Raiola, and Roberto Causin (2016). "Fusarium toxins in cereals: Occurrence, legislation, factors promoting the appearance and their management". In: *Molecules* 21.5. DOI: 10.3390/molecules21050627.

FiBL (2022). *Sortenliste - Getreide - Für den Bioanbau empfohlene Sorten, Ernte 2023*.

Fischer, Andreas and Urs Feller (1994). "Senescence and protein degradation in leaf segments of young winter wheat: influence of leaf age". In: *Journal of Experimental Botany* 45.1, pp. 103–109. DOI: 10.1093/jxb/45.1.103.

Fischer, R. A., D. Rees, K. D. Sayre, Z.-M. Lu, A. G. Condon, and A. Larque Saavedra (1998). "Wheat Yield Progress Associated with Higher Stomatal Conductance and Photosynthetic Rate, and Cooler Canopies". In: *Crop Science* 38.6, pp. 1467–1475. DOI: 10.2135/cropsci1998.0011183X003800060011x.

Fischer, Tony, Derek Byerlee, and Greg Edmeades (2014). *Crop yields and global food security: will yield increase continue to feed the world?* ACIAR monograph series 158. Canberra: ACIAR. ISBN: 978-1-925133-07-3.

Ford, Margaret A. and Gillian N. Thorne (1975). "Effects of variation in temperature and light intensity at different times on growth and yield of spring wheat". In: *Annals of Applied Biology* 80.3, pp. 283–299. DOI: 10.1111/j.1744-7348.1975.tb01634.x.

Francesconi, Sara, Antoine Harfouche, Mauro Maesano, and Giorgio Mariano Balestra (2021). "UAV-based thermal, RGB imaging and gene expression analysis allowed detection of Fusarium Head Blight and gave new insights into the physiological responses to the disease

in durum wheat". In: *Frontiers in Plant Science* 12.April, pp. 1–19. DOI: 10.3389/fpls.2021.628575.

Fuchs, M. and C. B. Tanner (1966). "Infrared Thermometry of Vegetation". In: *Agronomy Journal* 58.6, pp. 597–601. DOI: 10.2134/agronj1966.00021962005800060014x.

Gallet, Anne, René Flisch, Jean-Pierre Ryser, Emmanuel Frossard, and Sokrat Sinaj (2003). "Effect of phosphate fertilization on crop yield and soil phosphorus status". In: *Journal of Plant Nutrition and Soil Science* 166.5, pp. 568–578. DOI: 10.1002/jpln.200321081.

Gao, X., C. H. Hu, H. Z. Li, Y. J. Yao, M Meng, J. Dong, W C Zhao, Q. J. Chen, and X Y Li (2013). "Factors Affecting Pre-Harvest Sprouting Resistance in Wheat (Triticum Aestivum L.)" In: *J. Anim. Plant Sci.*

GDAL/OGR Contributors (2024). *Geospatial Data Abstraction software Library*. Publisher: Open Source Geospatial Foundation, Chicago, IL.

Gerard, Guillermo, Suchismita Mondal, Francisco Piñera-Chávez, Carolina Rivera-Amado, Gemma Molero, Jose Crossa, Julio Huerta-Espino, Govindan Velu, Hans Braun, Ravi Singh, and Leonardo Crespo-Herrera (2024). "Enhanced radiation use efficiency and grain filling rate as the main drivers of grain yield genetic gains in the CIMMYT elite spring wheat yield trial". In: *Scientific Reports* 14.1, p. 10975. DOI: 10.1038/s41598-024-60853-6.

Gerber, James S., Deepak K. Ray, David Makowski, Ethan E. Butler, Nathaniel D. Mueller, Paul C. West, Justin A. Johnson, Stephen Polasky, Leah H. Samberg, Stefan Siebert, and Lindsey Sloat (2024). "Global spatially explicit yield gap time trends reveal regions at risk of future crop yield stagnation". In: *Nature Food* 5.2, pp. 125–135. DOI: 10.1038/s43016-023-00913-8.

Gillespie, Alan R., Anne B. Kahle, and Richard E. Walker (1987). "Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques". In: *Remote Sensing of Environment* 22.3, pp. 343–365. DOI: 10.1016/0034-4257(87)90088-5.

Gilmour, Arthur R., Brian R. Cullis, and Arūnas P. Verbyla (1997). "Accounting for natural and extraneous variation in the analysis of field experiments". In: *Journal of Agricultural, Biological, and Environmental Statistics*, pp. 269–293. DOI: 10.2307/1400446.

Gitelson, Anatoly A., Yoram J. Kaufman, Robert Stark, and Don Rundquist (2002). "Novel algorithms for remote estimation of vegetation fraction". In: *Remote Sensing of Environment* 80.1, pp. 76–87. DOI: 10.1016/S0034-4257(01)00289-9.

Gitelson, Anatoly A. and Mark N. Merzlyak (1994). "Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves". In: *Journal of Photochemistry and Photobiology B: Biology* 22.3, pp. 247–252.

Gitelson, Anatoly A., Mark N. Merzlyak, and Olga B. Chivkunova (2001). "Optical properties and nondestructive estimation of anthocyanin content in plant leaves¶". In: *Photochemistry and photobiology* 74.1, pp. 38–45.

Godfray, H. Charles J. and Tara Garnett (2014). "Food security and sustainable intensification". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1639, p. 20120273. DOI: 10.1098/rstb.2012.0273.

Google (2024). *Preise | Cloud Storage*. de-x-mtfrom-en. URL: https://cloud.google.com/storage/pricing?hl=de (visited on 12/04/2024).

Gracia-Romero, Adrian, Shawn C. Kefauver, Omar Vergara-Díaz, Mainassara A. Zaman-Allah, Boddupalli M. Prasanna, Jill E. Cairns, and José L. Araus (2017). "Comparative performance of ground vs. Aerially assessed rgb and multispectral indices for early-growth evaluation of maize performance under phosphorus fertilization". In: *Frontiers in Plant Science* 8.November, pp. 1–13. DOI: 10.3389/fpls.2017.02004.

Graham, Eric A., Erin C. Riordan, Eric M. Yuen, Deborah Estrin, and Philip W. Rundel (2010). "Public Internet-connected cameras used as a cross-continental ground-based plant

phenology monitoring system". In: *Global Change Biology* 16.11, pp. 3014–3023. DOI: 10.1111/j.1365-2486.2010.02164.x.

GRDC (2025). *INVITA – Innovations in plant testing in Australia*. en-au. URL: https://grdc.com.au/grdc-investments/investments/investment (visited on 03/02/2025).

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre (2016). "Correlation and variable importance in random forests". In: *Statistics and Computing* 27.3, pp. 659–678. DOI: 10.1007/s11222-016-9646-1.

Grishina, Alyona, Oksana Sherstneva, Sergey Mysyagin, Anna Brilkina, and Vladimir Vodeneev (2024). "Detecting Plant Infections: Prospects for Chlorophyll Fluorescence Imaging". In: *Agronomy* 14.11, p. 2600. DOI: 10.3390/agronomy14112600.

Guo, Yahui, Shouzhi Chen, Yongshuo H. Fu, Yi Xiao, Wenxiang Wu, Hanxi Wang, and Kirsten De Beurs (2022). "Comparison of Multi-Methods for Identifying Maize Phenology Using PhenoCams". In: *Remote Sensing* 14.2, p. 244. DOI: 10.3390/rs14020244.

Harris, Charles R., K. Jarrod Millman, Stéfan J. Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, and Nathaniel J. Smith (2020). "Array programming with NumPy". In: *Nature* 585.7825, pp. 357–362.

Hartung, K. and H.-P. Piepho (2007). "Are ordinal rating scales better than percent ratings? a statistical and "psychological" view". In: *Euphytica* 155.1-2, pp. 15–26. DOI: 10.1007/s10681-006-9296-z.

Hasan, Umut, Mamat Sawut, and Shuisen Chen (2019). "Estimating the leaf area index of winter wheat based on unmanned aerial vehicle rgb-image parameters". In: *Sustainability (Switzerland)* 11.23, pp. 1–11. DOI: 10.3390/su11236829.

Herrera, Juan M., Lilia Levy Häner, Fabio Mascher, Jürg Hiltbrunner, Dario Fossati, Cécile Brabant, Raphaël Charles, and Didier Pellet (2020). "Lessons From 20 Years of Studies of Wheat Genotypes in Multiple Environments and Under Contrasting Production Systems". In: *Frontiers in Plant Science* 10, p. 1745. DOI: 10.3389/fpls.2019.01745.

Hershman, Donald E. (2011). "Black "sooty" head mold of wheat". In: *Plant Pathology Fact Sheet PPFS-AG-SG-07*.

Hickey, Lee T., Amber N. Hafeez, Hannah Robinson, Scott A. Jackson, Soraya C. M. Leal-Bertioli, Mark Tester, Caixia Gao, Ian D. Godwin, Ben J. Hayes, and Brande B. H. Wulff (2019). "Breeding crops to feed 10 billion". In: *Nature Biotechnology* 37.7, pp. 744–754. DOI: 10.1038/s41587-019-0152-9.

Holzkämper, Annelie, Dario Fossati, Jürg Hiltbrunner, and Jürg Fuhrer (2015). "Spatial and temporal trends in agro-climatic limitations to production potentials for grain maize and winter wheat in Switzerland". In: *Regional Environmental Change* 15.1, pp. 109–122. DOI: 10.1007/s10113-014-0627-7.

Hong, Qingqing, Ling Jiang, Zhenghua Zhang, Shu Ji, Chen Gu, Wei Mao, Wenxi Li, Tao Liu, Bin Li, and Changwei Tan (2022). "A Lightweight Model for Wheat Ear Fusarium Head Blight Detection Based on RGB Images". In: *Remote Sensing* 14.14, p. 3481. DOI: 10.3390/rs14143481.

Hörtensteiner, S. (2006). "Chlorophyll Degradation During Senescence". In: *Annual Review of Plant Biology* 57.1, pp. 55–77. DOI: 10.1146/annurev.arplant.57.032905.105212.

Hu, Pengcheng, Bangyou Zheng, Qiaomin Chen, Swaantje Grunefeld, Malini Roy Choudhury, Javier Fernandez, Andries Potgieter, and Scott C. Chapman (2024). "Estimating aboveground biomass dynamics of wheat at small spatial scale by integrating crop growth and radiative transfer models with satellite remote sensing data". In: *Remote Sensing of Environment* 311, p. 114277. DOI: 10.1016/j.rse.2024.114277.

Huang, Linsheng, Taikun Li, Chuanlong Ding, Jinling Zhao, Dongyan Zhang, and Guijun Yang (2020). "Diagnosis of the severity of fusarium head blight of wheat ears on the basis of image and spectral feature fusion". In: *Sensors (Switzerland)* 20.10. DOI: 10.3390/s20102887.

Huang, Qiang, Xu Wang, Qi Gao, Alberto. Carraro, Marco Sozzi, and Francesco Marinello (2024). "Indicators to Digitization Footprint and How to Get Digitization Footprint (Part 2)". In: *Computers and Electronics in Agriculture* 224, p. 109206. DOI: 10.1016/j.compag.2024.109206.

Huete, Alfredo R. (1988). "A soil-adjusted vegetation index (SAVI)". In: *Remote Sensing of Environment* 25.3, pp. 295–309. DOI: 10.1016/0034-4257(88)90106-X.

Huete, Alfredo R., K Didan, T Miura, E.P Rodriguez, X Gao, and L.G Ferreira (2002). "Overview of the radiometric and biophysical performance of the MODIS vegetation indices". In: *Remote Sensing of Environment* 83.1-2, pp. 195–213. DOI: 10.1016/S0034-4257(02)00096-2.

Hufkens, Koen, Trevor F. Keenan, Lawrence B. Flanagan, Russell L. Scott, Carl J. Bernacchi, Eva Joo, Nathaniel A. Brunsell, Joseph Verfaillie, and Andrew D. Richardson (2016). "Productivity of North American grasslands is increased under future climate scenarios despite rising aridity". In: *Nature Climate Change* 6.7. Publisher: Nature Publishing Group UK London, pp. 710–714.

Hufkens, Koen, Eli K. Melaas, Michael L. Mann, Timothy Foster, Francisco Ceballos, Miguel Robles, and Berber Kramer (2019). "Monitoring crop phenology using a smartphone based near-surface remote sensing approach". In: *Agricultural and Forest Meteorology* 265, pp. 327–337. DOI: 10.1016/j.agrformet.2018.11.002.

Hund, Andreas, Lukas Kronenberg, Jonas Anderegg, Kang Yu, and Achim Walter (2019). "Non-invasive field phenotyping of cereal development". In: *Burleigh Dodds Series in Agricultural Science*. Ed. by Frank Ordon. Burleigh Dodds Science Publishing, pp. 249–292. ISBN: 978-1-78676-244-3.

Hunt, E. Raymond, C. S. T. Daughtry, Jan U. H. Eitel, and Dan S. Long (2011). "Remote sensing leaf chlorophyll content using a visible band index". In: *Agronomy Journal* 103.4. ISBN: 0002-1962, pp. 1090–1099. DOI: 10.2134/agronj2010.0395.

Hunt, E. Raymond, Paul C Doraiswamy, James E McMurtrey, Craig S T Daughtry, Eileen M Perry, and Bakhyt Akhmedov (2013). "A visible band index for remote sensing leaf chlorophyll content at the canopy scale". In: *International Journal of Applied Earth Observation and Geoinformation* 21, pp. 103–112. DOI: https://doi.org/10.1016/j.jag.2012.07.020.

Ide, Reiko and Hiroyuki Oguma (2010). "Use of digital cameras for phenological observations". In: *Ecological Informatics* 5.5. Publisher: Elsevier B.V., pp. 339–347. DOI: 10.1016/j.ecoinf.2010.07.002.

Idso, S. B., R. D. Jackson, P. J. Pinter, R. J. Reginato, and J. L. Hatfield (1981). "Normalizing the stress-degree-day parameter for environmental variability". In: *Agricultural Meteorology* 24.C, pp. 45–55. DOI: 10.1016/0002-1571(81)90032-7.

InnoVar (2025). *InnoVar - Next-generation variety testing for improved cropping on European farmland - H2020 InnoVar Project*. en-US. URL: https://www.h2020innovar.eu/ (visited on 03/02/2025).

Invite (2025). *Invite - INnovations in plant VarIety Testing in Europe - H2020 INVITE Project*. en-GB. URL: https://www.h2020-invite.eu/ (visited on 03/02/2025).

Jacob, Frédéric, François Petitcolin, Thomas Schmugge, Eric Vermote, Andrew French, and Kenta Ogawa (2004). "Comparison of land surface emissivity and radiometric temperature derived from MODIS and ASTER sensors". In: *Remote Sensing of Environment* 90.2, pp. 137–152. DOI: 10.1016/j.rse.2003.11.015.

Jeger, M. J. and S. L.H. Viljanen-Rollinson (2001). "The use of the area under the disease-progress curve (AUDPC) to assess quantitative disease resistance in crop cultivars". In: *Theoretical and Applied Genetics* 102.1, pp. 32–40. DOI: 10.1007/s001220051615.

Jensen, J.R. (2009). *Remote Sensing of the Environment: An Earth Resource Perspective 2/e*. Pearson Education. ISBN: 978-81-317-1680-9.

Jensen, T., A. Apan, F. Young, and L. Zeller (2007). "Detecting the attributes of a wheat crop using digital imagery acquired from a low-altitude platform". In: *Computers and Electronics in Agriculture* 59.1-2, pp. 66–77. DOI: 10.1016/j.compag.2007.05.004.

Jia, Jiyu, Meng Xu, Shuikuan Bei, Hongzhi Zhang, Li Xiao, Yonghong Gao, Yongqiang Zhang, Lihan Sai, Lihua Xue, Junjie Lei, and Xu Qiao (2021). "Impact of reduced light intensity on wheat yield and quality: Implications for agroforestry systems". In: *Agroforestry Systems* 95.8, pp. 1689–1701. DOI: 10.1007/s10457-021-00668-w.

Jiang, Le and Shafiqul Islam (1999). "A methodology for estimation of surface evapotranspiration over large areas using remote sensing observations". In: *Geophysical Research Letters* 26.17, pp. 2773–2776. DOI: 10.1029/1999GL006049.

Jimenez-Berni, Jose A., David M. Deery, Pablo Rozas-Larraondo, Anthony (Tony) G. Condon, Greg J. Rebetzke, Richard A. James, William D. Bovill, Robert T. Furbank, and Xavier R. R. Sirault (2018). "High Throughput Determination of Plant Height, Ground Cover, and Above-Ground Biomass in Wheat with LiDAR". In: *Frontiers in Plant Science* 9, p. 237. DOI: 10.3389/fpls.2018.00237.

Jimenez-Berni, Jose A., P.J. Zarco-Tejada, G. Sepulcre-Cantó, E. Fereres, and F. Villalobos (2009). "Mapping canopy conductance and CWSI in olive orchards using high resolution thermal remote sensing imagery". In: *Remote Sensing of Environment* 113.11, pp. 2380–2388. DOI: 10.1016/j.rse.2009.06.018.

Jimenez-Berni, Jose A., Pablo J. Zarco-Tejada, Lola Suarez, and Elias Fereres (2009). "Thermal and narrowband multispectral remote sensing for vegetation monitoring from an unmanned aerial vehicle". In: *IEEE Transactions on Geoscience and Remote Sensing* 47.3, pp. 722–738. DOI: 10.1109/TGRS.2008.2010457.

Jin, Xiuliang, Shouyang Liu, Frédéric Baret, Matthieu Hemerlé, and Alexis Comar (2017). "Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery". In: *Remote Sensing of Environment* 198. Publisher: Elsevier Inc., pp. 105–114. DOI: 10.1016/j.rse.2017.06.007.

Jones, Hamlyn and Xavier Sirault (2014). "Scaling of Thermal Images at Different Spatial Resolution: The Mixed Pixel Problem". In: *Agronomy* 4.3, pp. 380–396. DOI: 10.3390/agronomy4030380.

Jones, Hamlyn G., Rachid Serraj, Brian R. Loveys, Lizhong Xiong, Ashley Wheaton, and Adam H. Price (2009). "Thermal infrared imaging of crop canopies for the remote diagnosis and quantification of plant responses to water stress in the field". In: *Functional Plant Biology* 36.11, p. 978. DOI: 10.1071/FP09123.

Jones, Hamlyn G., Manfred Stoll, Tiago Santos, Claudia de Sousa, M. Manuela Chaves, and Olga M. Grant (2002). "Use of infrared thermography for monitoring stomatal closure in the field: application to grapevine". In: *Journal of experimental botany* 53.378, pp. 2249–2260.

Jordan, Carl F. (1969). "Derivation of Leaf-Area Index from Quality of Light on the Forest Floor". In: *Ecology* 50.4, pp. 663–666. DOI: 10.2307/1936256.

Joshi, A. K., M. Kumari, V. P. Singh, C. M. Reddy, S. Kumar, J. Rane, and R. Chand (2007). "Stay green trait: variation, inheritance and its association with spot blotch resistance in spring wheat (Triticum aestivum L.)" In: *Euphytica* 153.1-2, pp. 59–71. DOI: 10.1007/s10681-006-9235-z.

Kahiluoto, Helena, Janne Kaseva, Jan Balek, Jørgen E. Olesen, Margarita Ruiz-Ramos, Anne Gobin, Kurt Christian Kersebaum, Jozef Takáč, Francoise Ruget, Roberto Ferrise, Pavol Bezak, Gemma Capellades, Camilla Dibari, Hanna Mäkinen, Claas Nendel, Domenico Ventrella, Alfredo Rodríguez, Marco Bindi, and Mirek Trnka (2019). "Decline in climate

resilience of european wheat". In: *Proceedings of the National Academy of Sciences of the United States of America* 116.1, pp. 123–128. DOI: 10.1073/pnas.1804387115.

Kamau, Hannah, Shahrear Roman, and Lisa Biber-Freudenberger (2023). "Nearly half of the world is suitable for diversified farming for sustainable intensification". In: *Communications Earth & Environment* 4.1, p. 446. DOI: 10.1038/s43247-023-01062-3.

Karlsson, Ida, Paula Persson, and Hanna Friberg (2021). "Fusarium Head Blight From a Microbiome Perspective". In: *Frontiers in Microbiology* 12, p. 628373. DOI: 10.3389/fmicb.2021.628373.

Kavaliauskas, Ardas, Renaldas Žydelis, Fabio Castaldi, Ona Auškalnienė, and Virmantas Povilaitis (2023). "Predicting Maize Theoretical Methane Yield in Combination with Ground and UAV Remote Data Using Machine Learning". In: *Plants* 12.9, p. 1823. DOI: 10.3390/plants12091823.

Kawashima, S. (1998). "An Algorithm for Estimating Chlorophyll Content in Leaves Using a Video Camera". In: *Annals of Botany* 81.1, pp. 49–54. DOI: 10.1006/anbo.1997.0544.

Keenan, T. F., B. Darby, E. Felts, O. Sonnentag, M. A. Friedl, K. Hufkens, J. O'Keefe, S. Klosterman, J. W. Munger, M. Toomey, and A. D. Richardson (2014). "Tracking forest phenology and seasonal physiology using digital repeat photography: a critical assessment". In: *Ecological Applications* 24.6, pp. 1478–1489. DOI: 10.1890/13-0652.1.

Keller, Beat, Shizue Matsubara, Uwe Rascher, Roland Pieruschka, Angelina Steier, Thorsten Kraska, and Onno Muller (2019). "Genotype Specific Photosynthesis x Environment Interactions Captured by Automated Fluorescence Canopy Scans Over Two Fluctuating Growing Seasons". In: *Frontiers in Plant Science* 10, p. 1482. DOI: 10.3389/fpls.2019.01482.

Kelly, Julia, Natascha Kljun, Per-Ola Olsson, Laura Mihai, Bengt Liljeblad, Per Weslien, Leif Klemedtsson, and Lars Eklundh (2019). "Challenges and best practices for deriving temperature data from an uncalibrated UAV thermal infrared camera". In: *Remote Sensing* 11.5, p. 567. DOI: 10.3390/rs11050567.

Kholová, Jana, Milan Oldřich Urban, James Cock, Jairo Arcos, Elizabeth Arnaud, Destan Aytekin, Vania Azevedo, Andrew P Barnes, Salvatore Ceccarelli, Paul Chavarriaga, Joshua N Cobb, David Connor, Mark Cooper, Peter Craufurd, Daniel Debouck, Robert Fungo, Stefania Grando, Graeme L Hammer, Carlos E Jara, Charlie Messina, Gloria Mosquera, Eileen Nchanji, Eng Hwa Ng, Steven Prager, Sindhujan Sankaran, Michael Selvaraj, François Tardieu, Philip Thornton, Sandra P Valdes-Gutierrez, Jacob Van Etten, Peter Wenzl, and Yunbi Xu (2021). "In pursuit of a better world: crop improvement and the CGIAR". In: *Journal of Experimental Botany* 72.14. Ed. by Mathew Reynolds, pp. 5158–5179. DOI: 10.1093/jxb/erab226.

Kipp, Sebastian, Bodo Mistele, and Urs Schmidhalter (2014). "Identification of stay-green and early senescence phenotypes in high-yielding winter wheat, and their relationship to grain yield and grain protein concentration using high-throughput phenotyping techniques". In: *Functional Plant Biology* 41.3, p. 227. DOI: 10.1071/FP13221.

Klosterman, S. T., K. Hufkens, J. M. Gray, E. Melaas, O. Sonnentag, I. Lavine, L. Mitchell, R. Norman, M. A. Friedl, and A. D. Richardson (2014). "Evaluating remote sensing of deciduous forest phenology at multiple spatial scales using PhenoCam imagery". In: *Biogeosciences* 11.16, pp. 4305–4320. DOI: 10.5194/bg-11-4305-2014.

Künzer, Claudia and Stefan Dech (2013). *Thermal Infrared Remote Sensing : Sensors, Methods, Applications*. Dordrecht : Springer.

Kurc, SA and LM Benton (2010). "Digital image-derived greenness links deep soil moisture to carbon uptake in a creosotebush-dominated shrubland". In: *Journal of Arid Environments* 74.5. Publisher: Elsevier, pp. 585–594.

Laidig, F., T. Drobek, and U. Meyer (2008). "Genotypic and environmental variability of yield for cultivars from 30 different crops in German official variety trials". In: *Plant Breeding* 127.6, pp. 541–547. DOI: 10.1111/j.1439-0523.2008.01564.x.

Lancashire, Peter D., H. Bleiholder, T. Van Den Boom, P. Langelüddeke, R. Stauss, Elfriede Weber, and A. Witzenberger (1991). "A uniform decimal code for growth stages of crops and weeds". In: *Annals of Applied Biology* 119.3, pp. 561–601. DOI: 10.1111/j.1744-7348.1991.tb04895.x.

Langer, Simon M., C. Friedrich H. Longin, and Tobias Würschum (2014). "Flowering time control in European winter wheat". In: *Frontiers in Plant Science* 5.OCT, pp. 1–11. DOI: 10.3389/fpls.2014.00537.

Lauterberg, Madita, Henning Tschiersch, Yusheng Zhao, Markus Kuhlmann, Ingo Mücke, Roberto Papa, Elena Bitocchi, and Kerstin Neumann (2024). "Implementation of theoretical non-photochemical quenching (NPQ(T)) to investigate NPQ of chickpea under drought stress with High-throughput Phenotyping". In: *Scientific Reports* 14.1, p. 13970. DOI: 10.1038/s41598-024-63372-6.

Lee, Soo Yee, Ahmed Mediani, Maulidiani Maulidiani, Alfi Khatib, Intan Safinar Ismail, Norhasnida Zawawi, and Faridah Abas (2017). "Comparison of partial least squares and random forests for evaluating relationship between phenolics and bioactivities of *Neptunia oleracea* ". In: *Journal of the Science of Food and Agriculture* 98.1, pp. 240–252. DOI: 10.1002/jsfa.8462.

Lepekhov, S. B. (2022). "Canopy temperature depression for droughtand heat stress tolerance in wheat breeding". In: *Vavilov Journal of Genetics and Breeding* 26.2, pp. 196–201. DOI: 10.18699/VJGB-22-24.

Levasseur-Garcia, Cecile (2018). "Updated overview of infrared spectroscopy methods for detecting mycotoxins on cereals (corn, wheat, and barley)". In: *Toxins* 10.1. DOI: 10.3390/toxins10010038.

Levy, Lilia, Numa Courvoisier, Sandro Rechsteiner, Juan Herrera, Cécile Brabant, Andreas Hund, Thomas Weissflog, Hansueli Dierauer, and Didier Pellet (2017). "Winterweizen: Bilanz aus 15 Jahren Sortenprüfung unter extensiven Anbaubedingungen". In: *Agrarforschung Schweiz* 8.7-8, pp. 300–309.

Li, Huawei, Dong Jiang, Bernd Wollenweber, Tingbo Dai, and Weixing Cao (2010). "Effects of shading on morphology, physiology and grain yield of winter wheat". In: *European Journal of Agronomy* 33.4, pp. 267–275. DOI: 10.1016/j.eja.2010.07.002.

Li, Lei, Qin Zhang, and Danfeng Huang (2014). "A Review of Imaging Techniques for Plant Phenotyping". In: *Sensors* 14.11, pp. 20078–20111. DOI: 10.3390/s141120078.

Li, Wei, Dong Li, Shouyang Liu, Frédéric Baret, Zhiyuan Ma, Can He, Timothy A. Warner, Caili Guo, Tao Cheng, Yan Zhu, Weixing Cao, and Xia Yao (2023). "RSARE: A physically-based vegetation index for estimating wheat green LAI to mitigate the impact of leaf chlorophyll content and residue-soil background". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 200, pp. 138–152. DOI: 10.1016/j.isprsjprs.2023.05.012.

Lillesand, Thomas M., Ralph W. Kiefer, and Jonathan W. Chipman (2015). *Remote sensing and image interpretation*. 7th edition. Hoboken, N.J: John Wiley. ISBN: 1-118-34328-X.

Liu, Yujie, Christoph Bachofen, Raphaël Wittwer, Gicele Silva Duarte, Qing Sun, Valentin H. Klaus, and Nina Buchmann (2022). "Using PhenoCams to track crop phenology and explain the effects of different cropping systems on yield". In: *Agricultural Systems* 195, p. 103306. DOI: https://doi.org/10.1016/j.agsy.2021.103306.

Longchamps, Louis and William Philpot (2023). "Full-Season Crop Phenology Monitoring Using Two-Dimensional Normalized Difference Pairs". In: *Remote Sensing* 15.23, p. 5565. DOI: 10.3390/rs15235565.

Lopes, Marta S. and Matthew P. Reynolds (2010). "Partitioning of assimilates to deeper roots is associated with cooler canopies and increased yield under drought in wheat". In: *Functional Plant Biology* 37.2, p. 147. DOI: 10.1071/FP09121.

López-Lozano, Raúl and Bettina Baruth (2019). "An evaluation framework to build a cost-efficient crop monitoring system. Experiences from the extension of the European crop monitoring system". In: *Agricultural Systems* 168, pp. 231–246. DOI: 10.1016/j.agsy.2018.04.002.

Lorence, Argelia and Karina Medina Jimenez, eds. (2022). *High-Throughput Plant Phenotyping: Methods and Protocols*. Vol. 2539. Methods in Molecular Biology. New York, NY: Springer US. ISBN: 978-1-07-162537-8.

Lorenz, K. (1986). "Effects of blackpoint on grain composition and baking quality of New Zealand wheat". In: *New Zealand Journal of Agricultural Research* 29.4, pp. 711–718. DOI: 10.1080/00288233.1986.10430468.

Louhaichi, Mounir, Michael M. Borman, and Douglas E. Johnson (2001). "Spatially Located Platform and Aerial Photography for Documentation of Grazing Impacts on Wheat". In: *Geocarto International* 16.1, pp. 65–70. DOI: 10.1080/10106040108542184.

Lowe, David G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2, pp. 91–110. DOI: 10.1023/B:VISI.0000029664.99615.94.

Lu, Ning, Jie Zhou, Zixu Han, Dong Li, Qiang Cao, Xia Yao, Yongchao Tian, Yan Zhu, Weixing Cao, and Tao Cheng (2019). "Improved estimation of aboveground biomass in wheat from RGB imagery and point cloud data acquired with a low-cost unmanned aerial vehicle system". In: *Plant Methods* 15.1, p. 17. DOI: 10.1186/s13007-019-0402-3.

Ludovisi, Riccardo, Flavia Tauro, Riccardo Salvati, Sacha Khoury, Giuseppe Scarascia Mugnozza, and Antoine Harfouche (2017). "Uav-based thermal imaging for high-throughput field phenotyping of black poplar response to drought". In: *Frontiers in Plant Science* 8.September, pp. 1–18. DOI: 10.3389/fpls.2017.01681.

Maes, Wouter H., Alfredo R. Huete, and Kathy Steppe (2017). "Optimizing the Processing of UAV-Based Thermal Imagery". In: *Remote Sensing* 9.5, p. 476. DOI: 10.3390/rs9050476.

Maes, Wouter H., T. Pashuysen, A. Trabucco, F. Veroustraete, and B. Muys (2011). "Does energy dissipation increase with ecosystem succession? Testing the ecosystem exergy theory combining theoretical simulations and thermal remote sensing observations". In: *Ecological Modelling* 222.23-24, pp. 3917–3941. DOI: 10.1016/j.ecolmodel.2011.08.028.

Maes, Wouter H. and Kathy Steppe (2012). "Estimating evapotranspiration and drought stress with ground-based thermal remote sensing in agriculture: a review". In: *Journal of Experimental Botany* 63.13, pp. 4671–4712. DOI: 10.1093/jxb/ers165.

Mahlein, Anne-Katrin, Elias Alisaac, Ali Al Masri, Jan Behmann, Heinz-Wilhelm Dehne, and Erich-Christian Oerke (2019). "Comparison and combination of thermal, fluorescence, and hyperspectral imaging for monitoring fusarium head blight of wheat on spikelet scale". In: *Sensors* 19.10, p. 2281. DOI: 10.3390/s19102281.

Mahrookashani, A., S. Siebert, H. Hüging, and F. Ewert (2017). "Independent and combined effects of high temperature and drought stress around anthesis on wheat". In: *Journal of Agronomy and Crop Science* 203.6, pp. 453–463. DOI: 10.1111/jac.12218.

Malbéteau, Yoann, Kasper Johansen, Bruno Aragon, Samir K. Al-Mashhawari, and Matthew F. McCabe (2021). "Overcoming the challenges of thermal infrared orthomosaics using a swath-based approach to correct for dynamic temperature and wind effects". In: *Remote Sensing* 13.16, p. 3255. DOI: 10.3390/rs13163255.

Mankins, John C. (1995). "Technology readiness levels". In: *White Paper, April* 6.1995. Publisher: NASA, p. 1995.

Mao, Wenhua, Yiming Wang, and Yueqing Wang (2003). "Real-time Detection of Between-row Weeds Using Machine Vision". In: ASABE Paper No. 031004. St. Joseph, MI: ASABE. DOI: 10.13031/2013.15381. URL: https://elibrary.asabe.org/abstract.asp?aid=15381&t=5.

Marinello, Francesco (2023). "Digitization Footprint". In: *Encyclopedia of Digital Agricultural Technologies*. Ed. by Qin Zhang. Cham: Springer International Publishing, pp. 356–363. ISBN: 978-3-031-24861-0.

Mason, R. and Ravi Singh (2014). "Considerations When Deploying Canopy Temperature to Select High Yielding Wheat Breeding Lines under Drought and Heat Stress". In: *Agronomy* 4.2, pp. 191–201. DOI: 10.3390/agronomy4020191.

McMaster, Gregory (1997). "Growing degree-days: one equation, two interpretations". In: *Agricultural and Forest Meteorology* 87.4, pp. 291–300. DOI: 10.1016/S0168-1923(97)00027-0.

McMullen, Marcia, Roger Jones, and Dale Gallenberg (1997). "Scab of wheat and barley: a re-emerging disease of devastating impact". In: *Plant disease* 81.12, pp. 1340–1348.

Mehmood, Tahir, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø (2012). "A review of variable selection methods in Partial Least Squares Regression". In: *Chemometrics and Intelligent Laboratory Systems* 118, pp. 62–69. DOI: 10.1016/j.chemolab.2012.07.010.

Meier, F., D. Scherer, J. Richters, and A. Christen (2011). "Atmospheric correction of thermal-infrared imagery of the 3-D urban environment acquired in oblique viewing geometry". In: *Atmospheric Measurement Techniques* 4.5, pp. 909–922. DOI: 10.5194/amt-4-909-2011.

Meng, Xiangchen, Jie Cheng, and Shunlin Liang (2017). "Estimating Land Surface Temperature from Feng Yun-3C/MERSI Data Using a New Land Surface Emissivity Scheme". In: *Remote Sensing* 9.12, p. 1247. DOI: 10.3390/rs9121247.

Merzlyak, Mark N., Anatoly A. Gitelson, Olga B. Chivkunova, and Victor Yu. Rakitin (1999). "Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening". In: *Physiologia Plantarum* 106.1, pp. 135–141. DOI: 10.1034/j.1399-3054.1999.106119.x.

Mesas-Carrascosa, Francisco-Javier, Fernando Pérez-Porras, Jose Meroño De Larriva, Carlos Mena Frau, Francisco Agüera-Vega, Fernando Carvajal-Ramírez, Patricio Martínez-Carricondo, and Alfonso García-Ferrer (2018). "Drift correction of lightweight microbolometer thermal sensors on-board unmanned aerial vehicles". In: *Remote Sensing* 10.4, p. 615. DOI: 10.3390/rs10040615.

Messina, Gaetano and Giuseppe Modica (2020). "Applications of UAV thermal imagery in precision agriculture: State of the art and future research outlook". In: *Remote Sensing* 12.9. DOI: 10.3390/RS12091491.

Mesterhazy, Akos (2020). "Updating the Breeding Philosophy of Wheat to Fusarium Head Blight (FHB): Resistance Components, QTL Identification, and Phenotyping—A Review". In: *Plants* 9.12, p. 1702. DOI: 10.3390/plants9121702.

Mevik, Bjørn-Helge and Ron Wehrens (2007). "The **pls** Package: Principal component and partial least squares regression in *R*". In: *Journal of Statistical Software* 18.2. DOI: 10.18637/jss.v018.i02.

Meyer, George E., T. Mehta, M. F. Kocher, D. A. Mortensen, and A. Samal (1998). "Textural imaging and discriminant analysis for distinguishingweeds for spot spraying". In: *Transactions of the ASAE* 41.4, pp. 1189–1197.

Meyer, George E. and João Camargo Neto (2008). "Verification of color vegetation indices for automated crop imaging applications". In: *Computers and Electronics in Agriculture* 63.2, pp. 282–293. DOI: 10.1016/j.compag.2008.03.009.

Michel, Sebastian, Franziska Löschenberger, Christian Ametz, and Hermann Bürstmayr (2023). "Improving the efficiency of multi-location field trials with complete and incomplete relationship information". In: *Euphytica* 219.1, p. 10. DOI: 10.1007/s10681-022-03142-5.

Michel, V. (2001). "La sélection de variétés de blé et de triticale résistantes aux maladies". fr. In: *Revue suisse d'agriculture (Switzerland)*.

Miedaner, T., C. J. R. Cumagun, and S. Chakraborty (2008). "Population Genetics of Three Important Head Blight Pathogens *Fusarium graminearum, F. pseudograminearum* and *F. culmorum*". In: *Journal of Phytopathology* 156.3, pp. 129–139. DOI: 10.1111/j.1439-0434.2007.01394.x.

Migliavacca, Mirco, Marta Galvagno, Edoardo Cremonese, Micol Rossini, Michele Meroni, Oliver Sonnentag, Sergio Cogliati, Giovanni Manca, Fabrizio Diotri, Lorenzo Busetto, Alessandro Cescatti, Roberto Colombo, Francesco Fava, Umberto Morra di Cella, Emiliano Pari, Consolata Siniscalco, and Andrew D. Richardson (2011). "Using digital repeat photography and eddy covariance data to model grassland phenology and photosynthetic CO2 uptake". In: *Agricultural and Forest Meteorology* 151.10. Publisher: Elsevier B.V., pp. 1325–1337. DOI: 10.1016/j.agrformet.2011.05.012.

Moll, Eckard, Kerstin Flath, and Hans-Peter Piehpho (2000). *Die Prüfung von Pflanzen auf ihre Widerstandsfähigkeit gegen Schadorganismen in der Biologischen Bundesanstalt - Teil 3*. Mitteilungen aus der Biologischen Bundesanstalt für Land- und Forstwirtschaft Berlin-Dahlem Heft 374. Berlin: Parey Buchverlag. ISBN: 3-8263-3256-3.

Montazeaud, Germain, Handan Karatoğma, Ibrahim Özturk, Pierre Roumet, Martin Ecarnot, Jose Crossa, Emel Özer, Fatih Özdemir, and Marta S. Lopes (2016). "Predicting wheat maturity and stay–green parameters by modeling spectral reflectance measurements and their contribution to grain yield under rainfed conditions". In: *Field Crops Research* 196, pp. 191–198. DOI: 10.1016/j.fcr.2016.06.021.

Mu, H., D. Jiang, B. Wollenweber, T. Dai, Q. Jing, and W. Cao (2010). "Long-term Low Radiation Decreases Leaf Photosynthesis, Photochemical Efficiency and Grain Yield in Winter Wheat". In: *Journal of Agronomy and Crop Science* 196.1, pp. 38–47. DOI: 10.1111/j.1439-037X.2009.00394.x.

Munkvold, Gary P. (2017). "Fusarium Species and Their Associated Mycotoxins". In: *Mycotoxigenic Fungi: Methods and Protocols*. Ed. by Antonio Moretti and Antonia Susca. New York, NY: Springer New York, pp. 51–106. ISBN: 978-1-4939-6707-0. DOI: 10.1007/978-1-4939-6707-0_4.

Mustafa, Ghulam, Hengbiao Zheng, Wei Li, Yuming Yin, Yongqing Wang, Meng Zhou, Peng Liu, Muhammad Bilal, Haiyan Jia, Guoqiang Li, Tao Cheng, Yongchao Tian, Weixing Cao, Yan Zhu, and Xia Yao (2023). "Fusarium head blight monitoring in wheat ears using machine learning and multimodal data from asymptomatic to symptomatic periods". In: *Frontiers in Plant Science* 13, p. 1102341. DOI: 10.3389/fpls.2022.1102341.

Naito, Hiroki, Satoshi Ogawa, Milton Orlando Valencia, Hiroki Mohri, Yutaka Urano, Fumiki Hosoi, Yo Shimizu, Alba Lucia Chavez, Manabu Ishitani, Michael Gomez Selvaraj, and Kenji Omasa (2017). "Estimating rice yield related traits and quantitative trait loci analysis under different nitrogen treatments using a simple tower-based field phenotyping system with modified single-lens reflex cameras". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 125, pp. 50–62. DOI: 10.1016/j.isprsjprs.2017.01.010.

Nicodemus, Fred Edwin (1977). *Geometrical considerations and nomenclature for reflectance*. Vol. 160. US Department of Commerce, National Bureau of Standards Washington, DC, USA. DOI: 10.6028/NBS.MONO.160.

Niedbała, Gniewko, Anna Tratwal, Magdalena Piekutowska, Tomasz Wojciechowski, and Jarosław Uglis (2022). "A Framework for Financing Post-Registration Variety Testing

System: A Case Study from Poland". In: *Agronomy* 12.2, p. 325. DOI: 10.3390/agronomy12020325.

Nugent, Paul W., Joseph A. Shaw, and Nathan J. Pust (2013). "Correcting for focal-plane-array temperature dependence in microbolometer infrared cameras lacking thermal stabilization". In: *Optical Engineering* 52.6, p. 061304. DOI: 10.1117/1.OE.52.6.061304.

Oakey, Helena, Arūnas Verbyla, Wayne Pitchford, Brian Cullis, and Haydn Kuchel (2006). "Joint modeling of additive and non-additive genetic line effects in single field trials". In: *Theoretical and Applied Genetics* 113.5, pp. 809–819. DOI: 10.1007/s00122-006-0333-z.

Oberholzer, Simon, Volker Prasuhn, and Andreas Hund (2017). "Crop water use under Swiss pedoclimatic conditions – Evaluation of lysimeter data covering a seven-year period". In: *Field Crops Research* 211, pp. 48–65. DOI: 10.1016/j.fcr.2017.06.003.

Oldenburg, Elisabeth and Frank Ellner (2015). "Distribution of disease symptoms and mycotoxins in maize ears infected by Fusarium culmorum and Fusarium graminearum". In: *Mycotoxin Research* 31.3, pp. 117–126. DOI: 10.1007/s12550-015-0222-x.

Padfield, Daniel and G. Matheso (2020). "Package 'Nls. multstart". In: *CRAN: Vienna, Austria*, pp. 1–5.

Parvej, Md. Rasel, David L. Holshouser, Robert J. Kratochvil, Cory M. Whaley, E. James Dunphy, Gregory W. Roth, and Giovani S. Faé (2020). "Early high-moisture wheat harvest improves double-crop system: I. Wheat yield and quality". In: *Crop Science* 60.5, pp. 2633–2649. DOI: 10.1002/csc2.20172.

Pask, Alistair J. D., J. Pietragalla, D. M. Mullan, and M. P. Reynolds (2012). *Physiological breeding II: a field guide to wheat phenotyping*. Cimmyt. ISBN: 978-970-648-182-5.

Peiris, Kamaranga H.S., William W. Bockus, and Floyd E. Dowell (2016). "Near-infrared spectroscopic evaluation of single-kernel deoxynivalenol accumulation and fusarium head blight resistance components in wheat". In: *Cereal Chemistry* 93.1, pp. 25–31. DOI: 10.1094/CCHEM-03-15-0057-R.

Peiris, Kamaranga H.S., M. O. Pumphrey, Y. Dong, E. B. Maghirang, W. Berzonsky, and Floyd E. Dowell (2010). "Near-infrared spectroscopic method for identification of Fusarium head blight damage and prediction of deoxynivalenol in single wheat kernels". In: *Cereal Chemistry* 87.6, pp. 511–517. DOI: 10.1094/CCHEM-01-10-0006.

Pellan, Lucile, Cheikh Ahmeth Tidiane Dieye, Noël Durand, Angélique Fontana, Sabine Schorr-Galindo, and Caroline Strub (2021). "Biocontrol Agents Reduce Progression and Mycotoxin Production of Fusarium graminearum in Spikelets and Straws of Wheat". In: *Toxins* 13.9, p. 597. DOI: 10.3390/toxins13090597.

Perich, Gregor, Andreas Hund, Jonas Anderegg, Lukas Roth, Martin P. Boer, Achim Walter, Frank Liebisch, and Helge Aasen (2020). "Assessment of multi-image unmanned aerial vehicle based high-throughput field phenotyping of canopy temperature". In: *Frontiers in Plant Science* 11.February, pp. 1–17. DOI: 10.3389/fpls.2020.00150.

Piepho, H. P. and E. R. Williams (2010). "Linear variance models for plant breeding trials". In: *Plant Breeding* 129.1, pp. 1–8. DOI: 10.1111/j.1439-0523.2009.01654.x.

Piepho, Hans-Peter, Jens Möhring, Torben Schulz-Streeck, and Joseph O. Ogutu (2012). "A stage-wise approach for the analysis of multi-environment trials: Stage-wise analysis of trials". In: *Biometrical Journal* 54.6, pp. 844–860. DOI: 10.1002/bimj.201100219.

Pinto, Francisco, Mainassara Zaman-Allah, Matthew P. Reynolds, and Urs Schulthess (2023). "Satellite imagery for high-throughput phenotyping in breeding plots". In: *Frontiers in Plant Science* 14, p. 1114670. DOI: 10.3389/fpls.2023.1114670.

Poursafar, A., Y. Ghosta, and M. Javan-Nikkhah (2016). "A taxonomic study on Stemphylium species associated with black (sooty) head mold of wheat and barley in Iran". In: *Mycologia Iranica* 3.2. DOI: 10.22043/mi.2017.26183.

Prashar, Ankush and Hamlyn Jones (2014). "Infra-Red Thermography as a High-Throughput Tool for Field Phenotyping". In: *Agronomy* 4.3, pp. 397–417. DOI: 10.3390/agronomy4030397.

PSI Photon Systems Instruments (2021). *FluorPen FP 110 Instruction Guide*. URL: https://handheld.psi.cz/products/fluorpen-and-par-fluorpen/#download.

QGIS Development Team (2022). *QGIS geographic information system*. URL: https://www.qgis.org.

Qiu, Ruicheng, Ce Yang, Ali Moghimi, Man Zhang, Brian J. Steffenson, and Cory D. Hirsch (2019). "Detection of Fusarium Head Blight in wheat using a deep neural network and color imaging". In: *Remote Sensing* 11.22. DOI: 10.3390/rs11222658.

R Development Core Team (2022). *R: A language and environment for statistical computing*. Place: Vienna, Austria. URL: http://www.r-project.org.

Ray, Deepak K., Navin Ramankutty, Nathaniel D. Mueller, Paul C. West, and Jonathan A. Foley (2012). "Recent patterns of crop yield growth and stagnation". In: *Nature Communications* 3.1, p. 1293. DOI: 10.1038/ncomms2296.

Rebetzke, Greg J., Allan R. Rattey, Graham D. Farquhar, Richard A. Richards, and Anthony (Tony) G. Condon (2013). "Genomic regions for canopy temperature and their genetic association with stomatal conductance and grain yield in wheat". In: *Functional Plant Biology* 40.1, p. 14. DOI: 10.1071/FP12184.

Reynolds, Daniel, Joshua Ball, Alan Bauer, Robert Davey, Simon Griffiths, and Ji Zhou (2019). "CropSight: A scalable and open-source information management system for distributed plant phenotyping and IoT-based crop management". In: *GigaScience* 8.3. Publisher: Oxford University Press, pp. 1–11. DOI: 10.1093/gigascience/giz009.

Reynolds, Matthew P., Andrew Borrell, H. J. Braun, G. O. Edmeades, Richard Flavell, Jeff Gwyn, David Jordan, K. V. Pixley, and G. J. Rebetzke (2019). "Translational research for climate resilient, higher yielding crops". In: *Crop Breeding, Genetics and Genomics*. DOI: 10.20900/cbgg20190016.

Reynolds, Matthew P., Scott C. Chapman, Leonardo Crespo-Herrera, Gemma Molero, Suchismita Mondal, Diego N.L. Pequeno, Francisco Pinto, Francisco J. Pinera-Chavez, Jesse Poland, Carolina Rivera-Amado, Carolina Saint Pierre, and Sivakumar Sukumaran (2020). "Breeder friendly phenotyping". In: *Plant Science* 295, p. 110396. DOI: 10.1016/j.plantsci.2019.110396.

Reynolds, Matthew P., Alistair J. D. Pask, and D. M. Mullan (2012). *Physiological breeding I: interdisciplinary approaches to improve crop adaptation*. CIMMYT.

Ribeiro-Gomes, Krishna, David Hernández-López, José Ortega, Rocío Ballesteros, Tomás Poblete, and Miguel Moreno (2017). "Uncooled thermal camera calibration and optimization of the photogrammetry process for UAV applications in agriculture". In: *Sensors* 17.10, p. 2173. DOI: 10.3390/s17102173.

Richardson, Andrew D, Trevor F Keenan, Mirco Migliavacca, Youngryel Ryu, Oliver Sonnentag, and Michael Toomey (2013). "Climate change, phenology, and phenological control of vegetation feedbacks to the climate system". In: *Agricultural and Forest Meteorology* 169. Publisher: Elsevier, pp. 156–173.

Richardson, Andrew D. (2019). "Tracking seasonal rhythms of plants in diverse ecosystems with digital camera imagery". In: *New Phytologist* 222.4, pp. 1742–1750. DOI: 10.1111/nph.15591.

Richardson, Andrew D., Bobby H. Braswell, David Y. Hollinger, Julian P. Jenkins, and Scott V. Ollinger (2009). "Near-surface remote sensing of spatial and temporal variation in canopy phenology". In: *Ecological Applications* 19.6. Publisher: Wiley Online Library, pp. 1417–1428.

Richardson, Andrew D., Koen Hufkens, Tom Milliman, Donald M. Aubrecht, Min Chen, Josh M. Gray, Miriam R. Johnston, Trevor F. Keenan, Stephen T. Klosterman, Margaret Kosmala, Eli K. Melaas, Mark A. Friedl, and Steve Frolking (2018). "Tracking vegetation phenology across diverse North American biomes using PhenoCam imagery". In: *Scientific Data* 5.1, p. 180028. DOI: 10.1038/sdata.2018.28.

Richardson, Andrew D., Julian P. Jenkins, Bobby H. Braswell, David Y. Hollinger, Scott V. Ollinger, and Marie Louise Smith (2007). "Use of digital webcam images to track spring green-up in a deciduous broadleaf forest". In: *Oecologia* 152.2, pp. 323–334. DOI: 10.1007/s00442-006-0657-z.

Rife, Trevor W. and Jesse A. Poland (2014). "Field book: an open-source application for field data collection on android". In: *Crop Science* 54.4. Publisher: Wiley Online Library, pp. 1624–1627.

Ripley, Brian and Jim Ramsey (2024). *Package 'pspline'*. URL: https://cran.r-project.org/web/packages/pspline/ (visited on 04/05/2025).

Roche, Dominique (2015). "Stomatal Conductance Is Essential for Higher Yield Potential of C_3 Crops". In: *Critical Reviews in Plant Sciences* 34.4, pp. 429–453. DOI: 10.1080/07352689.2015.1023677.

Rodríguez-Álvarez, María Xosé, Martin P. Boer, . van Eeuwijk, and Paul H. C. Eilers (2018). "Correcting for spatial heterogeneity in plant breeding experiments with P-splines". In: *Spatial Statistics* 23, pp. 52–71. DOI: 10.1016/j.spasta.2017.10.003.

Rogger, Julian, Andreas Hund, Dario Fossati, and Annelie Holzkämper (2021). "Can Swiss wheat varieties escape future heat stress?" In: *European Journal of Agronomy* 131, p. 126394. DOI: 10.1016/j.eja.2021.126394.

Romano, Giuseppe, Shamaila Zia, Wolfram Spreer, Ciro Sanchez, Jill Cairns, Jose Luis Araus, and Joachim Müller (2011). "Use of thermography for high throughput phenotyping of tropical maize adaptation in water stress". In: *Computers and Electronics in Agriculture* 79.1, pp. 67–74. DOI: 10.1016/j.compag.2011.08.011.

Roth, Lukas (2021). "Development of drone-based phenotyping methodologies to support physiological plant breeding of wheat and soybean". Publisher: ETH Zurich. PhD thesis. ETH Zürich.

Roth, Lukas, Helge Aasen, Achim Walter, and Frank Liebisch (2018). "Extracting leaf area index using viewing geometry effects—A new perspective on high-resolution unmanned aerial system photography". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 141, pp. 161–175. DOI: 10.1016/j.isprsjprs.2018.04.012.

Roth, Lukas, Martina Binder, Norbert Kirchgessner, Flavian Tschurr, Steven Yates, Andreas Hund, Lukas Kronenberg, and Achim Walter (2024). "From Neglecting to Including Cultivar-Specific Per Se Temperature Responses: Extending the Concept of Thermal Time in Field Crops". In: *Plant Phenomics* 6, p. 0185. DOI: 10.34133/plantphenomics.0185.

Roth, Lukas, Moritz Camenzind, Helge Aasen, Lukas Kronenberg, Christoph Barendregt, Karl-Heinz Camp, Achim Walter, Norbert Kirchgessner, and Andreas Hund (2020). "Repeated multiview imaging for estimating seedling tiller counts of wheat genotypes using drones". In: *Plant Phenomics* 2020, pp. 2020/3729715. DOI: 10.34133/2020/3729715.

Roth, Lukas, María Xosé Rodríguez-Álvarez, Fred van Eeuwijk, Hans-Peter Piepho, and Andreas Hund (2021). "Phenomics data processing: A plot-level model for repeated measurements to extract the timing of key stages and quantities at defined time points". In: *Field Crops Research* 274, p. 108314. DOI: 10.1016/j.fcr.2021.108314.

Rotili, Diego Hernan, Peter De Voil, Joseph Eyre, Loretta Serafin, Darren Aisthorpe, Gustavo Angel Maddonni, and Daniel Rodríguez (2020). "Untangling genotype x management interactions in multi-environment on-farm experimentation". In: *Field Crops Research* 255, p. 107900. DOI: 10.1016/j.fcr.2020.107900.

Rouse, John Wilson, Rüdiger H. Haas, John A. Schell, and Donald W. Deering (1974). "Monitoring vegetation systems in the Great Plains with ERTS". In: *NASA Spec. Publ* 351.1, p. 309.

Rubio, E., V. Caselles, and C. Badenas (1997). "Emissivity Measurements of Several Soils and Vegetation Types in the 8-14/ m Wave Band: Analysis of Two Field Methods". In: *Remote Sensing of Environment* 59.3, pp. 490–521.

Sadeghi-Tehran, Pouria, Kasra Sabermanesh, Nicolas Virlet, and Malcolm J. Hawkesford (2017). "Automated Method to Determine Two Critical Growth Stages of Wheat: Heading and Flowering". In: *Frontiers in Plant Science* 8.February, pp. 1–14. DOI: 10.3389/fpls.2017.00252.

Schaepman-Strub, G., M.E. Schaepman, T.H. Painter, S. Dangel, and J.V. Martonchik (2006). "Reflectance quantities in optical remote sensing—definitions and case studies". In: *Remote Sensing of Environment* 103.1, pp. 27–42. DOI: 10.1016/j.rse.2006.03.002.

Schauberger, Bernhard, Tamara Ben-Ari, David Makowski, Tomomichi Kato, Hiromi Kato, and Philippe Ciais (2018). "Yield trends, variability and stagnation analysis of major crops in France over more than a century". In: *Scientific Reports* 8.1, p. 16865. DOI: 10.1038/s41598-018-35351-1.

Schils, René et al. (2018). "Cereal yield gaps across Europe". In: *European Journal of Agronomy* 101, pp. 109–120. DOI: 10.1016/j.eja.2018.09.003.

Schlang, Norbert, Ulrike Steiner, Heinz Wilhelm Dehne, Jiro Murakami, Etienne Duveiller, and Erich Christian Oerke (2008). "Spatial distribution of fusarium head blight pathogens and associated mycotoxins in wheat fields". In: *Cereal Research Communications* 36.SUPPL. 6, pp. 573–577. DOI: 10.1556/CRC.36.2008.Suppl.B.47.

Schläpfer, D., R. Richter, C. Popp, and P. Nygren (2022). "Droacor ® - Thermal: Automated temperature / emissivity retrieval for drone based hyperspectral imaging data". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B3-2022, pp. 429–434. DOI: 10.5194/isprs-archives-XLIII-B3-2022-429-2022.

Schwarz, Gideon (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6.2, pp. 461–464.

Segal, D. (1982). "Theoretical basis for differentiation of ferric-iron bearing minerals, using Landsat MSS data". In: *Proceedings of symposium for remote sensing of environment, 2nd Thematic Conference on Remote Sensing for Exploratory Geology, Fort Worth, TX.* Vol. 949, p. 951.

Segarra, Joel, Fatima Zahra Rezzouk, Nieves Aparicio, Jon González-Torralba, Iker Aranjuelo, Adrian Gracia-Romero, Jose Luis Araus, and Shawn C. Kefauver (2023). "Multiscale assessment of ground, aerial and satellite spectral data for monitoring wheat grain nitrogen content". In: *Information Processing in Agriculture* 10.4, pp. 504–522. DOI: 10.1016/j.inpa.2022.05.004.

Sellaro, Romina, María Crepy, Santiago Ariel Trupkin, Elizabeth Karayekov, Ana Sabrina Buchovsky, Constanza Rossi, and Jorge José Casal (2010). "Cryptochrome as a Sensor of the Blue/Green Ratio of Natural Radiation in Arabidopsis". In: *Plant Physiology* 154.1, pp. 401–409. DOI: 10.1104/pp.110.160820.

Shiferaw, Bekele, Melinda Smale, Hans Joachim Braun, Etienne Duveiller, Mathew Reynolds, and Geoffrey Muricho (2013). "Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security". In: *Food Security* 5.3, pp. 291–317. DOI: 10.1007/s12571-013-0263-y.

Sobrino, J., J. Jimenez-Munoz, and W. Verhoef (2005). "Canopy directional emissivity: Comparison between models". In: *Remote Sensing of Environment* 99.3, pp. 304–314. DOI: 10.1016/j.rse.2005.09.005.

Sonnentag, Oliver, Koen Hufkens, Cory Teshera-Sterne, Adam M. Young, Mark Friedl, Bobby H. Braswell, Thomas Milliman, John O'Keefe, and Andrew D. Richardson (2012). "Digital repeat photography for phenological research in forest ecosystems". In: *Agricultural and Forest Meteorology* 152, pp. 159–177. DOI: 10.1016/j.agrformet.2011.09.009.

Stoica, P. and Y. Selen (2004). "Model-order selection". In: *IEEE Signal Processing Magazine* 21.4, pp. 36–47. DOI: 10.1109/MSP.2004.1311138.

Strasser, R. J., A. Srivastava, and M. Tsimilli-Michael (2000). "Analysis of the Fluorescence Transiet as a tool to characterize and screen photosynthetic samples." In: *Chlorophyll a fluorescence: a signature of photosynthesis*, pp. 443–480.

Strebel, Silvan, Lilia Levy Häner, Michaël Mattin, Noémie Schaad, Romina Morisoli, Malgorzata Watroba, Marion Girard, Numa Courvoisier, Julien Berberat, and Raphaël Grandgirard (2022). *Liste der empfohlenen Getreidesorten für die Ernte 2023*.

Strebel, Silvan, Lilia Levy Häner, Malgorzata Watroba, Marion Girard, Anne-Valentine de Jong, Vincent Jaunin, Raphaël Grandgirard, and Nicolas Linder (2024). *Liste der empfohlenen Getreidesorten für die Ernte 2025*. de.

Su, Wen Hao, Jiajing Zhang, Ce Yang, Rae Page, Tamas Szinyei, Cory D. Hirsch, and Brian J. Steffenson (2021). "Automatic evaluation of wheat resistance to fusarium head blight using dual mask-rcnn deep learning frameworks in computer vision". In: *Remote Sensing* 13.1, pp. 1–20. DOI: 10.3390/rs13010026.

Sugita, Michiaki, Tetsuya Hiyama, and Tomohiko Ikukawa (1996). "Determination of canopy emissivity: how reliable is it?" In: *Agricultural and Forest Meteorology* 81.3-4, pp. 229–239. DOI: 10.1016/0168-1923(95)02313-5.

Sun, Dawei, Kelly Robbins, Nicolas Morales, Qingyao Shu, and Haiyan Cen (2022). "Advances in optical phenotyping of cereal crops". In: *Trends in Plant Science* 27.2, pp. 191–208. DOI: 10.1016/j.tplants.2021.07.015.

Sunic, Katarina, Lidija Brkljacic, Rosemary Vukovic, Zorana Katanic, Branka Salopek-Sondi, and Valentina Spanic (2023). "Fusarium Head Blight Infection Induced Responses of Six Winter Wheat Varieties in Ascorbate–Glutathione Pathway, Photosynthetic Efficiency and Stress Hormones". In: *Plants* 12.21, p. 3720. DOI: 10.3390/plants12213720.

Swiss Federal Council (2013). *Verordnung über die Direktzahlungen an die Landwirtschaft (Direktzahlungsverordnung, dzv)*.

Tafesse, Endale Geta, Thomas D. Warkentin, Steve Shirtliffe, Scott Noble, and Rosalind Bueckert (2022). "Leaf Pigments, Surface Wax and Spectral Vegetation Indices for Heat Stress Resistance in Pea". In: *Agronomy* 12.3, p. 739. DOI: 10.3390/agronomy12030739.

Tang, Zhehan, Yufang Jin, Maria Mar Alsina, Andrew J. McElrone, Nicolas Bambach, and William P. Kustas (2022). "Vine water status mapping with multispectral UAV imagery and machine learning". In: *Irrigation Science* 40.4-5, pp. 715–730. DOI: 10.1007/s00271-022-00788-w.

Taylor, Shawn D. and Dawn M. Browning (2021). "Classification of daily crop phenology in PhenoCams using deep learning and hidden markov models". In: *Remote Sensing*.

Tester, Mark and Peter Langridge (2010). "Breeding Technologies to Increase Crop Production in a Changing World". In: *Science* 327.5967, pp. 818–822. DOI: 10.1126/science.1183700.

Trail, Frances (2009). "For Blighted Waves of Grain: *Fusarium graminearum* in the Postgenomics Era". In: *Plant Physiology* 149.1, pp. 103–110. DOI: 10.1104/pp.108.129684.

Treier, Simon, Juan M. Herrera, Andreas Hund, Norbert Kirchgessner, Helge Aasen, Achim Walter, and Lukas Roth (2024). "Improving drone-based uncalibrated estimates of wheat canopy temperature in plot experiments by accounting for confounding factors in a multi-view analysis". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 218, pp. 721–741. DOI: 10.1016/j.isprsjprs.2024.09.015.

Tschurr, Flavian, Lukas Roth, Nicola Storni, Olivia Zumsteg, Achim Walter, and Jonas Anderegg (2024). "Temporal resolution trumps spectral resolution in UAV-based monitoring of cereal senescence dynamics". In: *Plant Methods* 20.1, p. 188. DOI: 10.1186/s13007-024-01308-x.

Tucker, Compton J. (1979). "Red and photographic infrared linear combinations for monitoring vegetation". In: *Remote Sensing of Environment* 8.2, pp. 127–150. DOI: https://doi.org/10.1016/0034-4257(79)90013-0.

Valcke, R. (2021). "Can chlorophyll fluorescence imaging make the invisible visible?" In: *Photosynthetica* 59.SPECIAL ISSUE, pp. 381–398. DOI: 10.32615/ps.2021.017.

Van Dijk, Michiel, Tom Morley, Marie Luise Rau, and Yashar Saghai (2021). "A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050". In: *Nature Food* 2.7, pp. 494–501. DOI: 10.1038/s43016-021-00322-9.

van Rossum, Guido and Drake, Fred L. (2009). *Python 3 Reference Manual*. Place: Scotts Valley, CA.

Velazco, Julio G., María Xosé Rodríguez-Álvarez, Martin P. Boer, David R. Jordan, Paul H. C. Eilers, Marcos Malosetti, and Fred A. Van Eeuwijk (2017). "Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model". In: *Theoretical and Applied Genetics* 130.7, pp. 1375–1392. DOI: 10.1007/s00122-017-2894-4.

Velumani, Kaaviya, Simon Madec, Benoit De Solan, Raul Lopez-Lozano, Jocelyn Gillet, Jeremy Labrosse, Stephane Jezequel, Alexis Comar, and Frédéric Baret (2020). "An automatic method based on daily in situ images and deep learning to date wheat heading stage". In: *Field Crops Research* 252, p. 107793. DOI: 10.1016/j.fcr.2020.107793.

Vincke, Damien, Damien Eylenbosch, Guillaume Jacquemin, Anne Chandelier, Juan Antonio Fernández Pierna, François Stevens, Vincent Baeten, Benoît Mercatoris, and Philippe Vermeulen (2023). "Near infrared hyperspectral imaging method to assess Fusarium Head Blight infection on winter wheat ears". In: *Microchemical Journal* 191, p. 108812. DOI: 10.1016/j.microc.2023.108812.

Voss-Fels, Kai P., Andreas Stahl, Benjamin Wittkop, Carolin Lichthardt, Sabrina Nagler, Till Rose, Tsu Wei Chen, Holger Zetzsche, Sylvia Seddig, Mirza Majid Baig, Agim Ballvora, Matthias Frisch, Elizabeth Ross, Ben J. Hayes, Matthew J. Hayden, Frank Ordon, Jens Leon, Henning Kage, Wolfgang Friedt, Hartmut Stützel, and Rod J. Snowdon (2019). "Breeding improves wheat productivity under contrasting agrochemical input levels". In: *Nature Plants* 5.7, pp. 706–714. DOI: 10.1038/s41477-019-0445-5.

Walter, Achim, Frank Liebisch, and Andreas Hund (2015). "Plant phenotyping: from bean weighing to image analysis". In: *Plant Methods* 11.14. ISBN: 1746-4811. DOI: 10.1186/s13007-015-0056-8.

Wang, Dunliang, Rui Li, Bo Zhu, Tao Liu, Chengming Sun, and Wenshan Guo (2022). "Estimation of Wheat Plant Height and Biomass by Combining UAV Imagery and Elevation Data". In: *Agriculture* 13.1, p. 9. DOI: 10.3390/agriculture13010009.

Wang, Falv, Mao Yang, Longfei Ma, Tong Zhang, Weilong Qin, Wei Li, Yinghua Zhang, Zhencai Sun, Zhimin Wang, Fei Li, and Kang Yu (2022). "Estimation of Above-Ground Biomass of Winter Wheat Based on Consumer-Grade Multi-Spectral UAV". In: *Remote Sensing* 14.5, p. 1251. DOI: 10.3390/rs14051251.

Wang, Hesong, Gensuo Jia, Howard E. Epstein, Huichen Zhao, and Anzhi Zhang (2020). "Integrating a PhenoCam-derived vegetation index into a light use efficiency model to estimate daily gross primary production in a semi-arid grassland". In: *Agricultural and Forest Meteorology* 288-289.March. DOI: 10.1016/j.agrformet.2020.107983.

Wang, Ziwei, Ji Zhou, Jin Ma, Yong Wang, Shaomin Liu, Lirong Ding, Wenbin Tang, Nuradili Pakezhamu, and Lingxuan Meng (2023). "Removing temperature drift and temporal variation in thermal infrared images of a UAV uncooled thermal infrared imager". In:

ISPRS Journal of Photogrammetry and Remote Sensing 203, pp. 392–411. DOI: 10.1016/j.isprsjprs.2023.08.011.

WBF (2021). *Verordnung des WBF über Vermehrungsmaterial von Ackerpflanzen-, Futterpflanzen- und Gemüsearten*. Tech. rep. Das Eidgenössische Departement für Wirtschaft, Bildung und Forschung (WBF).

Welbank, P. J., K. J. Witts, and Gillian N. Thorne (1968). "Effect of Radiation and Temperature on Efficiency of Cereal Leaves during Grain Growth". In: *Annals of Botany* 32.1, pp. 79–95. DOI: 10.1093/oxfordjournals.aob.a084201.

Wilhelm, W. W. and Gregory S. McMaster (1996). "Spikelet and Floret Naming Scheme for Grasses with a Spike Inflorescence". In: *Crop Science* 36.4, pp. 1071–1073. DOI: 10.2135/cropsci1996.0011183X0036000400044x.

Wilson, W., B. Dahl, and W. Nganje (2018). "Economic costs of Fusarium Head Blight, scab and deoxynivalenol". In: *World Mycotoxin Journal* 11.2, pp. 291–302. DOI: 10.3920/WMJ2017.2204.

Windels, C. E. (2000). "Economic and social impacts of Fusarium head blight: Changing farms and rural communities in the Northern Great Plains". In: *Phytopathology* 90.1, pp. 17–21. DOI: 10.1094/PHYTO.2000.90.1.17.

Woebbecke, D. M., G. E. Meyer, K. Von Bargen, and D. A. Mortensen (1995). "Color Indices for Weed Identification Under Various Soil, Residue, and Lighting Conditions". In: *Transactions of the ASAE* 38.1. DOI: 10.13031/2013.27838.

Wu, Lang (2010). *Mixed effects models for complex data*. eng. Monographs on statistics and applied probability ; 113. Boca Raton: Chapman & Hall/CRC Press. ISBN: 0-429-14251-X.

Xu, Rui and Changying Li (2022). "A Review of High-Throughput Field Phenotyping Systems: Focusing on Ground Robots". In: *Plant Phenomics* 2022, p. 9760269. DOI: 10.34133/2022/9760269.

Yan, Sheng-nan, Zhao-yu Yu, Wei Gao, Xu-yang Wang, Jia-jia Cao, Jie Lu, Chuan-xi Ma, Cheng Chang, and Hai-ping Zhang (2023). "Dissecting the genetic basis of grain color and pre-harvest sprouting resistance in common wheat by association analysis". In: *Journal of Integrative Agriculture* 22.9, pp. 2617–2631. DOI: 10.1016/j.jia.2023.04.017.

Yang, Chin Jian, Joanne Russell, Ian Mackay, and Wayne Powell (2024). "Opportunities to Improve the Recommendation of Plant Varieties under the Recommended List (RL) System". In: *Agronomy* 14.10, p. 2267. DOI: 10.3390/agronomy14102267.

Yang, Feng, Zhongqiang Liu, Yuxi Wang, Xiaofeng Wang, Qiusi Zhang, Yanyun Han, Xiangyu Zhao, Shouhui Pan, Shuo Yang, Shufeng Wang, Qi Zhang, Jun Qiu, and Kaiyi Wang (2023). "A variety test platform for the standardization and data quality improvement of crop variety tests". In: *Frontiers in Plant Science* 14, p. 1077196. DOI: 10.3389/fpls.2023.1077196.

Yang, Hong, Baodi Dong, Yakai Wang, Yunzhou Qiao, Changhai Shi, Lele Jin, and Mengyu Liu (2020). "Photosynthetic base of reduced grain yield by shading stress during the early reproductive stage of two wheat cultivars". In: *Scientific Reports* 10.1, p. 14353. DOI: 10.1038/s41598-020-71268-4.

Yang, Mingzhe, Zhipeng Wang, Kaiwei Liu, Yingqi Rong, Bing Yuan, and Jiang Zhang (2024). "Finding emergence in data by maximizing effective information". In: *National Science Review* 12.1, nwae279. DOI: 10.1093/nsr/nwae279.

Yuan, Wenan and Weiyun Hua (2022). "A case study of vignetting nonuniformity in UAV-based uncooled thermal cameras". In: *Drones* 6.12, p. 394. DOI: 10.3390/drones6120394.

Yue, Jibo, Guijun Yang, Qingjiu Tian, Haikuan Feng, Kaijian Xu, and Chengquan Zhou (2019). "Estimate of winter-wheat above-ground biomass based on UAV ultrahigh-ground-resolution image textures and vegetation indices". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 150.February. Publisher: Elsevier, pp. 226–244. DOI: 10.1016/j.isprsjprs.2019.02.022.

Zarco-Tejada, P.J., V. González-Dugo, and J.A.J. Berni (2012). "Fluorescence, temperature and narrow-band indices acquired from a UAV platform for water stress detection using a micro-hyperspectral imager and a thermal camera". In: *Remote Sensing of Environment* 117, pp. 322–337. DOI: 10.1016/j.rse.2011.10.007.

Zarco-Tejada, P.J., V. González-Dugo, L.E. Williams, L. Suárez, J.A.J. Berni, D. Goldhamer, and E. Fereres (2013). "A PRI-based water stress index combining structural and chlorophyll effects: Assessment using diurnal narrow-band airborne imagery and the CWSI thermal index". In: *Remote Sensing of Environment* 138, pp. 38–50. DOI: 10.1016/j.rse.2013.07.024.

Zetzsche, Holger, Wolfgang Friedt, and Frank Ordon (2020). "Breeding progress for pathogen resistance is a second major driver for yield increase in German winter wheat at contrasting N levels". In: *Scientific Reports* 10.1, p. 20374. DOI: 10.1038/s41598-020-77200-0.

Zhang, Chu, Lei Zhou, Qinlin Xiao, Xiulin Bai, Baohua Wu, Na Wu, Yiying Zhao, Junmin Wang, and Lei Feng (2022). "End-to-End Fusion of Hyperspectral and Chlorophyll Fluorescence Imaging to Identify Rice Stresses". In: *Plant Phenomics* 2022, p. 9851096. DOI: 10.34133/2022/9851096.

Zhang, Hansu, Linsheng Huang, Wenjiang Huang, Yingying Dong, Shizhuang Weng, Jinling Zhao, Huiqin Ma, and Linyi Liu (2022). "Detection of wheat Fusarium head blight using UAV-based spectral and image feature fusion". In: *Frontiers in Plant Science* 13, p. 1004427. DOI: 10.3389/fpls.2022.1004427.

Zhang, Jiayi, Xiaolei Qiu, Yueting Wu, Yan Zhu, Qiang Cao, Xiaojun Liu, and Weixing Cao (2021). "Combining texture, color, and vegetation indices from fixed-wing UAS imagery to estimate wheat growth parameters using multivariate regression methods". In: *Computers and Electronics in Agriculture* 185, p. 106138. DOI: 10.1016/j.compag.2021.106138.

Zheng, Xiaopo, Zhao-Liang Li, Xia Zhang, and Guofei Shang (2019). "Quantification of the adjacency effect on measurements in the thermal infrared region". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.12, pp. 9674–9687. DOI: 10.1109/TGRS.2019.2928525.

Zhou, Yong, Hao Tang, Meng-Ping Cheng, Kwame O. Dankwa, Zhong-Xu Chen, Zhan-Yi Li, Shang Gao, Ya-Xi Liu, Qian-Tao Jiang, Xiu-Jin Lan, Zhi-En Pu, Yu-Ming Wei, You-Liang Zheng, Lee T. Hickey, and Ji-Rui Wang (2017). "Genome-Wide Association Study for Pre-harvest Sprouting Resistance in a Large Germplasm Collection of Chinese Wheat Landraces". In: *Frontiers in Plant Science* 08. DOI: 10.3389/fpls.2017.00401.

Zhu, Yanjun, Zhiguo Cao, Hao Lu, Yanan Li, and Yang Xiao (2016). "In-field automatic observation of wheat heading stage using computer vision". In: *Biosystems Engineering* 143. Publisher: Elsevier Ltd, pp. 28–41. DOI: 10.1016/j.biosystemseng.2015.12.015.

Zorn, Alexander, Tomke Musa, and Markus Lips (2018). "Was kostet die Vermeidung des Pilzgifts Deoxynivalenol im Weizenanbau?" de. In: *Agrarforschung Schweiz*.

S1 Supplementary Materials - Improving drone-based uncalibrated estimates of wheat canopy temperature in plot experiments by accounting for confounding factors in a multi-view analysis

S1.1 Experimental design of EuVar

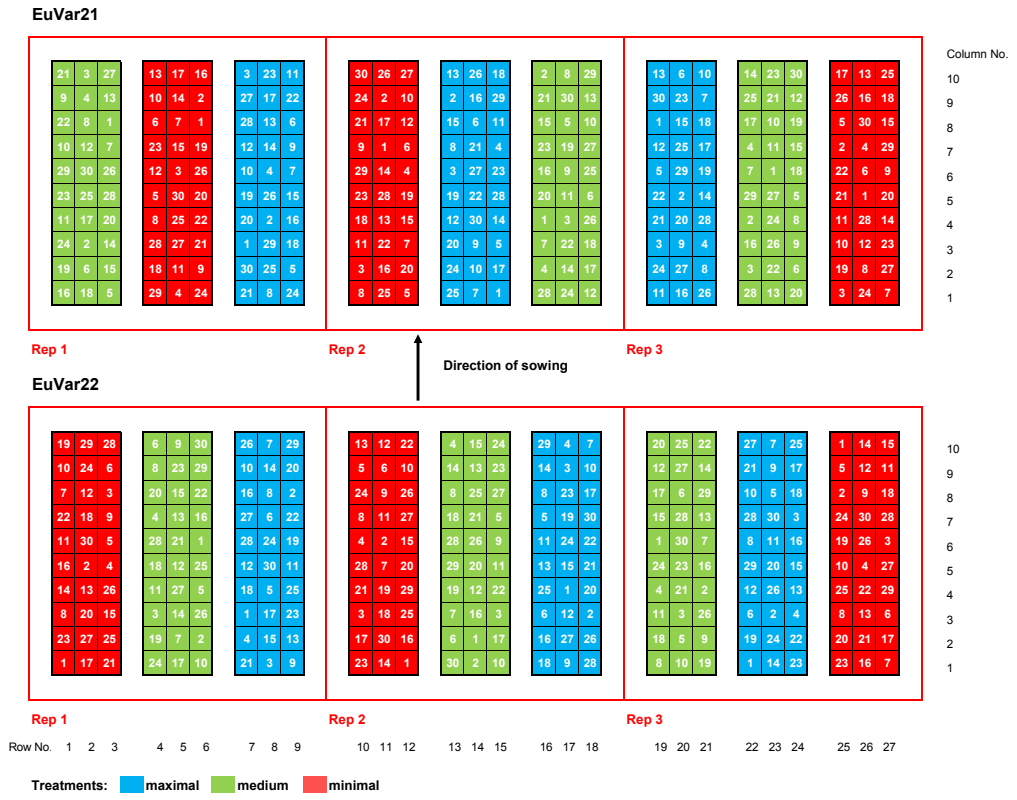


FIGURE S1.1: The experimental design of EuVar for the two years. The numbers inside the blocks indicate the genotypes.

S1.2 Details on Field treatments

Table S1.1: Overview of trial treatments and most important field interventions for all trials. "too wet" indicates that treatments were intended but could not be applied as conditions were too wet and heavy machinery could not enter the field.

				Herbicides					Fertilization [kg/ha]			
Experiment	**Treatment**	**Sowing date**	**Harvest date**	Monocot. 1st	Monocot. 2nd	Dicot.	**Growth regulator**	**Fungicide**	N	CaO	MgO	SO_3
EuVar21	Minimal	2020-10-22	2021-07-20	Archipel®	too wet	too wet	-	-	140	15	9	-
	Medium						Moddus®	-				
	Maximal						Moddus®	Amistar®				
EuVar22	Minimal	2021-10-15	2022-06-30	Archipel®	Othello Star®	Cleave®/ Express Max®	-	-	140	23	36	30
	Medium						Moddus®	-				
	Maximal						Moddus®	Amistar®				

Table S1.2: Chemical compositions of field treatments and quantities applied.

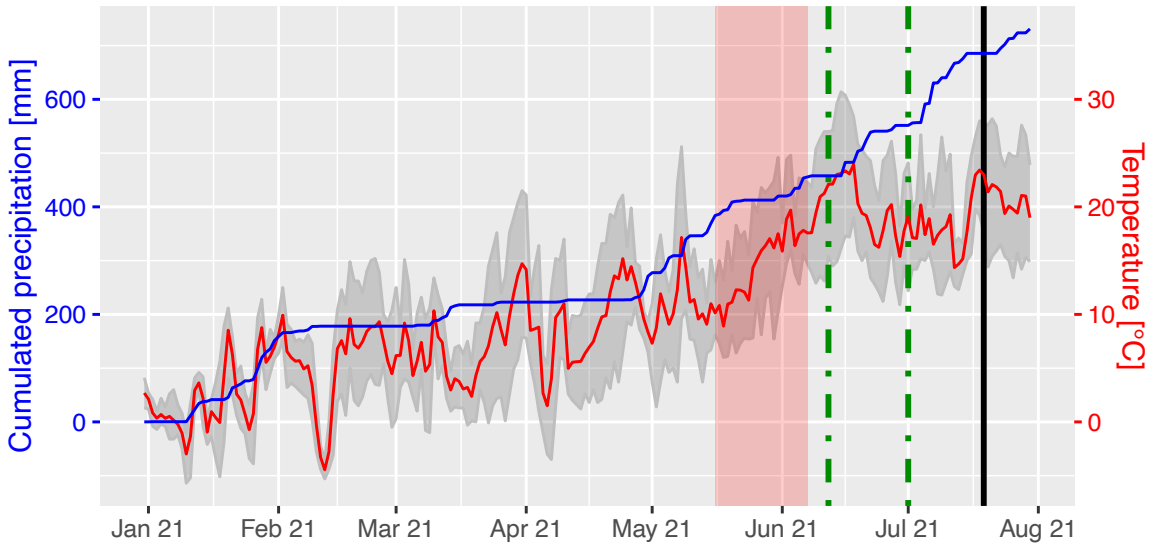
Procuct	**Active ingredient(s)**	**Application rate [g/ha]**	**Producer**
Archipel®	Iodosulfuron-methyl-sodium	9	Syngenta
	Mesosulfuron-methyl	9	
	Mefenpyr-diethyl	27	
Moddus®	Trinexapac-ethyl	125	Syngenta
Amistar®	Azoxystrobin	200	Syngenta
	Cyproconazole	80	
Othello Star®	Iodosulfuron-methyl-sodium	9	Bayer
	Mesosulfuron-methyl	9	
	Mefenpyr-diethyl	27	
	Thiencarbazone-methyl	7.5	
Cleave®	Fluroxypyr	90	Syngenta
	Fluroxypyr-meptyl	130	
	Florasulam	23	
Express Max®	Metsulfuron-methyl	5	Syngenta
	Tribenuron-methyl	5	

S1.3 Overview on flights

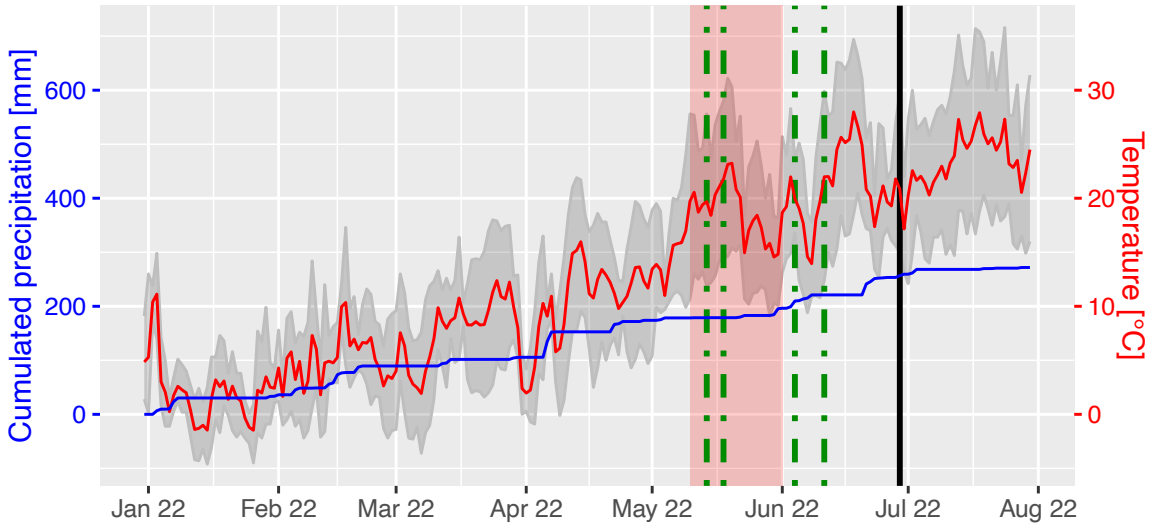
TABLE S1.3: Overview on number of thermal measurements per date.

Year	Date	No. of flights
2021	2021-06-12	13
	2021-07-01	9
2022	2022-05-14	3
	2022-05-18	3
	2022-06-04	6
	2022-06-11	5

S1.4 Environmental conditions



(a) Weather overview of growing seasons 2021



(b) Weather overview of growing seasons 2022

FIGURE S1.2: (a) and (b) show the general weather conditions during the growing seasons of the years 2021 and 2022 from January until after harvest. The red line shows the mean air temperature measured at 2 m above the ground and the shades indicate daily air temperature minima and maxima. Green dashed lines represent the measurement dates. Cumulative precipitation is shown as rising blue line. During the period shaded in red, heading was observed in the field and harvest dates are marked by black lines.

S1.5 Flight campaigns

Optimal conditions for TIR imaging surveys are a clear blue sky, hot temperatures and little or no wind (Perich et al., 2020). In 2021, weather conditions were often sub-optimal for flying (rain and cloud cover throughout the growing season) (Fig. S1.2a). On rare days with suitable

weather, as many flights as possible were done on one experiment before, during and after solar noon.

2022 was a very hot and dry year (Fig. S1.2b). The phenological window suitable for flying was short but within this window, conditions were often suitable for TIR imaging. Consequently, in season 2022, flights were conducted on more days but on individual days, with less flights per day than in 2021. Based on results of 2021, flights were restricted to times after 12:00 in 2022. Fig. S1.2 provides an overview on when TIR measurements were conducted within the growing season while S1.9 and S1.10 show, on what time of day the flights were conducted during the single days. In total, 39 flights were conducted (Table S1.3).

S1.6 Camera settings and flight planning

Flights were conducted with a DJI Matrice 200 drone (SZ DJI Technology Co. Ltd., China) that carried a DJI Zenmuse XT TIR sensor equipped with a 9 mm, f/1.4 lens. Pixel resolution of 640 x 512 was achieved by individual uncooled VOx microbolometers arranged in a focal plane array. The field of view (FOV) was 69° x 56°. Temperature was measured in the wavelength range from 7.5− 13.5 µm and thermal sensitivity was < 50 mK. In high gain mode, the camera could measure temperature in the range from -25 °C to +135 °C. The absolute measurement accuracy was ± 10 °C according to manufacturer. The external parameters were set in DJI Pilot software (SZ DJI Technology Co. Ltd., China) to the same values for all flights using DJI default settings. Scene emissivity was set to 100 %, background and air temperature were set to 22 °C. The sensor provided the option of periodical flat field correction throughout measurement to reduce the noise of non-uniformity effects. This periodic compensation would reduce comprehensibility of drift effects. In addition, Kelly et al. (2019) showed that non-uniformity correction alone was not sufficient to correct for drift effect during flights. Flat field correction was therefore deactivated following Mesas-Carrascosa et al. (2018).

Flights were planned with DJI Pilot software. The drone flew over the plots at a height of approximately 40 m on a path that was defined as a way-point mission. This allowed for a ground sampling distance (GSD) of about 5.2 cm/pixel. With a plot width of 1.5 m, this GSD allowed to have more than 20 rows of pixels within plots after excluding border areas of the plots while still allowing for relatively short flights. Exposure interval was 2 s. These settings resulted in an image pattern where each spot in the trial was recorded at least on 9 images from different perspectives. The camera was pointing toward the ground orthogonally (*i.e.* in nadir orientation).

While mission planing often is done by defining a minimal front- and side overlap, this was not possible due to software restrictions for the drone−sensor combination used. In addition, using way-point flight planning with a fixed exposure interval allowed for a manually defined camera heading throughout the flight. Therefore, the heading of drone and TIR camera remained relatively stable throughout the flight and did not change with flight path direction changes. Flight speed was limited at $4\,\mathrm{m\,s^{-1}}$.

S1.7 Flight operation

In 2021, the camera was turned on at least 15 min before each flight to allow the temperature signal to stabilize. In 2022, an additional set of batteries was used and the stabilization period was increased to 30 min. In situations, where the battery was not sufficient anymore to complete all flights, the temperature stabilization was not repeated after a rapid battery change. After the first flight campaigns in 2021, a rather strong drift of apparent temperature was noticed that seemed to be particularly strong during the beginning of flights. To reduce

initial drift, the drone was further hovered above the wheat field for about one minute in addition to previous temperature stabilization on the ground before the measurement flight sequences were started.

S1.8 Thermal ground control points

GCPs were produced following Perich et al. (2020) by gluing triangles of 2 mm thick aluminum sheets on polystyrene foam plates. These plates had an extent of 1 m x 0.5 m or 0.5 m x 0.5 m. Unlike in Perich et al. (2020), the aluminum sheets were left blank as was done in other TIR surveys (e.g. Mesas-Carrascosa et al., 2018; Aragon et al., 2020) and not painted black. This avoided large temperature gradients in the FOV and reduced possible adjacency effects of hot objects (Aragon et al., 2020; Zheng et al., 2019).

S1.9 Georeference images

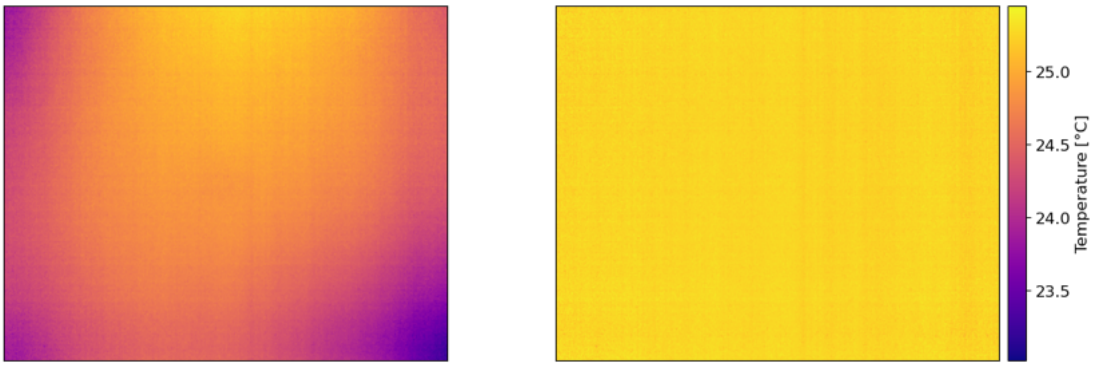
The 8-bit JPEGs of the radiometric image as well as the RGB images were aligned in the structure-from-motion-based software Agisoft Metashape Professional (Agisoft LLC, St. Petersburg, Russia). TIR images feature a low spatial resolution and are therefore difficult to georeference. No precise GPS device to measure GCP positions was available and an indirect referencing approach was used. One RGB project served as a reference project and was referenced by the positioning information from the drone available for each image in the meta data. The GCP coordinates were extracted from this project and used to reference all other projects of one year. Conventional GCPs are difficult to detect on TIR images and on one RGB project, the RGB GCPs were visible together with the thermal GCPs. The locations of the thermal GCPs were then extracted from this RGB project and used to reference the thermal projects. This allowed for a correct geographic orientation and a absolute positioning precision within 2 m horizontally and vertically according to a quality check in Qgis (QGIS Development Team, 2022). With this procedure all TIR flights were georeferenced in the Cartesian Swiss coordinate system EPSG:2056 (CH1903+LV95) which allowed to precisely superimpose the aligned images of the different flights. Relative positioning precision between flights was estimated to be 15 cm or smaller based on marker position error estimates in Agisoft.

The 8-bit JPEGs were preferred over the 14-bit TIFF images in the process of aligning images as they provide better contrast and contain meta information on TIR camera position and orientation during triggering, which allows to get a valid alignment more reliably. However, these 8-bit JPEG are just a non-linear, visually augmented interpretation of TIR with a value range of 0 to 255 and could not be used for analysis of temperature. As pixel position remained consistent between the two formats, 8-bit JPEG images were replaced by 14-bit TIFF files after alignment for further temperature analysis.

S1.10 Vignetting correction procedure

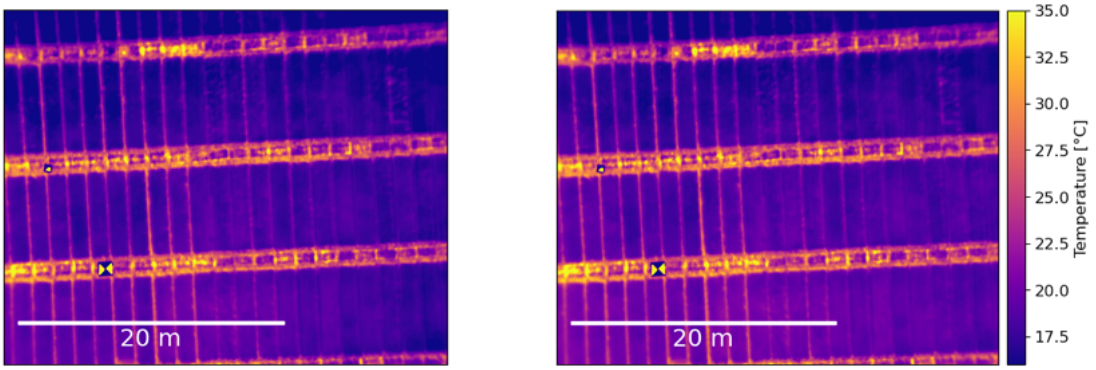
To generate the base for an overall vignetting correction, the drone was located indoors with the TIR sensor pointing at a hard foam PVC sheet. The distance between sensor and sheet was about 40 cm and the sheet fully covered the FOV of the camera. Ambient air temperature was 22 °C and there was no direct light on the PVC sheet. The ambient light in the room was reduced to mitigate artifacts from light. The PVS foam was put inside the room 5 hours prior to use to reach temperature equilibrium.

The camera was started to stabilize. After 1 h, TIR images of the PCV sheet were taken at an interval of 5 s during more than 30 min. A vignetting correction image was then calculated as the pixel-wise mean of these 413 images in Python 3.8.



(a) Before correction (b) After correction

FIGURE S1.3: (a) shows an image of a homogeneous PVC sheet that was part of the set used to create an vignetting correction image. A vignetting pattern is clearly visible with a cooling trend toward the edges. (b) shows the same image after correction was applied. The vignetting was clearly mitigated and the image of the PVS sheet now appears flat with almost no trends visible.



(a) Before correction (b) After correction

FIGURE S1.4: (a & b) show an example image taken during a fligth before and after vignetting correction.

S1.11 Generalized heritability formula framework

In generalized heritability, the effective dimensions ED_g are divided by the difference between the number of genotypes m_g and the number of zero eigenvalues ζ_g:

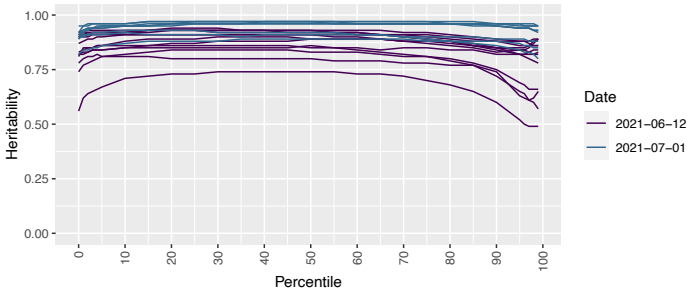
$$H^2_{genral.} = \frac{ED_g}{(m_g - \zeta_g)}, \tag{S1.1}$$

with

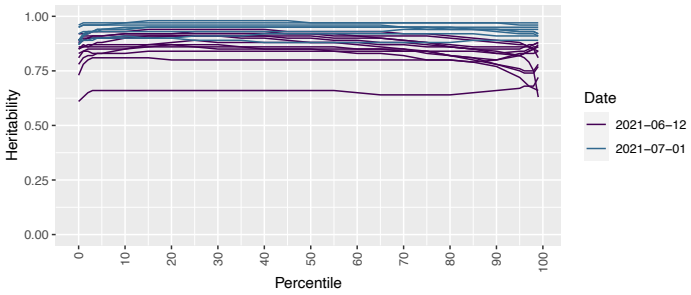
$$ED_g = (m_g - 1)\frac{\sigma^2_g}{(\sigma^2_g + \frac{\sigma^2_e}{r})}. \tag{S1.2}$$

S1.12 Blending mode selection

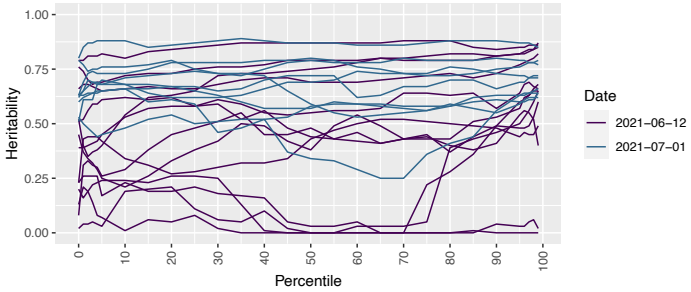
EuVar21



(a) Blending mode: Average



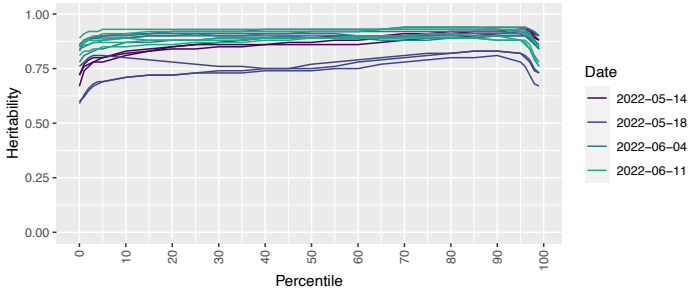
(b) Blending mode: Mosaic



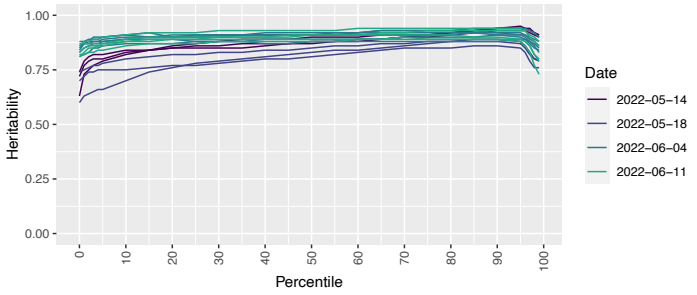
(c) Blending mode: Disabled

FIGURE S1.5: (a - c) show the heritability of the orthomosaic method for each pixel value percentile for each flight conducted on EuVar21. The orthomosaics were created with three different blending modes average (a), mosaic (b) and disabled (c).

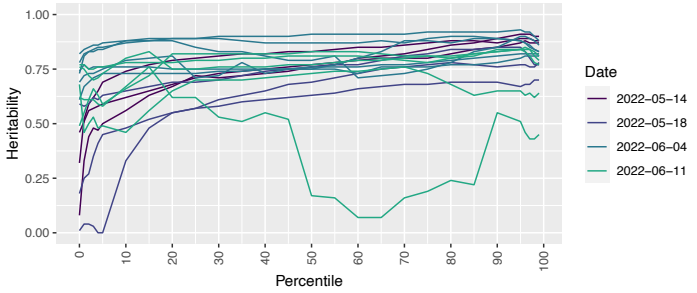
EuVar22



(a) Blending mode: Average



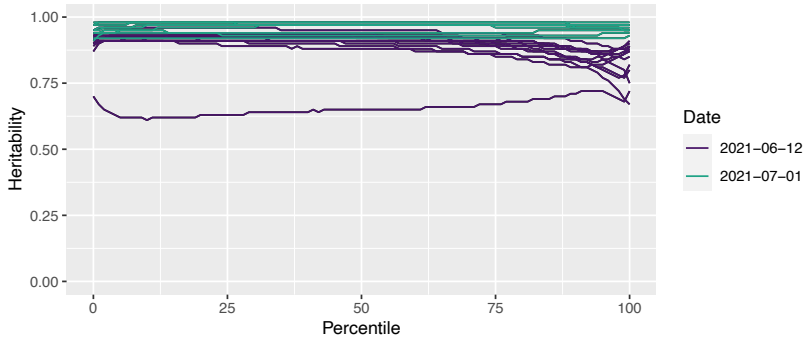
(b) Blending mode: Mosaic



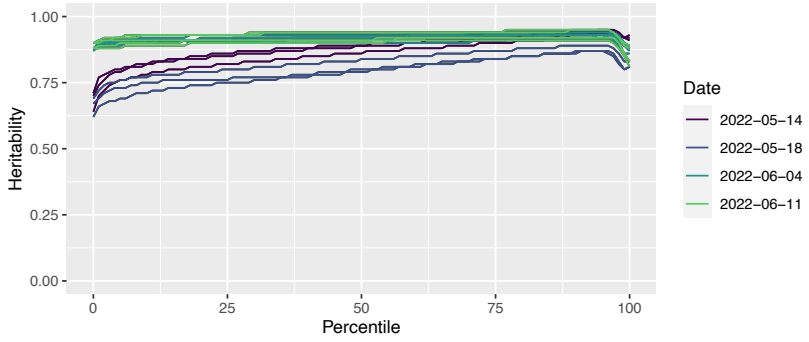
(c) Blending mode: Disabled

FIGURE S1.6: (a - c) show the heritability of the orthomosaic method for each pixel value percentile for each flight conducted on EuVar22. The orthomosaics were created with three different blending modes average (a), mosaic (b) and disabled (c).

S1.13 Multi-view percentile choice



(a) EuVar21



(b) EuVar22

FIGURE S1.7: (a & b) show the heritability of the multi-view method for each pixel value percentile for each flight conducted on EuVar.

S1.14 Correlations

EuVar21

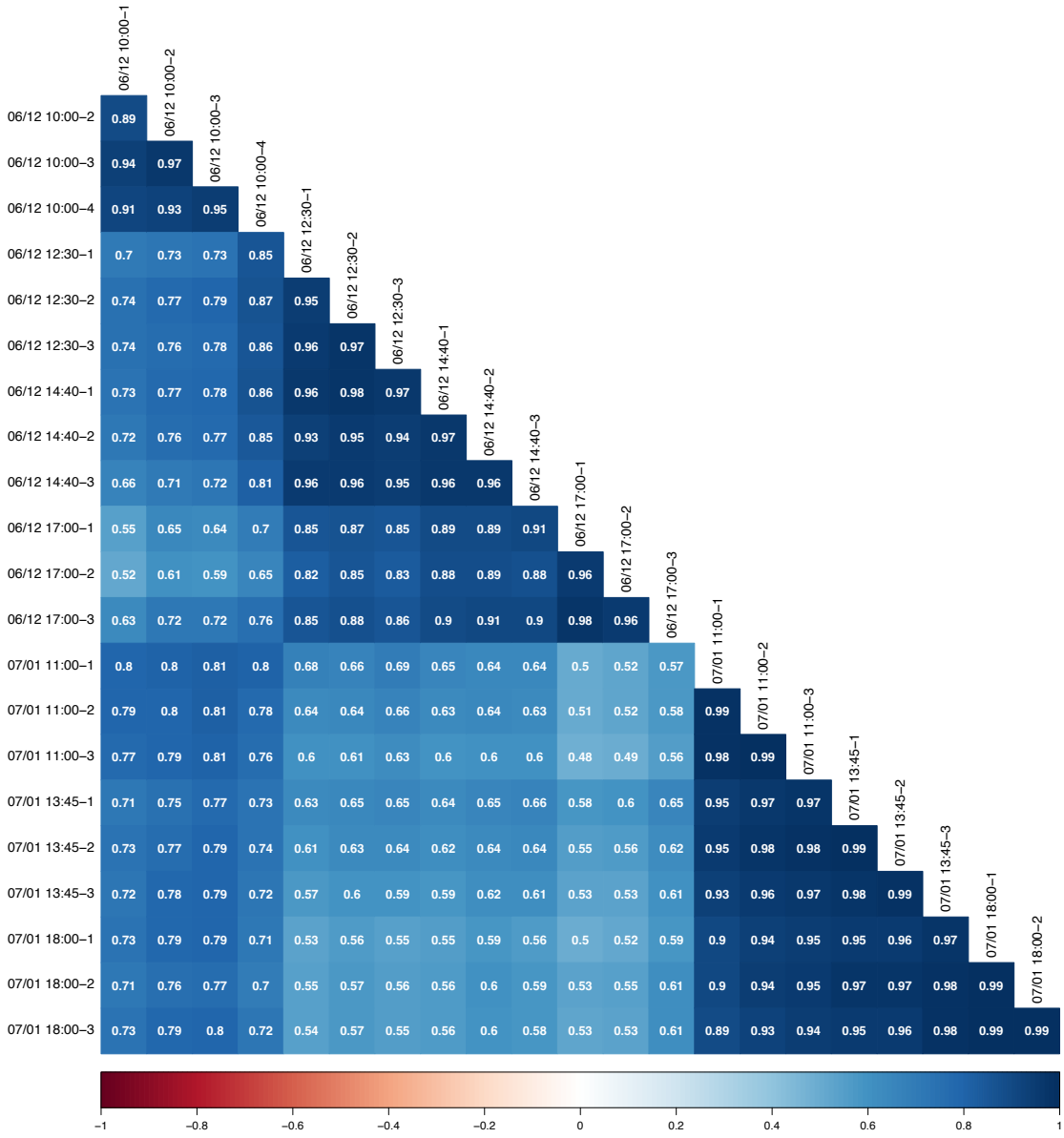


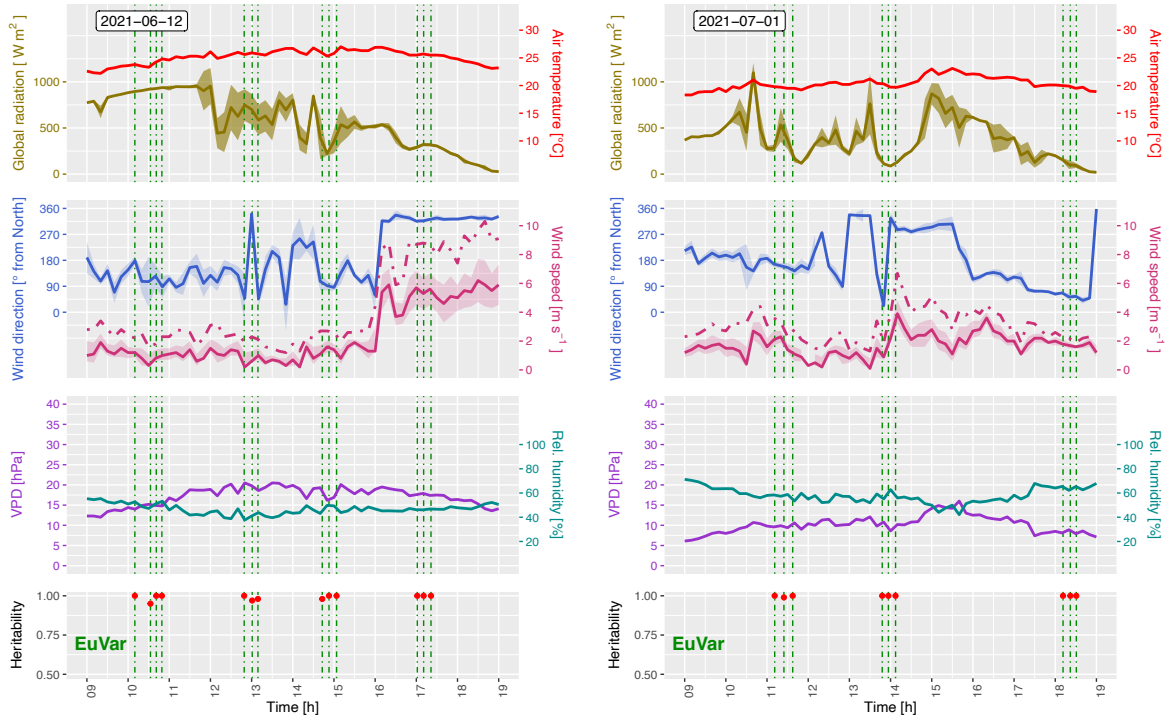
FIGURE S1.8: Pearson's correlations of plot-based CT measurements within EuVar21 were derived after plot-wise CT estimation with the most complex mixed model "MM Trigger + RowDir + SunDir + Sensor" and subsequent fitting with SpATS. All correlation are significant at $P < 0.001$.

S1.15 Campaign-wise genotype ranking consistency

TABLE S1.4: The sd of genotype ranking within campaigns σ_{gen_r} was calculated for all the methods and models. Mean and median values for each method are shown for each year before and after spatial correction in SpATS. (*The values of the method "SpATS (one-stage)" before spatial correction correspond to unadjusted mean values as for the method "Agg. - Mean".)

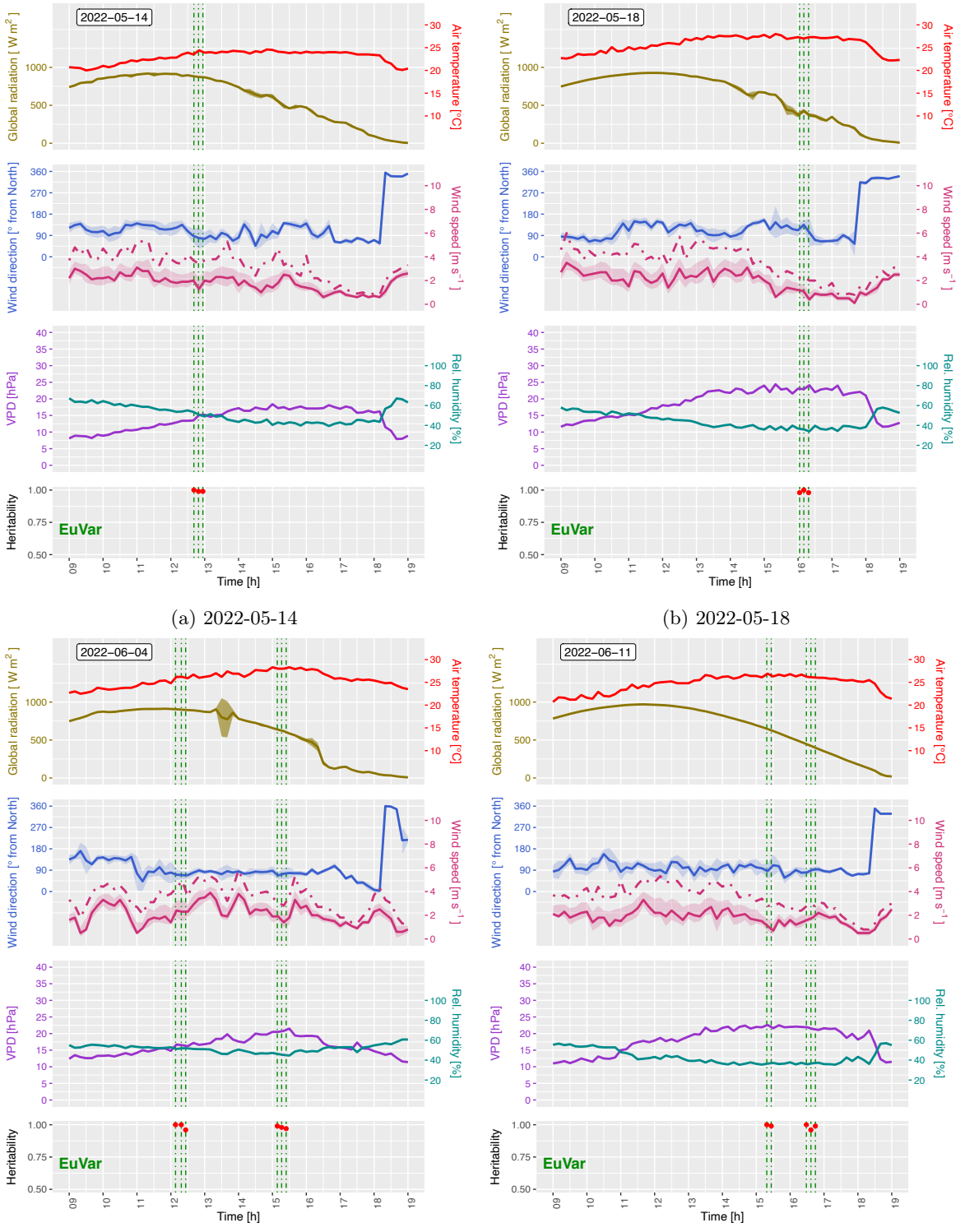
Year	SpATS	Method/Model	mean σ_{gen_r}	median σ_{gen_r}
2021	before	Ortho	6.54	6.52
		Agg. - Median	6.60	6.44
		Agg. - Mean	6.35	6.31
		LM	7.41	7.02
		SpATS (one-stage)	6.35*	6.31*
		MM Trigger	4.00	3.67
		MM Trigger+ RowDir+SunDir+ Sensor	4.06	3.81
	after	Ortho	4.61	4.44
		Agg. - Median	5.22	5.00
		Agg. - Mean	4.58	4.45
		LM	4.30	4.04
		SpATS (one-stage)	3.97	3.50
		MM Trigger	3.94	3.49
		MM Trigger+ RowDir+SunDir+ Sensor	3.97	3.60
2022	before	Ortho	7.21	7.02
		Agg. - Median	7.48	7.54
		Agg. - Mean	7.36	7.33
		LM	5.00	5.10
		SpATS (one-stage)	7.36*	7.33*
		MM Trigger	3.16	3.02
		MM Trigger+ RowDir+SunDir+ Sensor	3.12	3.06
	after	Ortho	3.46	3.40
		Agg. - Median	3.72	3.54
		Agg. - Mean	3.41	3.29
		LM	3.32	3.14
		SpATS (one-stage)	3.13	3.02
		MM Trigger	3.07	2.79
		MM Trigger+ RowDir+SunDir+ Sensor	3.02	2.72

S1.16 Weather data



(a) 2021-06-12 (b) 2021-07-01

FIGURE S1.9: (a - b) show the detailed weather conditions on measurement days in 2021. Solid lines show means, shades are means ± SD and dashed lines the maxima for 10 min intervals. The vertical lines indicate the different flights of EuVar and heritabilities are indicated at the bottom of the weather charts.



(a) 2022-05-14

(b) 2022-05-18

(c) 2022-06-04

(d) Weather 2022-06-11

FIGURE S1.10: (a - d) show the detailed weather conditions on measurement days in 2022. Solid lines show means, shades are means ± SD and dashed lines the maxima for 10 min intervals. The vertical lines indicate the different flights of EuVar and heritabilities are indicated at the bottom of the weather charts.

S1.17 Considerations on vignetting correction and stabilization

Vignetting correction is proposed to improve accuracy of TIR measurements and to mitigate problems during alignment of TIR images (Yuan and Hua, 2022). In this work, no influence of vignetting correction on the success of alignment was observed. Vignetting correction also did not increase the correlations of plot-wise CT estimates between flights or heritability.

The vignetting correction effect on correlation and heritability is yet not well understood. Yuan and Hua (2022) propose to use a single TIR image taken shortly after landing for applying vignetting correction individually to every flight. They state that wind direction and speed do not change vignetting and non-uniformity patterns. However they used just two opposing wind directions (left and right from the sensor) and let the signal stabilize for several minutes. As mentioned for the stabilization procedure, the temperature keeps oscillating significantly during the flights. While it was not possible to estimate vignetting and non-uniformity effects within flights, it is assumed that the strong oscillation during the flight also leads to changes in the vignetting patterns at least short-term. It was shown that changes in thermal drift happened usually within less than 1 min from each other and an equilibrium was never reached. This is limiting the potential of applying the same vignetting correction to all images across entire flights.

Different stabilization procedures are suggested to reduce non-uniformity effects and vignetting (e.g. Jimenez-Berni, P. J. Zarco-Tejada, et al., 2009; Kelly et al., 2019; Yuan and Hua, 2022), the problem is that surrounding conditions of the drone keep changing. According to our personal experience, the most elaborate initial stabilization procedures remain very limited at mitigating non-uniformity effects and vignetting once the drone took off.

For the measures considered here, the multi-view method seemed to deal with vignetting in a way that does not make ex-ante vignetting correction prerequisite and could be dropped for further TIR analysis for most experiments. Kelly et al. (2019) mention that vignetting is more problematic in single image analysis than when working with orthomosaics and seemingly also with a multi-view approach. Nevertheless, vignetting correction shows patterns very similar to CT patterns related to viewing geometry. Therefore, vignetting correction in a probabilistic manner might help to avoid an overestimation of the importance of viewing geometry related covariates in mixed models.

S1.18 Mixed pixels and zonal data aggregation by specific percentiles

With a GSD of 5.2 cm, individual wheat plants could not be recognized in this study, as ears, culms and leaves of wheat are smaller than the GSD. Consequently, pixels in images are mixed pixels, containing TIR information from both, different wheat organs and background soil. Between the sowing rows and on spots with poor plant development, the contribution of background to the value of a single pixel is larger than within sowing rows. When aggregating pixel values per ROI by simply calculating the mean, genotype-specific canopy cover values will bias the CT estimate, estimating CT to be more biased towards the temperature of the background when genotype specific canopy cover is lower. Using specific percentiles allows to compensate for mixed-pixel problems to a certain extent. In accordance with Perich et al. (2020) it was shown that the 50^{th} percentile is suitable for most of the situations.

Deery, Rebetzke, Jimenez-Berni, James, et al. (2016) proposed to use a function that excludes the hottest and coolest pixels. They argued that, depending on the daytime and meteorological conditions, soil might be hotter or cooler than the canopy. By limiting the analysis to a range of central percentiles, the disturbing influence of soil can be covered in both situations. Following the same argumentation, Perich et al. (2020) used the median

as aggregation function, which does not require assuming an upper and lower threshold. In this work, the two approaches were combined: By using an empirically determined specific percentile for each year, the influence of soil signal was minimized.

S1.19 SpATS Code

The code below is an example for a one-stage SpATS model where just trigger timing is considered in addition to the spatial model and experimental design factors. An "f" at the end of a variable name indicates that the variables were defined as factors.

```
SpATS_fit <- SpATS(response = TempImgPlot,
          random = ~ Rowf + Colf + Plotf + TriggerTimingf + Genotypef:Treatmentf,
          fixed = ~ Treatmentf + Repf,
          spatial = ~PSANOVA(Row, Col, nseg = c(nX, nY), nest.div = c(1,1)),
          genotype =``Genotypef'', genotype.as.random = TRUE,
          data = df_for_correction,
          weights = df_for_correction$weights,
          control = list(maxit = 100, tolerance = 1e-03, monitoring = 0))
```

Table S1.5: Variable explenation from SpATS code example

Variable	Explanation
TempImgPlot	Aggregated temperature for each plot per image
Genotypef	Genotype as factor
Treatmentf	Agricultural treatment as factor
Plotf	Unique plot ID as factor
Row	Plot sequence perpendicular to direction of sowing
Col	Plot sequence in direction of sowing, *i.e.*, along columns
Rowf	Row as factor
Colf	Column as factor
Repf	Replication as factor
TriggerTimingf	Time stamp of image in seconds after flight start as factor
nX & nY	number of segments for smoothing splines

S1.20 ASReml-R Code

The code below is an example code for an ASReml-R model that includes trigger timing as well as viewing geometry in direction of the sun and in sowing row direction. Knot points have to be set so the models do not become too heavy. An "f" at the end of a variable name indicates that the variables were defined as factors.

```
as.base <- asreml(
  fixed = TempImgPlot ~ Treatment,
  random = ~ Genotypef +
    ar1(Rowf):ar1(Colf) +
    id(Colf) + id(Rowf) +
    id(Repf) +
    id(Treatment):id(Repf) +
    id(Plotf) +
    spl(Col):spl(Row) +
    spl(TriggerTiming) +
    spl(Longitudinal_dist_from_ex_pos):spl(Lateral_dist_from_ex_pos) +
    spl(Longitudinal_dist_sun_direction):spl(Lateral_dist_sun_direction),
  residual = ~ id(units),
  knot.points = list(
    Longitudinal_dist_from_ex_pos = seq(
      min(df_single_flight$Longitudinal_dist_from_ex_pos),
      max(df_single_flight$Longitudinal_dist_from_ex_pos),
```

```
      length.out = 10
    ),
    Lateral_dist_from_ex_pos = seq(
      min(df_single_flight$Lateral_dist_from_ex_pos),
      max(df_single_flight$Lateral_dist_from_ex_pos),
      length.out = 10
    ),
    Longitudinal_dist_sun_direction = seq(
      min(df_single_flight$Longitudinal_dist_sun_direction),
      max(df_single_flight$Longitudinal_dist_sun_direction),
      length.out = 10
    ),
    Lateral_dist_sun_direction = seq(
      min(df_single_flight$Lateral_dist_sun_direction),
      max(df_single_flight$Lateral_dist_sun_direction),
      length.out = 10
    ),
    TriggerTiming = seq(0, max(df_single_flight$TriggerTiming), 4)
  ),
  data = df_single_flight
)
```

TABLE S1.6: Variable explenation from ASReml-R code example

Variable	**Explanation**
TempImgPlot	Aggregated temperature for each plot per image
Genotypef	Genotype as factor
Treatment	Agricultural treatment if applied
Plotf	Unique plot ID as factor
Row	Plot sequence perpendicular to direction of sowing
Col	Plot sequence in direction of sowing, *i.e.*, along columns
Rowf	Row as factor
Colf	Col as factor
Repf	Replication as factor
TriggerTiming	Time stamp of image in seconds after flight start
Longitudinal_dist_from_ex_pos	Distance of the plot center from the drone in direction of sowing
Lateral_dist_from_ex_pos	Distance of the plot center from the drone perpendicular to direction of sowing
Longitudinal_dist_sun_direction	Distance of the plot center from the drone in sun direction
Lateral_dist_sun_direction	Distance of the plot center from the drone perpendicular to sun direction

S2 Supplementary Materials - Analysis of variance and its sources in UAV-based multi-view thermal imaging of wheat plots

S2.1 Experimental design - EuVar

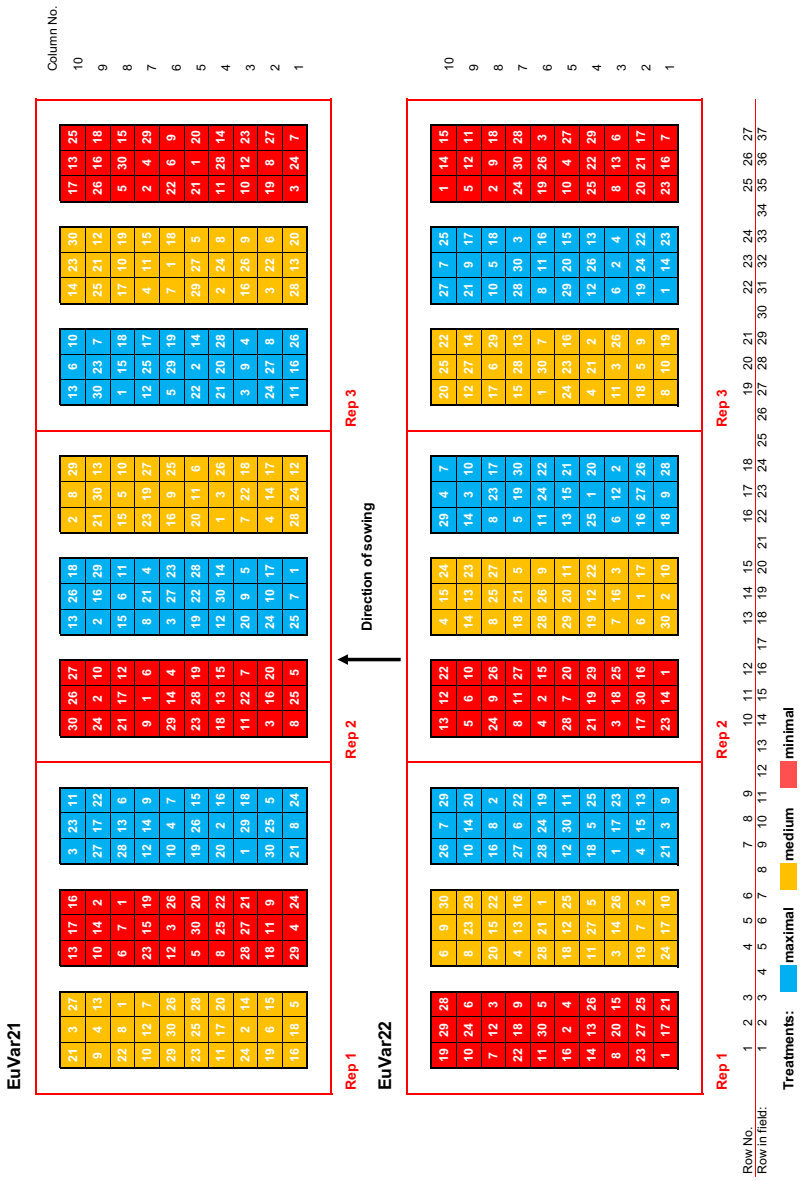


FIGURE S2.1: The experimental design of EuVar for the two years. The numbers inside the blocks indicate the genotypes.

S2.2 Experimental design - SwiVar

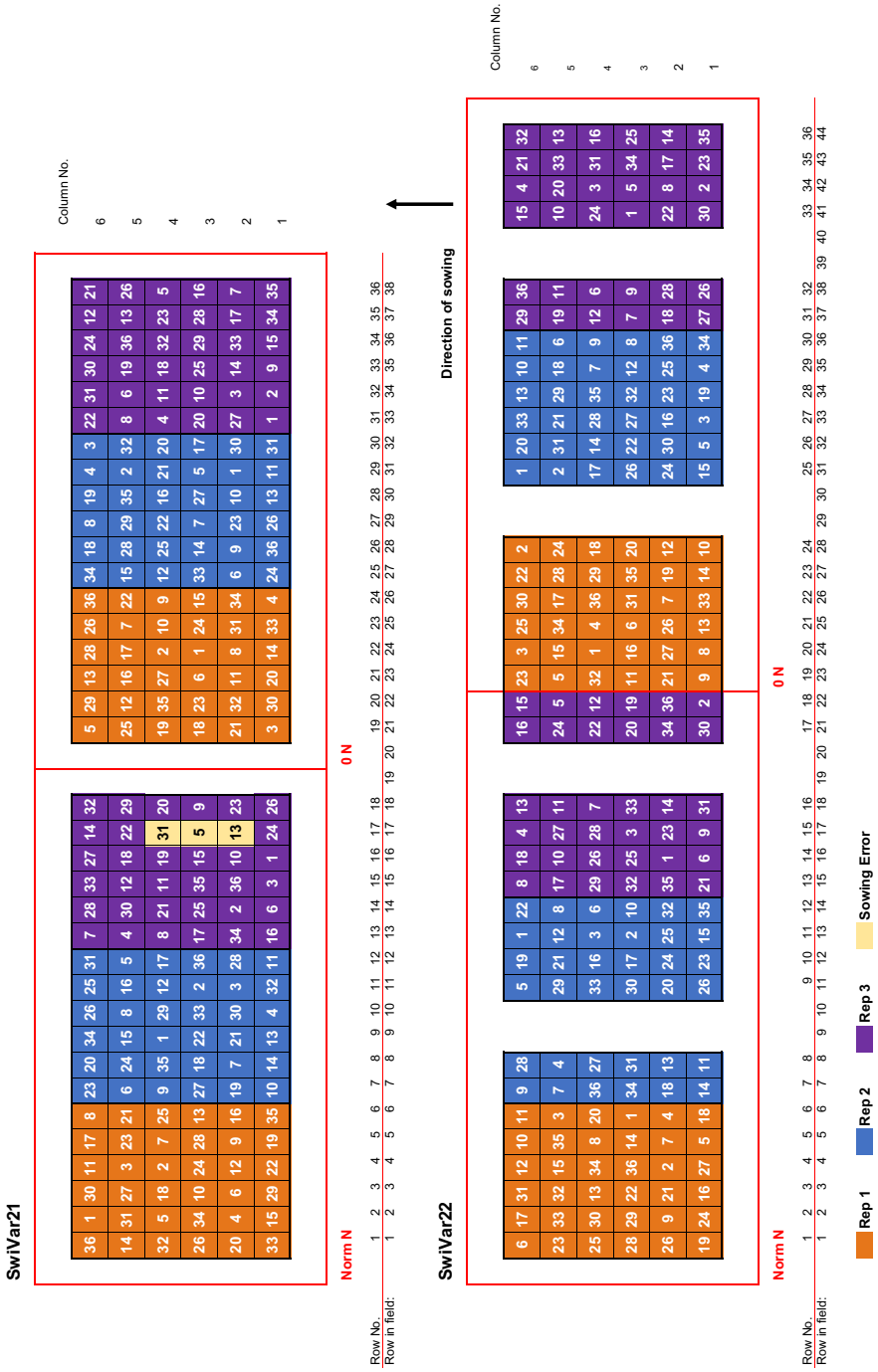


FIGURE S2.2: The experimental design of SwiVar for the two years. The numbers inside the blocks indicate the genotypes.

S2.3 Details on Field treatments

TABLE S2.1: Overview of trial treatments and most important field interventions for all trials. "too wet" indicates that treatments were intended but could no be applied as conditions were to heavy machinery could not enter the field.

				Herbicides					**Fertilization (kg/ha)**			
Experiment	**Treatment**	**Sowing date**	**Harvest date**	Monocot. 1st	Monocot. 2nd	Dicot.	**Growth regulator**	**Fungicide**	N	CaO	MgO	SO_3
EuVar21	Minimal	2020-10-22	2021-07-20	Archipel®	too wet	too wet	-	-	140	15	32	30
	Medium						Moddus®	-				
	Maximal						Moddus®	Amistar®				
SwiVar21	Fertilized	2020-11-07	2021-07-20	Archipel®	too wet	too wet	-	-	140	15	32	30
	Not fertilized						-	-	-	-	23	30
EuVar22	Minimal	2021-10-15	2022-06-30	Archipel®	Othello Star®	Cleave®/ Express Max®	-	-	140	23	37	30
	Medium						Moddus®	-				
	Maximal						Moddus®	Amistar®				
SwiVar22	Fertilized	2021-10-15	2022-07-06	Archipel®	Othello Star®	Cleave®/ Express Max®	-	-	140	5	26	30
	Not fertilized						-	-	-	-	23	30

TABLE S2.2: Details on fertilizer application with split N applications.

				Fertilization (kg/ha)			
	Treatment	**N Application split**	**Date**	N	CaO	MgO	SO_3
EuVar21	All	-	2021-02-22		-	23	30
		1	2021-02-23	50	-	-	-
		2	2021-03-24	60	10	6	-
		3	2021-04-20	30	5	3	-
SwiVar21	Fertilized	-	2021-02-22	-	-	23	30
		1	2021-02-23	50	-	-	-
		2	2021-03-24	60	-	-	-
		3	2021-04-20	30	-	-	-
	Not fertilized	-	2021-02-22	-	-	23	30
EuVar22	All	1	2022-02-08	50	8	5	-
		-	2022-03-03	-	-	23	30
		2	2022-02-23	60	10	6	-
		3	2022-04-28	30	5	3	-
SwiVar22	Fertilized	-	2022-03-03	-	-	23	30
		1	2022-03-11	50	-	-	-
		2	2022-03-30	60	-	-	-
		3	2022-04-28	30	-	-	-
	Not fertilized	-	2022-03-03	-	-	23	30

TABLE S2.3: Chemical compositions of field treatments and quantities applied.

Procuct	**Active ingredients**	**Application rate (g/ha)**	**Producer**
Archipel®	Iodosulfuron-methyl-sodium	9	Syngenta
	Mesosulfuron-methyl	9	
	Mefenpyr-diethyl	27	
Moddus®	Trinexapac-ethyl	125	Syngenta
Amistar®	Azoxystrobin	200	Syngenta
	Cyproconazole	80	
Othello Star®	Iodosulfuron-methyl-sodium	9	Bayer
	Mesosulfuron-methyl	9	
	Mefenpyr-diethyl	27	
	Thiencarbazone-methyl	7.5	
Cleave®	Fluroxypyr	90	Syngenta
	Fluroxypyr-meptyl	130	
	Florasulam	23	
Express Max®	Metsulfuron-methyl	5	Syngenta
	Tribenuron-methyl	5	

S2.4 Overview on flights

Table S2.4: Overview on number of thermal measurements per date and project

Year	Project	Date	No. of flights
2021	EuVar	2021-06-12	13
	SwiVar	2021-06-19	15
	SwiVar	2021-06-28	13
	EuVar	2021-07-01	9
2022	EuVar	2022-05-14	3
	SwiVar		6
	EuVar	2022-05-18	3
	SwiVar		6
	EuVar	2022-06-04	6
	SwiVar		3
	EuVar	2022-06-11	5
	SwiVar		3
	SwiVar	2022-06-14	8
	SwiVar	2022-06-18	6

S2.5 Flight campaigns

Optimal conditions for TIR imaging surveys are a clear blue sky, warm temperatures, and little or no wind (Perich et al., 2020). In 2021, the weather conditions were often suboptimal for flying (rain and cloud cover throughout the growing season) (Fig. S2.5a). On rare days with suitable weather, as many flights as possible were conducted on one experiment before, during, and after solar noon.

2022 was a very hot and dry year (Fig. S2.5b). The phenological window suitable for flying was short, but within this window conditions were often suitable for TIR imaging. Consequently, in season 2022, flights were conducted on more days but on individual days, with fewer flights per day than in 2021. Flights were carried out between late morning and mid-afternoon as suggested by Deery, Rebetzke, Jimenez-Berni, James, et al. (2016), and Perich et al. (2020), with only some exceptions, where flights were also taken later in the day. Based on the results of 2021, flights were restricted to times after 12:00 in 2022. Fig. S2.5 provides an overview on when TIR measurements were conducted within the growing season, while Figs. S2.6 and S2.7 show, at what time of day, the flights were conducted during the single days. In total, 99 flights were performed (Table S2.4).

S2.6 Camera settings and flight planning

The flights were carried out with a DJI Matrice 200 drone (SZ DJI Technology Co. Ltd., China) that carried a DJI Zenmuse XT TIR sensor equipped with a 9 mm, f/1.4 lens. Pixel resolution of 640 x 512 was achieved by individual uncooled VOx microbolometers arranged in a focal plane array. The field of view (FOV) was 69° x 56°. The temperature was measured

in the wavelength range of 7.5−13.5 µm and the thermal sensitivity was < 50 mK. In high gain mode, the camera could measure the temperature in the range of -25 °C to +135 °C. The absolute measurement accuracy was ± 10 °C according to the manufacturer. The external parameters were set in the DJI Pilot software (SZ DJI Technology Co. Ltd., China) to the same values for all flights using the default DJI settings. The scene emissivity was set to 100 %, background and air temperature were set to 22 °C. The sensor provided the option of periodical flat-field correction throughout measurement to reduce the noise of non-uniformity effects. This periodic compensation would reduce the comprehensibility of drift effects. In addition, Kelly et al. (2019) showed that the non-uniformity correction alone was not sufficient to correct for the drift effect during flights. The flat field correction was therefore deactivated following Mesas-Carrascosa et al. (2018).

The flights were planned with DJI Pilot software. The exposure interval was 2 s. While mission planing often is done by defining a minimal front- and side overlap, this was not possible due to software restrictions for the drone−sensor combination used. In addition, using way-point flight planning with a fixed exposure interval allowed for a manually defined camera heading throughout the flight. Therefore, the heading of drone and TIR camera remained relatively stable throughout the flight and did not change with flight path direction changes. The flight speed was limited to 4 m s^{-1}.

S2.7 DEM creation

TIR images often do not provide enough spatial detail to generate DEMs of sufficient quality (*e.g.* Malbéteau et al., 2021; Treier et al., 2024). Thermal images have lower pixel resolution and contrast compared to RGB images (Boesch, 2017). TIR based DEMs may therefore appear flat with no distinct plot pattern. Thus, DEMs were also based on the RGB data of Micasense RedEdge-MX Dual camera (MicaSense Inc., Seattle, Washington, USA) which allows for more spatial detail.

DEMs were created on the basis of aligned images in Agisoft Metashape and were derived from thermal data in 2021, but not in 2022, when DEMs were generated from RGB data. Both methods allowed generating DEMs of sufficient positioning precision (positioning RMSE vertical: 2.5 cm, horizontal: 1.5 cm based on Agisoft alignment error estimates for ground control points). For each year, a representative DEM was chosen that was created from images taken after the wheat stem elongation phase and before early senescence, when the canopy height remained stable. The quality of the DEMs was checked by visually inspecting the plausibility of the positioning of the masks projected on single images in multi-view pre-processing. The projected masks needed to be centered within plots and rectangular in shape. For EuVar21, the DEM was based on the second flight of the thermal campaign flown on 2021-06-12 at 12:30 and for SwiVar21, the DEM was based on the third flight of the thermal campaign flown on 2021-06-19 at 16:30 with a flight height of 40 m for both flights. The ground sampling distance (GSD) of the TIR images was 5.15 cm/pix and the spatial resolution of the DEMs was 41 cm/pix and 16 cm/pix for EuVar21 and SwiVar21 respectively. With this coarse resolution, inconsistencies such as holes in the DEM could be leveled out. The DEMs used in 2022 were based on flights with the Micasense sensor at a flight height of 40 m at 2022-06-04 and 2022-05-18 for EuVar22 and SwiVar22 respectively. The GSD of the images was 2.71 cm/pix. The DEMs of 2022 did not exhibit holes, and the spatial resolution of the DEM was set to 2.71 cm/pix too.

S2.8 TIR image pre-processing

Radiometric JPEG format contains an 8-bit gray scale JPEG image as well as a 14-bit array with digital numbers (DN), which represent the magnitude of TIR radiation (Kelly et al., 2019). The DNs in the 14-bit arrays of the radiometric JPEGs were transformed to TIFF files representing temperature in $°C$ x 1000 by using a Python 3.8 script (van Rossum, Guido and Drake, Fred L., 2009) and a modified version of the Flir Image Extractor (`https://github.com/ITVRoC/FlirImageExtractor`), which allowed for batched processing.

Plot masks were created for each plot in Qgis 3.16 (QGIS Development Team, 2022), to determine the ROIs from which data was used for analysis. To account for border effects in the field and for inaccuracies of georeferencing and superimposition of different flights, a border buffer of 25 cm was applied to all masks on plot width. On plot length, the buffer was up to 1 m, leaving at least a surface of 2.1 m^2 to be analyzed in each plot. The plot masks were saved to GeoJSON format.

Imaging techniques deliver pixel values in a 2-D space. In order to evaluate experimental units, pixels within ROIs in this 2-D space must be analyzed. Usually, this is done using zonal statistics, that is, the pixels within ROIs are reduced to single values using statistical aggregation functions. In this work, an empirically determined specific percentile for each year was used.

The procedure for finding an optimal percentile was described in Treier et al. (2024). In short, for each percentile, heritabilities were calculated in a simplified mixed model in SpATS (Rodríguez-Álvarez et al., 2018). The resulting percentile-heritability relations were plotted for graphical comparison. Two quantitative criteria were used to select the percentiles: Select a percentile in the center of a percentile region where (1) the heritability is close to the maximum, and (2) closely adjacent percentiles have similar heritabilities, *i.e.* the heritability is stable in the respective percentile region. For each experiment in each year, the optimal percentile was determined. The values within the ROIs were reduced to a single value by using the optimal percentile. One value per measurement (for multiple measurements per plot) was then used as plot-wise CT value in further analysis. The same percentile was used for the aggregation of all flights on one experiment within one year.

S2.9 Multi-view pre-processing

The camera positions (longitude, latitude, height) and orientations (pitch, roll, yaw) at the moment of triggering of individual images were estimated in an indirect sensor orientation approach (Benassi et al., 2017) in Agisoft Metashape after aligning images. Using the estimated trigger positions, the single images were projected onto the DEMs by ray tracing as described in Roth, Aasen, et al. (2018), Roth, Camenzind, et al. (2020) and Treier et al. (2024). This allowed for the projection from geographic coordinates (*e.g.* EPSG:2056 reference system) to image coordinates. As a result, plot masks of ROIs were created for each trigger position (*i.e.* for each image) where at least one plot was entirely inside the field of view (FOV) of the camera. As coordinates were identical for 8-bit JPEG images and 14-bit intensity value arrays, the image-wise masks could be directly applied to the temperature TIFF files. This approach of identifying the ROIs for each plot on every single image is referred to as multi-view. For each plot on each TIF file, all percentiles were extracted with a Python 3.8 script and saved to a CSV file.

S2.10 Flight operation

In 2021, the camera was turned on at least 15 min before each flight to allow the temperature signal to stabilize. In 2022, an additional set of batteries was used and the stabilization period was increased to 30 min. In situations where the battery was not sufficient anymore to complete all flights, the temperature stabilization was not repeated after a rapid battery change. After the first flight campaigns in 2021, a rather strong drift of apparent temperature was noticed that seemed to be particularly strong during the beginning of flights. To further reduce initial drift, the drone was hovered above the wheat field for about one minute in addition to the previous temperature stabilization on the ground before the measurement flight sequence was started.

S2.11 Thermal ground control points

GCPs were produced following Perich et al. (2020) by gluing triangles of 2 mm thick aluminum sheets on polystyrene foam plates. These plates had an extent of 1 m x 0.5 m or 0.5 m x 0.5 m. Unlike in Perich et al. (2020), the aluminum sheets were left blank as was done in other TIR surveys (*e.g.* Mesas-Carrascosa et al., 2018; Aragon et al., 2020) and not painted black. This avoided large temperature gradients in the FOV and reduced possible adjacency effects of hot objects (Aragon et al., 2020; Zheng et al., 2019).

S2.12 Georeferencing images

The 8-bit JPEGs of the radiometric image as well as the RGB images were aligned in the structure-from-motion-based software Agisoft Metashape Professional (Agisoft LLC, St. Petersburg, Russia). TIR images feature a low spatial resolution and are therefore difficult to georeference. No precise GPS device was available to measure GCP positions, and an indirect referencing approach was used. One RGB project served as a reference project and was referenced by the positioning information of the drone available for each image in the meta-data. The GCP coordinates were extracted from this project and used to reference all other projects of one year. Conventional GCPs are difficult to detect in TIR images, and in one RGB project, the RGB GCPs were visible together with the thermal GCPs. The locations of the thermal GCPs were then extracted from this RGB project and used to reference the thermal projects. This allowed for a correct geographic orientation and a absolute positioning precision within 2 m horizontally and vertically according to a quality check in Qgis (QGIS Development Team, 2022). With this procedure all TIR flights were georeferenced in the Cartesian Swiss coordinate system EPSG:2056 (CH1903 + LV95), which allowed one to precisely superimpose the aligned images of the different flights. The relative positioning precision between flights was estimated to be 15 cm or smaller based on marker position error estimates in Agisoft.

The 8-bit JPEGs were preferred over the 14-bit TIFF images in the process of aligning images, as they provide better contrast and contain meta information on TIR camera position and orientation during triggering, which was allowing for a valid alignment more reliably. However, these 8-bit JPEG are just a nonlinear, visually augmented interpretation of TIR with a value range of 0 to 255 and could not be used for analysis of temperature. As the pixel position remained consistent between the two formats, 8-bit JPEG images were replaced by 14-bit TIFF files after alignment for further temperature analysis.

S2.13 Covariates related to viewing geometry

TABLE S2.5: List of all covariates calculated from multi-view data

Covariate	Description
Angle Sun-Plot-Drone	The angle between sun, plot and drone
Azimuth drone	The Azimuth of the drone, seen from the plot (horizontal planar clockwise angle from north)
Azimuth sun	The Azimuth of the sun, seen from the plot
Azimuth diff	Difference between the two Azimuth angles of sun and drone
Elevation sun	The vertical angles from the horizon to the sun
Elevation drone	The vertical angles from the horizon to the drone
Lateral angle row dir.	Lateral angle of the plot relative to the drone in sowing row direction
Lateral angle sun dir.	Lateral angle of the plot relative to the drone in sun direction
Longitudinal angle row dir.	Longitudinal angle of the plot relative to the drone in sowing row direction
Longitudinal angle sun dir.	Longitudinal angle of the plot relative to the drone in sun direction
Lateral dist row dir.	Lateral distance of the plot relative to the drone in sowing row direction
Lateral dist sun dir.	Lateral distance of the plot relative to the drone in sun direction (i.e. orthogonal to principle plane of the sun)
Longitudinal dist row dir.	Longitudinal distance of the plot relative to the drone in sowing row direction
Longitudinal dist sun dir.	Longitudinal distance of the plot relative to the drone in sun direction (i.e. in the principle plane of the sun)
Trigger timing	The time stamp when each TIR image was taken
Sensor x	X coordinate of the plot center on the sensor plane (image coordinates)
Sensor y	Y coordinate of the plot center on the sensor plane (image coordinates)
Total dist.	Total distance between drone and plot center

S2.14 Spectral properties of the Micasense RedEdge-MX Dual Camera System

TABLE S2.6: Specification of the ten bands of the Micasense RedEdge-MX Dual Camera System

Micasense band-name	Band variable	Center wave-length (nm)	Band width (nm)	Micasence Band Suffix
Coastal Blue	$Blue_{444}$	444	28	6
Blue	$Blue_{475}$	475	32	1
Green	$Green_{531}$	531	14	7
Green	$Green_{560}$	560	27	2
Red	Red_{650}	650	16	8
Red	Red_{668}	668	14	3
Red Edge	Red_Edge_{705}	705	10	9
Red Edge	Red_Edge_{717}	717	12	5
Red Edge	Red_Edge_{740}	740	18	10
Near IR	NIR_{842}	842	57	4

S2.15 Multispectral measurements

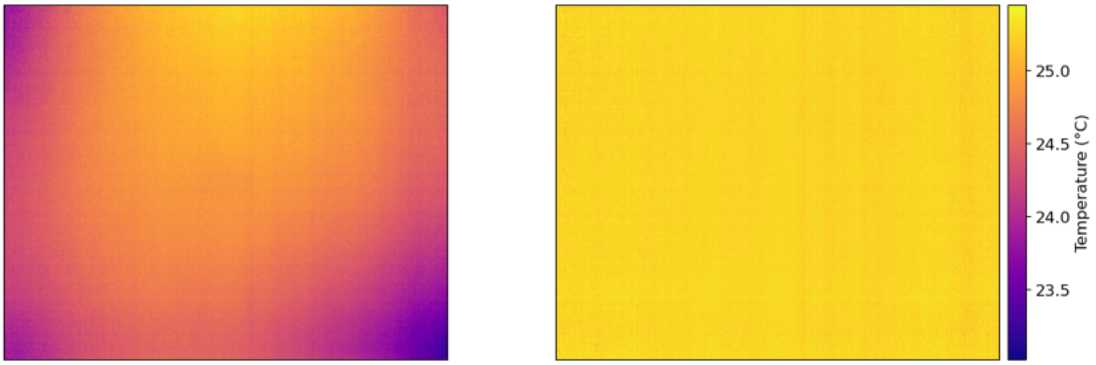
The sensor was carried by a DJI Inspire 2 drone (SZ DJI Technology Co. Ltd., China). The flight height was 60 meter in 2021 and 40 meter in 2022 resulting in a ground sampling distance (GSD) of 3.98 cm and 2.71 cm, respectively. The side overlap was set to 80 %, the flight speed was limited to $5\,\mathrm{m\,s^{-1}}$ and an image was taken every 2 s, resulting in a front overlap of approximately 70 % and 60 % for the two flight heights, respectively.

S2.16 Vignetting correction

This procedure was explained in detail in Treier et al. (2024) and for the sake of clarity, the method is described here again. To generate a vignetting correction image, the drone was located indoors with the thermal sensor pointing to a hard foam PVC sheet. The distance between the sensor and the sheet was about 40 cm and the sheet completely covered the FOV of the camera. The ambient temperature was 22 °C and there was no direct light on the PVC

sheet. The ambient light in the room was reduced (turned off in the respective section of the room) to mitigate artifacts of light. The PVC sheet was placed inside the room 5 hours prior to use to reach temperature equilibrium.

The camera was started to stabilize. After 1 h, TIR images of the PCV sheet were taken at an interval of 5 s for more than 30 min. A vignetting correction image was then calculated as the pixel-wise mean of these 413 images in Python 3.8. The pixel valvues of the resulting correction image were subtracted from corresponding pixel values of all TIR images of all flights to obtain vignetting-corrected images. Fig. S2.3 shows an image that was taken after the camera was running for more than 70 min before and after correction.



(a) Before correction (b) After correction

FIGURE S2.3: (a) shows an image of a homogeneous PVC sheet that was part of the set used to create a vignetting correction image. A vignetting pattern is clearly visible with a cooling trend toward the edges. (b) shows the same image after correction was applied. Vignetting was clearly mitigated and the image of the PVC sheet now appears flat with almost no trends visible.

S2.17 Fan experiment to determine the influence of wind

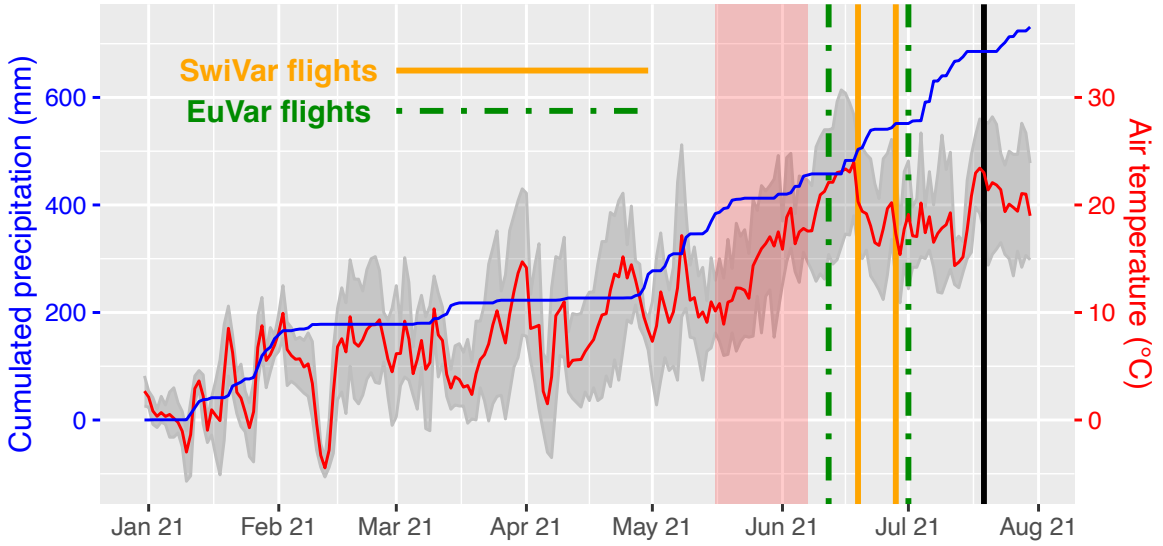
The drone with the thermal camera was placed indoors at an ambient air temperature of 20 °C and the thermal camera pointed to a hard foam PVC sheet (Fig. S2.4) similar to what was done for the vignetting correction. The distance between the sensor and the sheet was about 120 cm. At a distance of 90 cm and an angle of about 45° a fan was placed, pointing in the direction of the camera. The fan generated a wind speed of about 3 - 3.3 $\mathrm{m\,s^{-1}}$ at the sensor. At a distance of 67 cm and an angle of 90°, a Philips Attralux spot (230V, 150W) pointed to the camera as an artificial source of heat. The spot did not point inside the FOV of the camera but just heated it up from the side. The ambient light in the room was reduced to minimize disturbances from other sources of light.

The camera was started for stabilization and images were taken from the beginning. To examine whether sudden and strong temperature gradients have a sustained influence on subsequent TIR readings, warm and hot disturbance objects (hands at body temperature and a water cooker with boiling water) were introduced into the scene for several seconds 35 min after camera startup. Each disturbance was repeated three times with a period of 5 min for stabilization after each disturbance. 75 min after camera startup, the heating lamp was started and after another 5 min, the fans was turned on and off at an interval of 5 min. The heating lamp was turned off again 5 min after the last fan iteration.

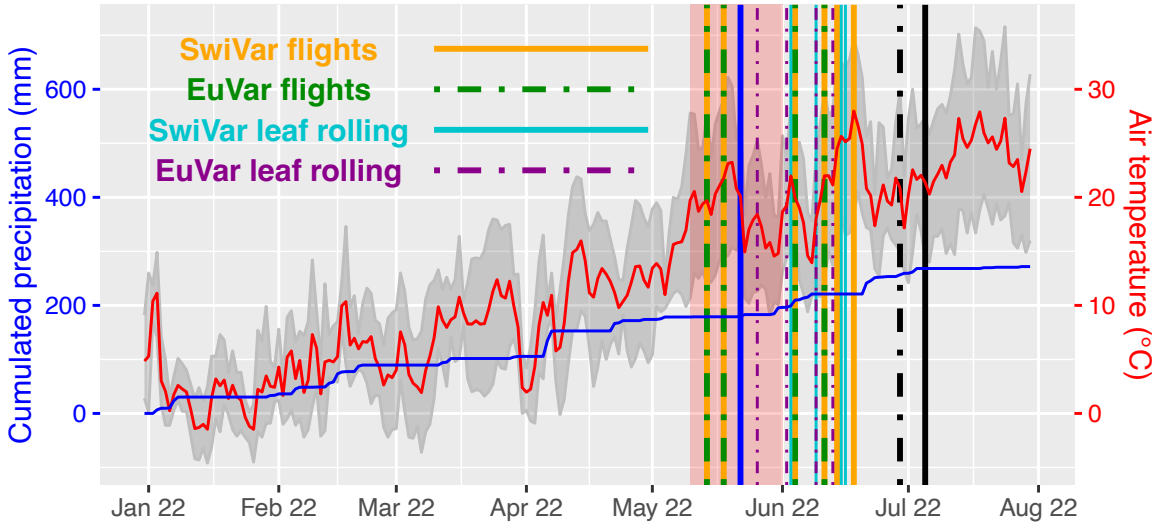
On the TIR images, a polygon was defined, covering just the PVC sheet. From within this polygon, the mean temperatures and standard deviation of pixel-wise temperatures were extracted with a Python 3.8 script.

FIGURE S2.4: Setup of fan experiment. The drone was pointing at a PVC sheet. From an oblique frontal angle, the fan was blowing in the direction of the sensor. From the side, a lamp was heating the sensor without directly pointing into the FOV of the camera.

S2.18 Environmental conditions and timing of measurements



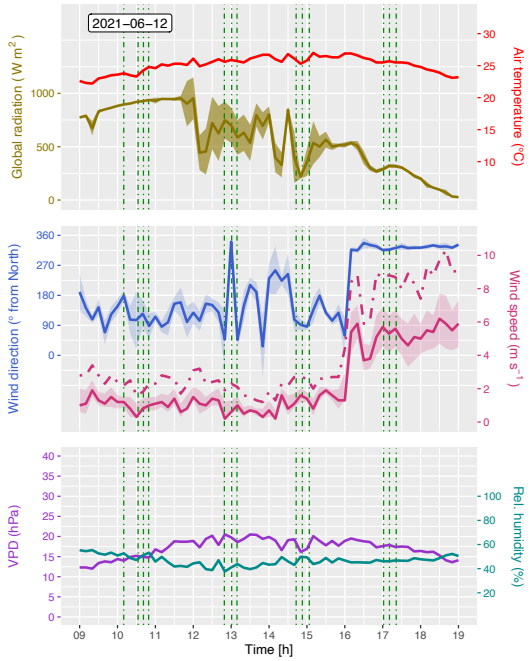
(a) Weather overview of growing seasons 2021



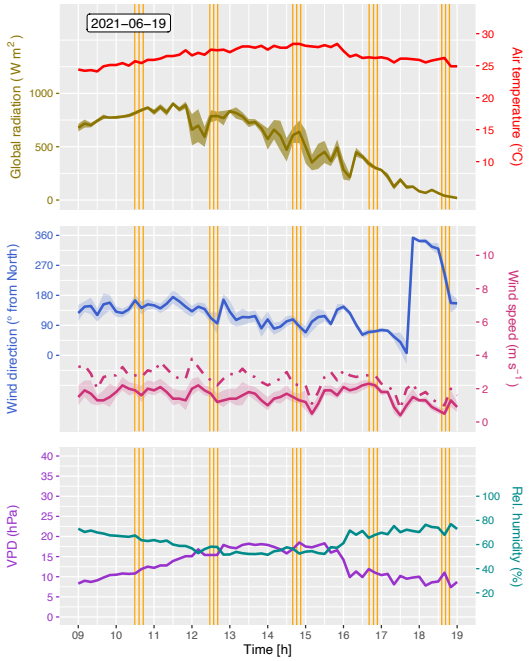
(b) Weather overview of growing seasons 2022

FIGURE S2.5: (a) and (b) show the general weather conditions during the growing seasons of 2021 and 2022 from January until after harvest. Red shows the mean air temperature, and the shades indicate daily temperature minima and maxima. The orange lines and green dashed lines represent the flight dates of SwiVar and EuVar, respectively. Cumulative precipitation is shown as a rising blue line, and the vertical blue line indicates an irrigation intervention for SwiVar22 (30 mm of water). During the period shaded in red, heading was observed in the field. Cyan and purple lines indicate flag leaf rolling ratings in 2022. Harvest dates are marked by black lines (the dashed black line in 2022 is the harvest date of EuVar22 which was harvested before SwiVar22).

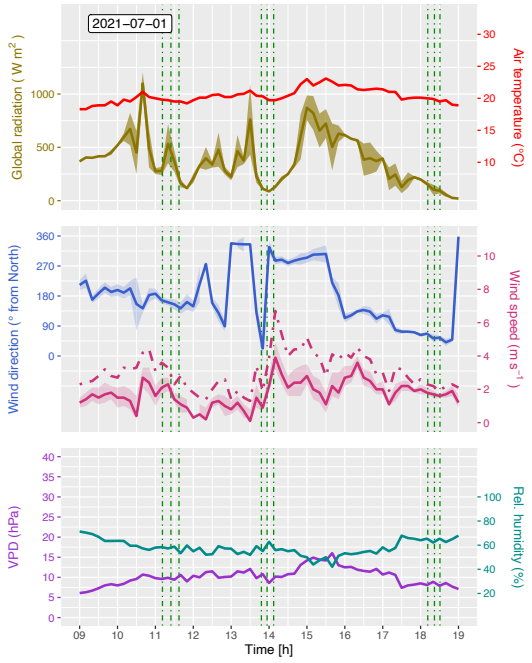
S2.19 Weather data



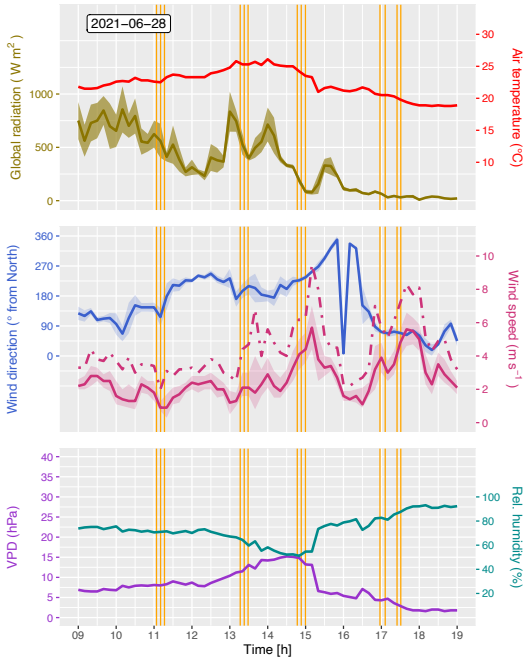
(a) 2021-06-12



(b) 2021-06-19

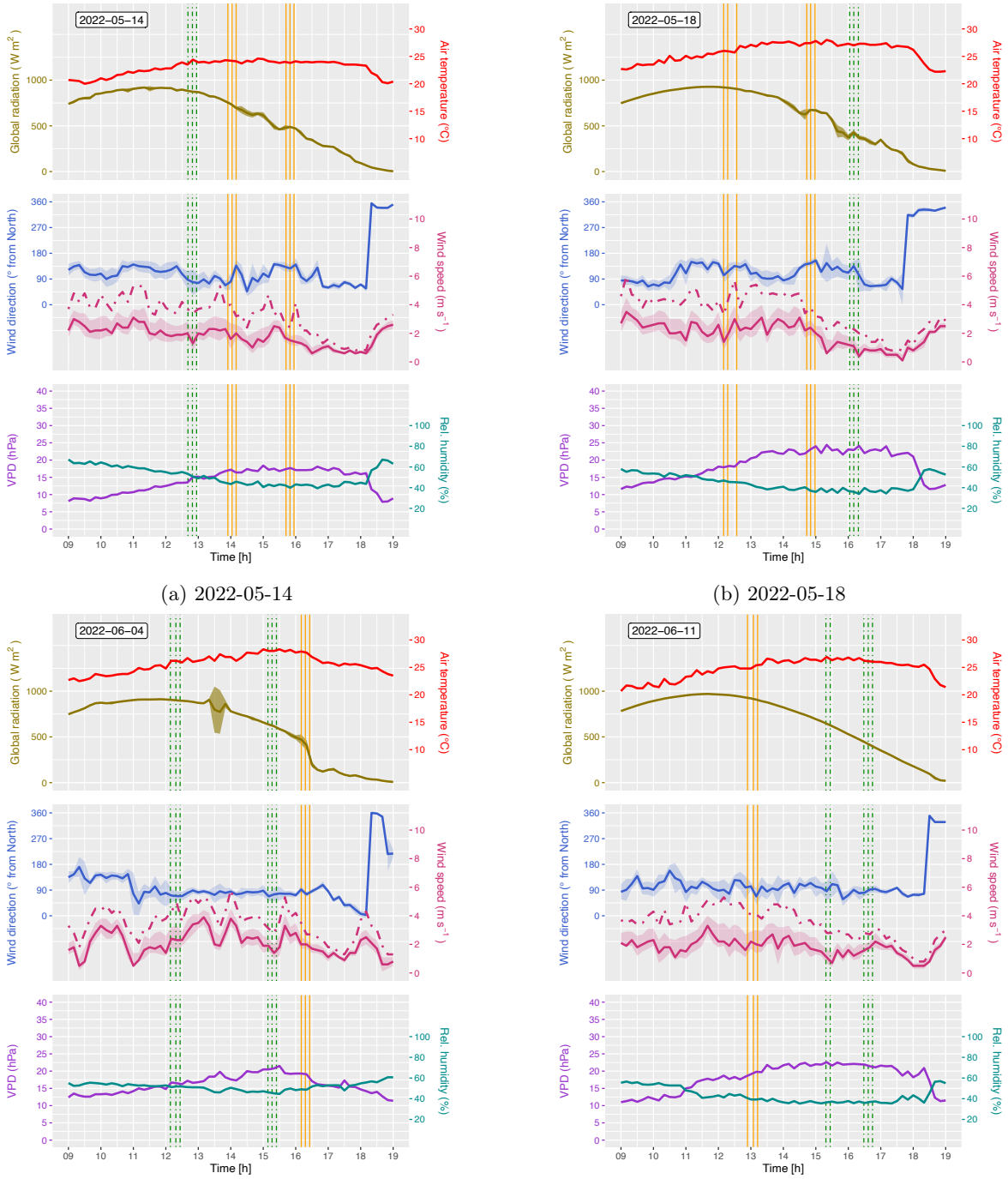


(c) 2021-07-01



(d) 2021-06-28

FIGURE S2.6: (a - c) show the detailed weather conditions on measurement days in 2021. Solid lines show means, shades are means ± SD and dashed lines show the maxima for 10 min intervals. The vertical lines indicate the different flights of EuVar (green) and SwiVar (yellow).



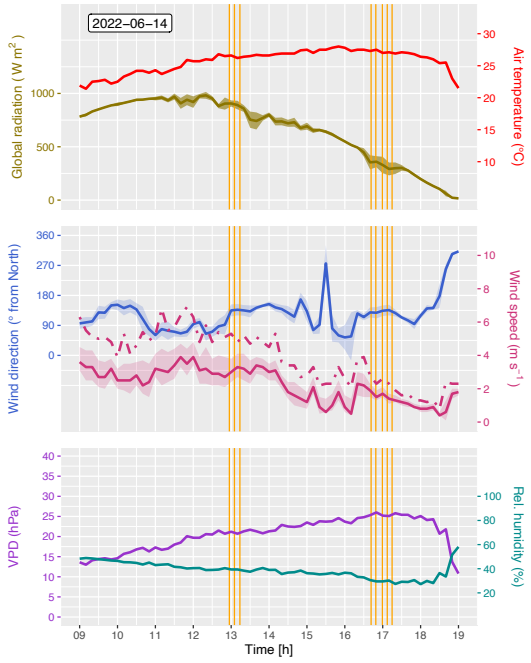
(a) 2022-05-14

(b) 2022-05-18

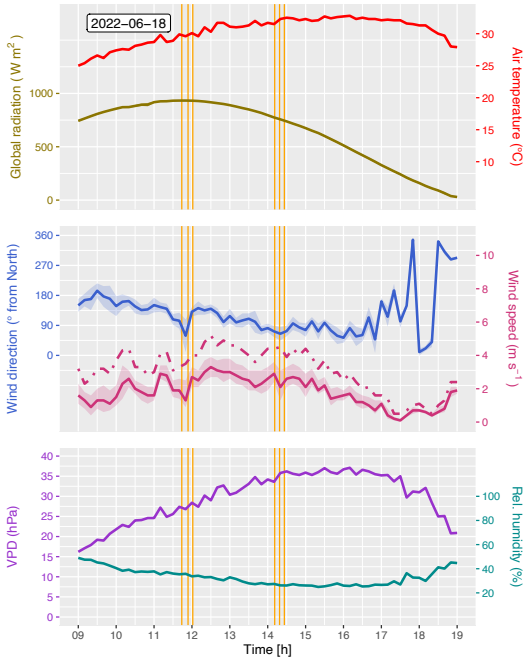
(c) 2022-06-04

(d) 2022-06-11

FIGURE S2.7: (a - f) show the detailed weather conditions on days of measurements in 2022.



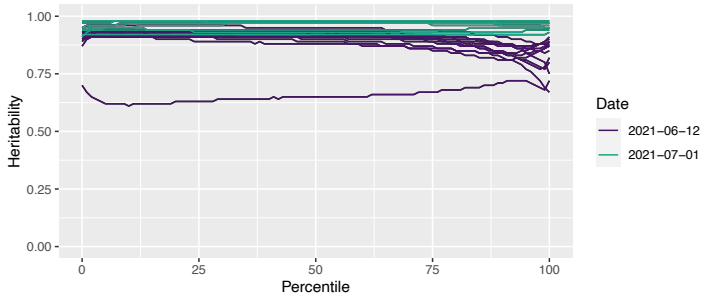
(e) 2022-06-14



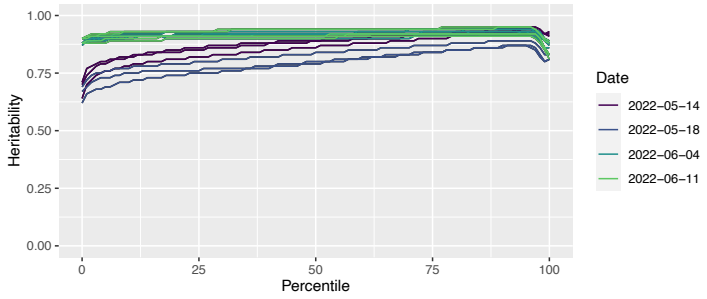
(f) 2022-06-18

FIGURE S2.7: (a - f) show the detailed weather conditions on days of measurements in 2022. (cont.)

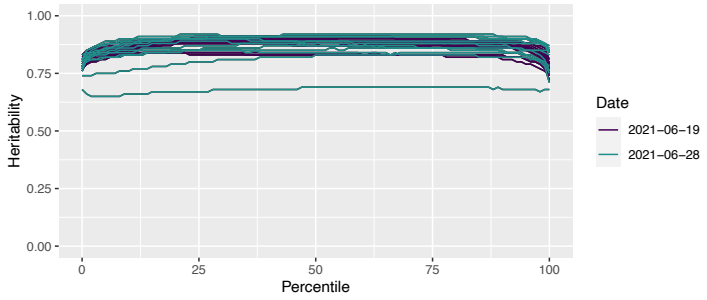
S2.20 Multi-view percentile selection



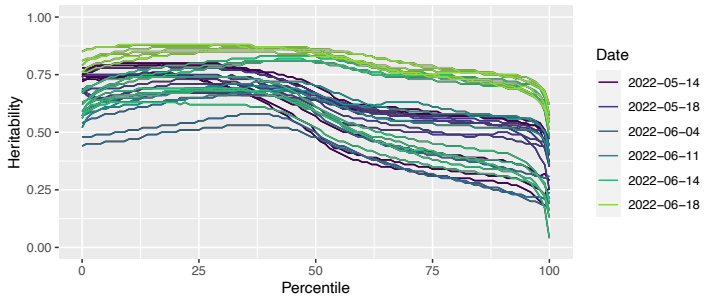
(a) EuVar21



(b) EuVar22



(c) SwiVar21



(d) SwiVar22

FIGURE S2.8: (a - d) show the heritability of the multi-view method for each pixel value percentile for each flight conducted on EuVar and SwiVar

S2.21 Correction steps EuVar

S2.21.0.1 No correction applied - plot-wise means

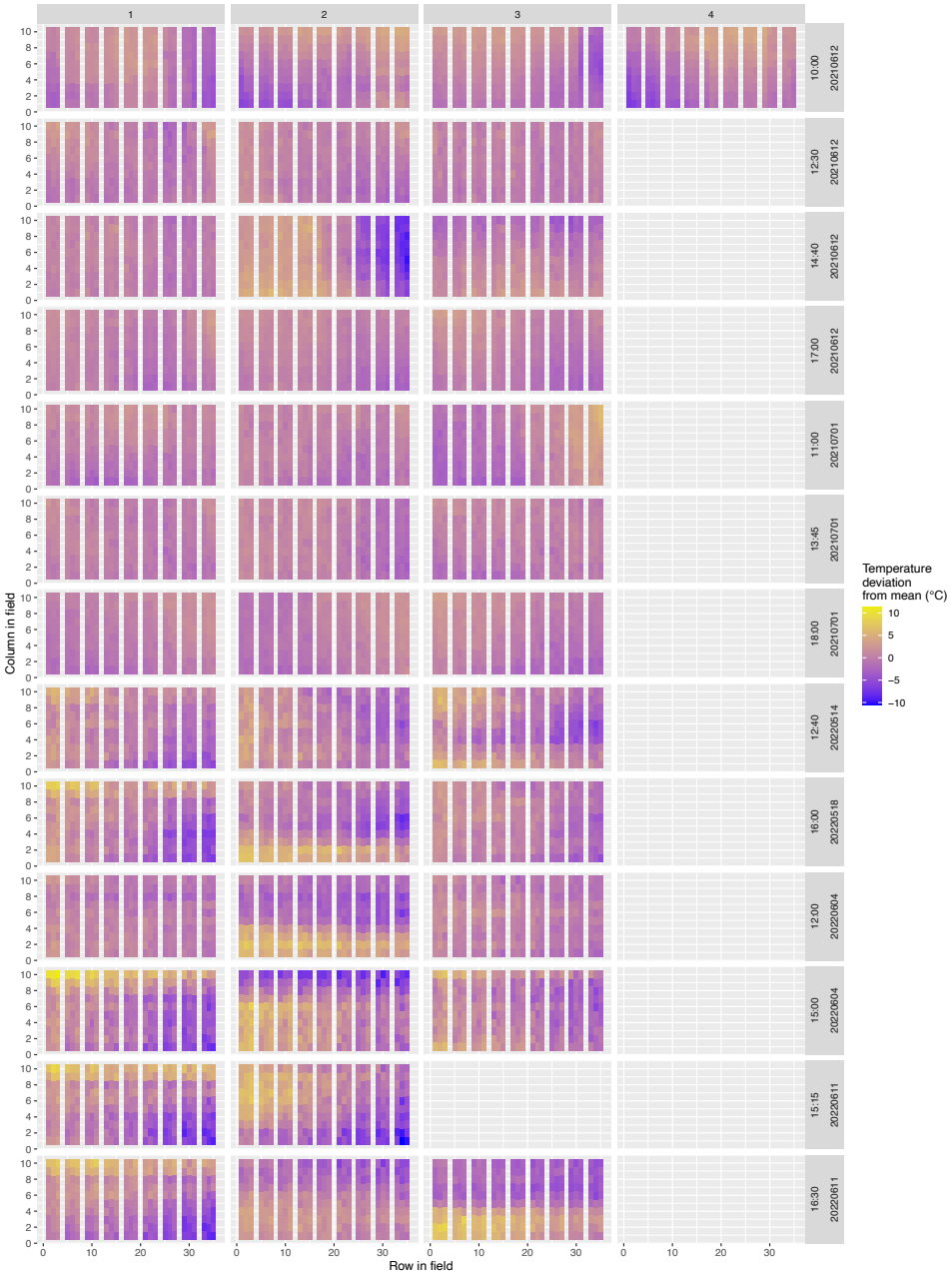


FIGURE S2.9: Unadjusted plot-wise means of EuVar. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.21.0.2 Temporal trend estimation

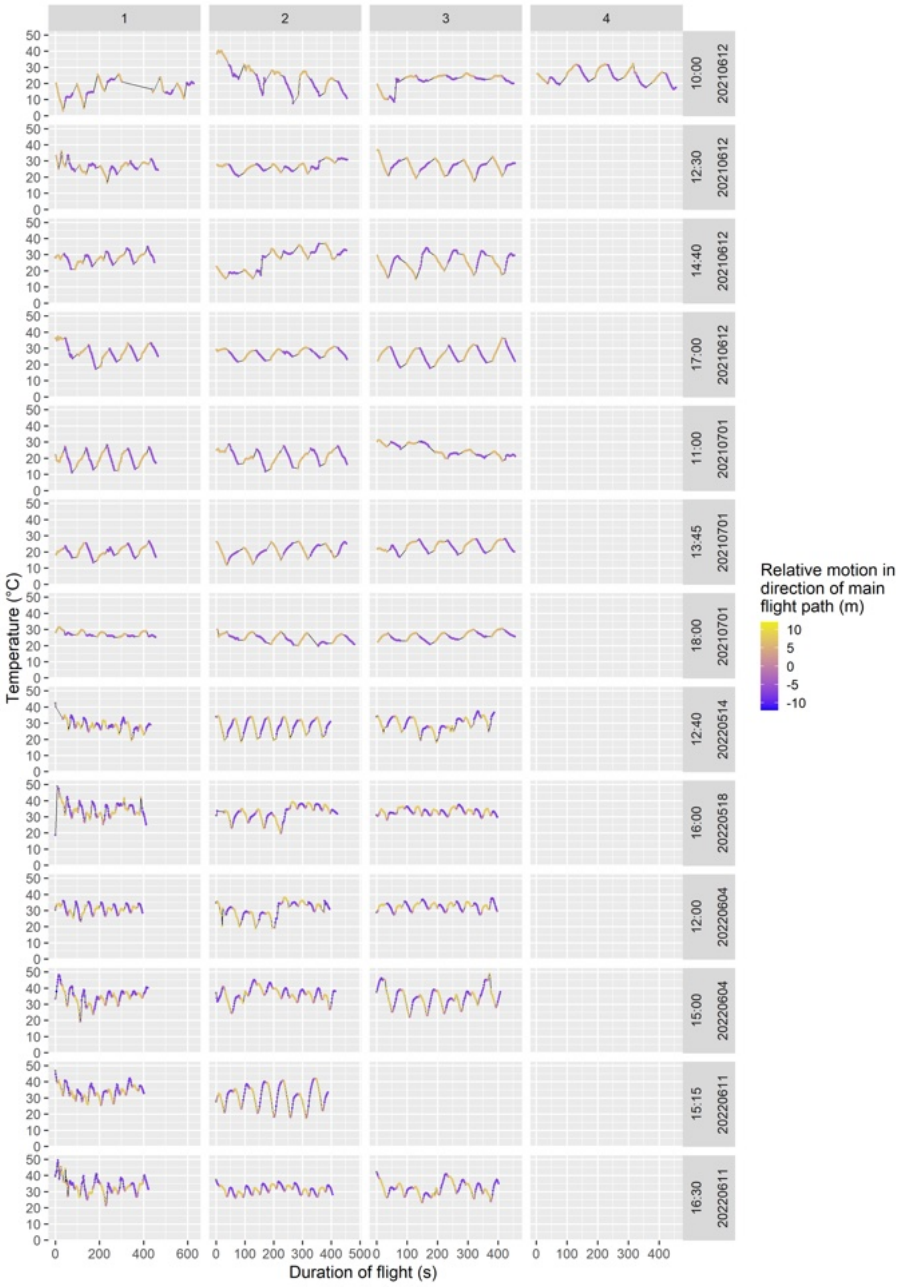


FIGURE S2.10: Estimated thermal drift of TIR measurements throughout the duration of fights for all flights of EuVar. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. The colors indicate the motion in direction of the main fight path. Purple indicates fights in one direction and yellow in the opposite direction of the flight path grid. For gray points, temporal drift was modeled but there was no corresponding measurement of motion along the main fight path.

S2.21.0.3 Temporal correction applied

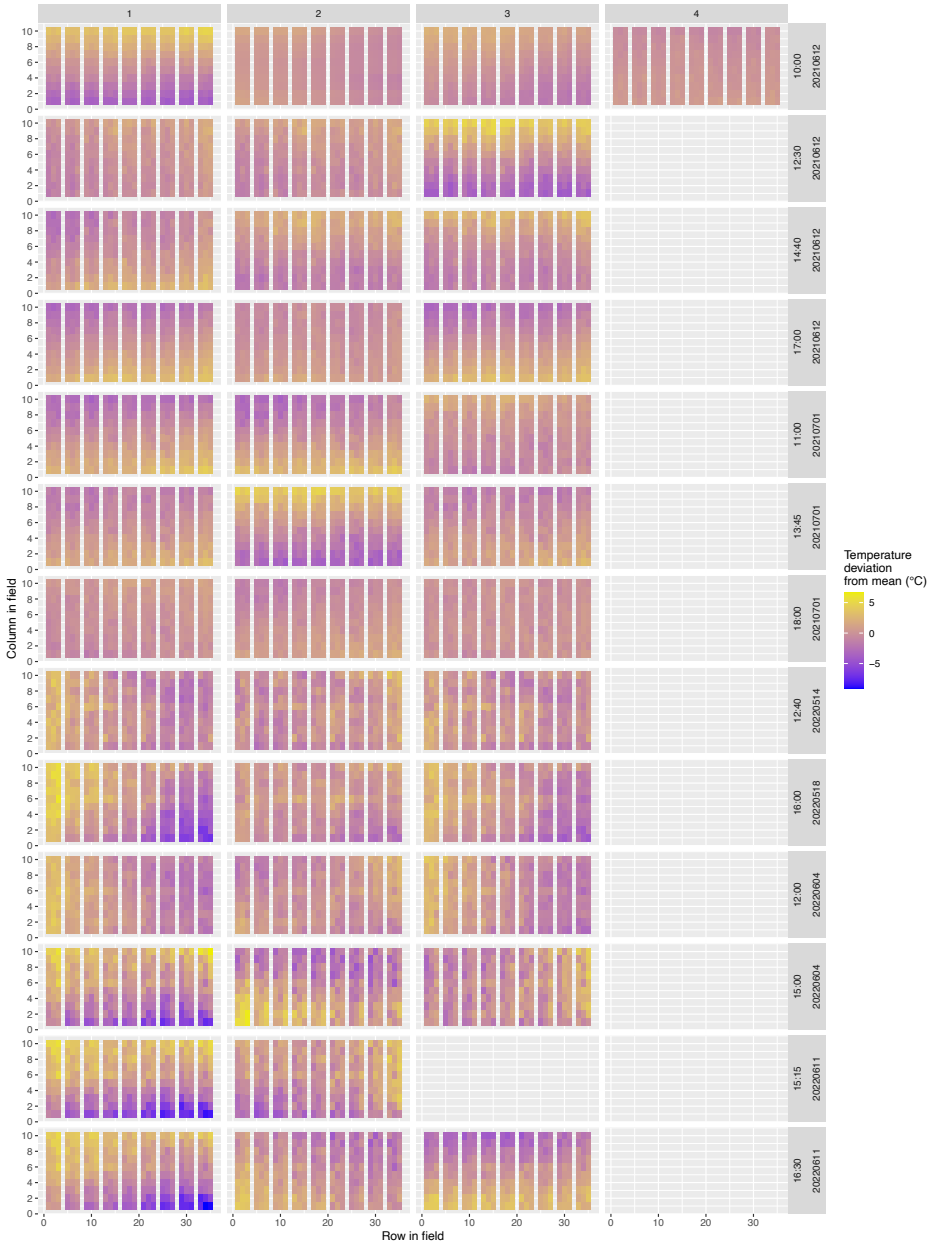


FIGURE S2.11: Adjusted plot-wise means of EuVar after a temporal correction. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.21.0.4 Temporal and spatial correction applied

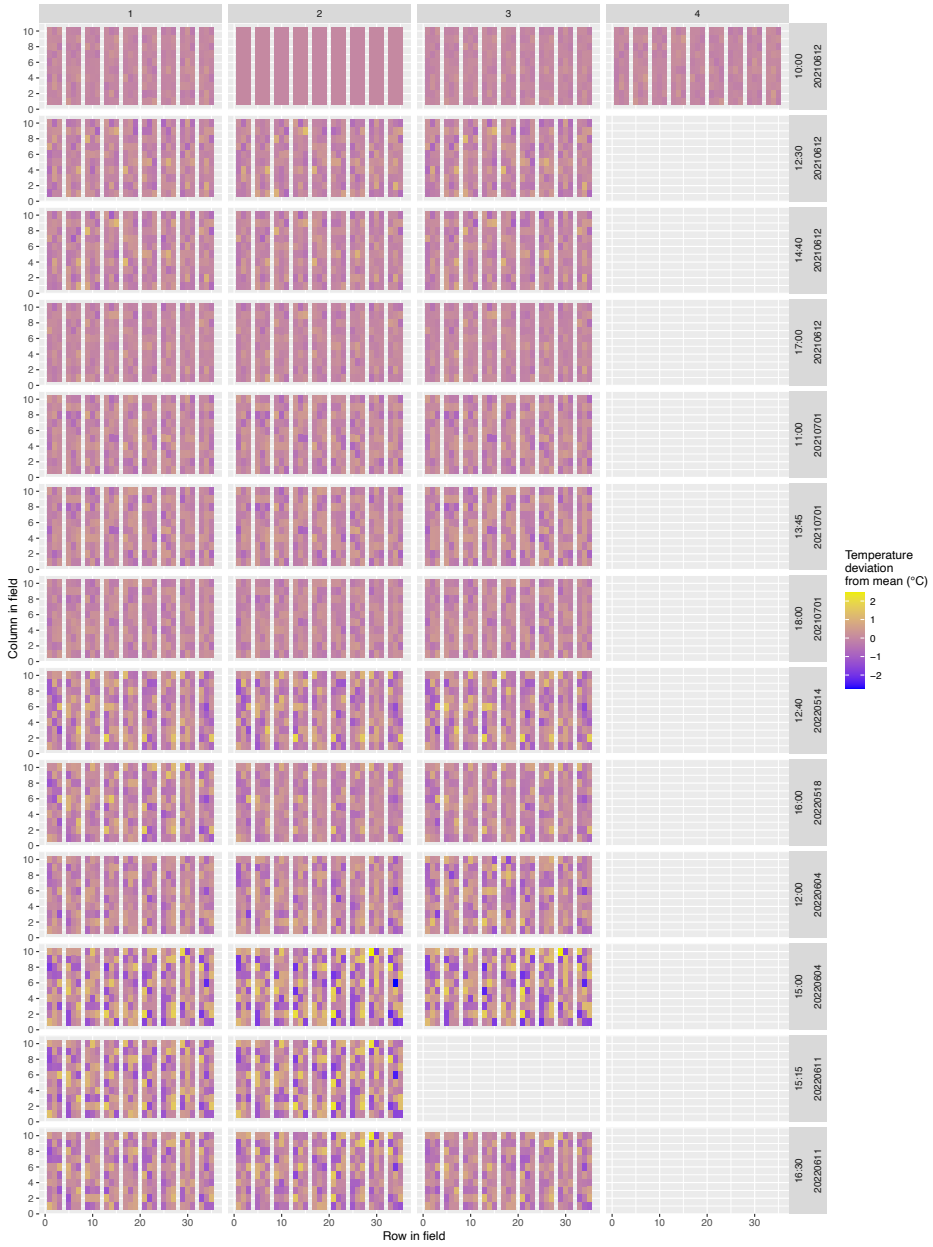


FIGURE S2.12: Adjusted plot-wise means of EuVar after a temporal and spatial correction. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.21.0.5 Genotypic effect (correction applied for effects of trigger timing, field heterogeneity, plots and treatment regimens)

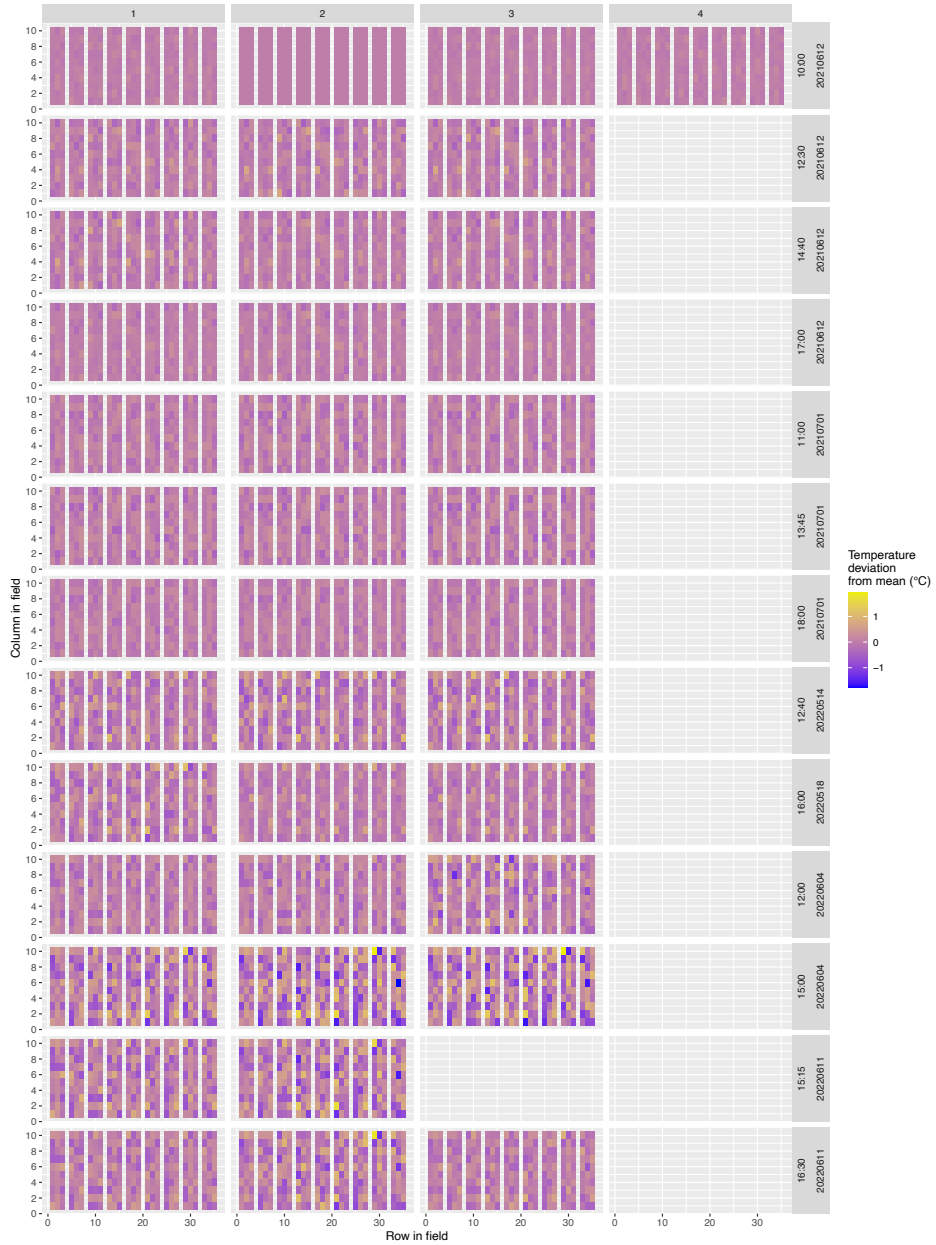


FIGURE S2.13: Estimated effect for the single genotypes and genotype-treatment interactions for all flights flown on EuVar. Temporal correction, spatial correction and treatment deflation were applied. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.21.0.6 Treatment effect (correction applied for effects of trigger timing, field heterogeneity, plots and genotypes)

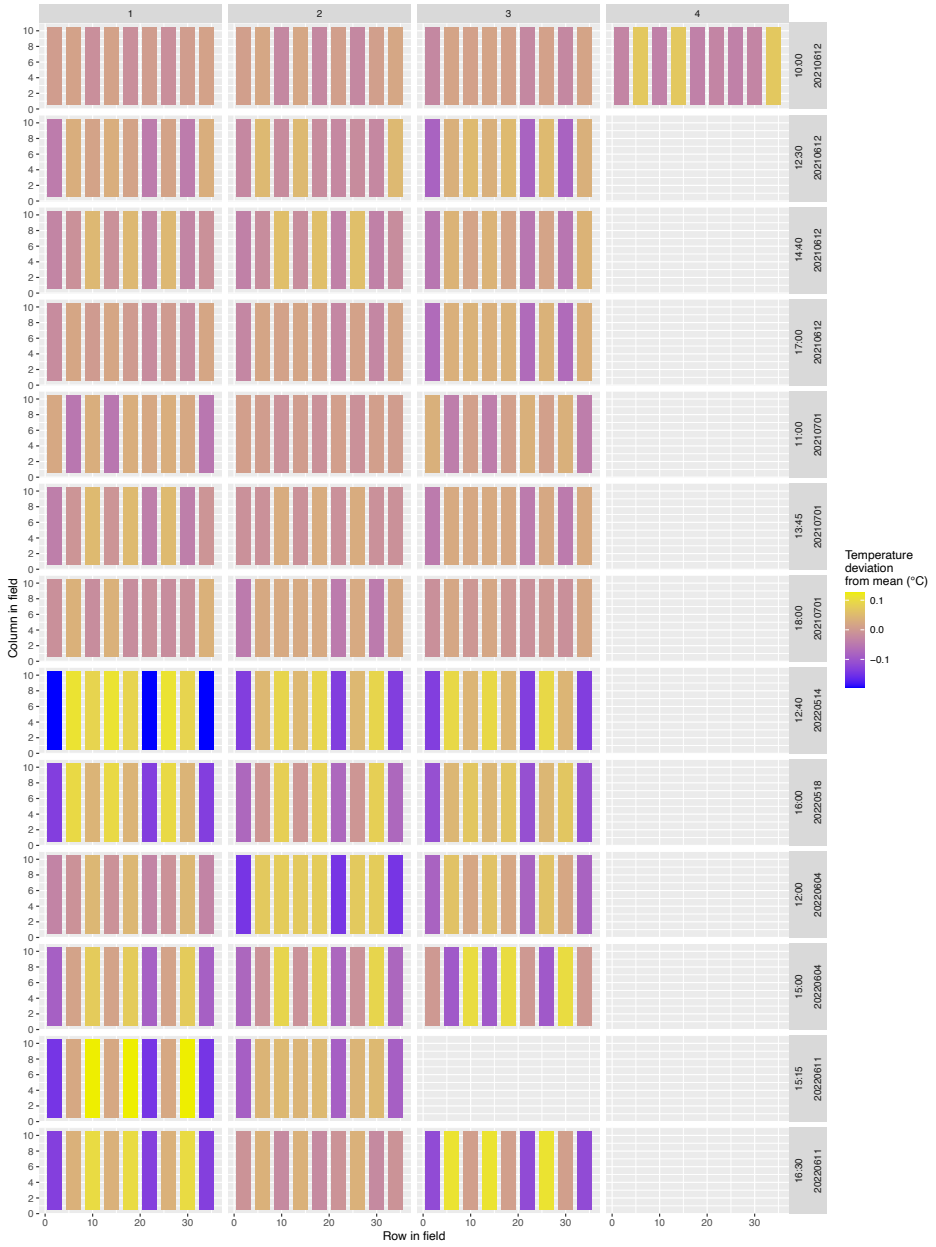


FIGURE S2.14: Estimated effect for the treatment regimens for all flights flown on EuVar. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.22 Correction steps SwiVar

S2.22.0.1 No correction applied - plot-wise means

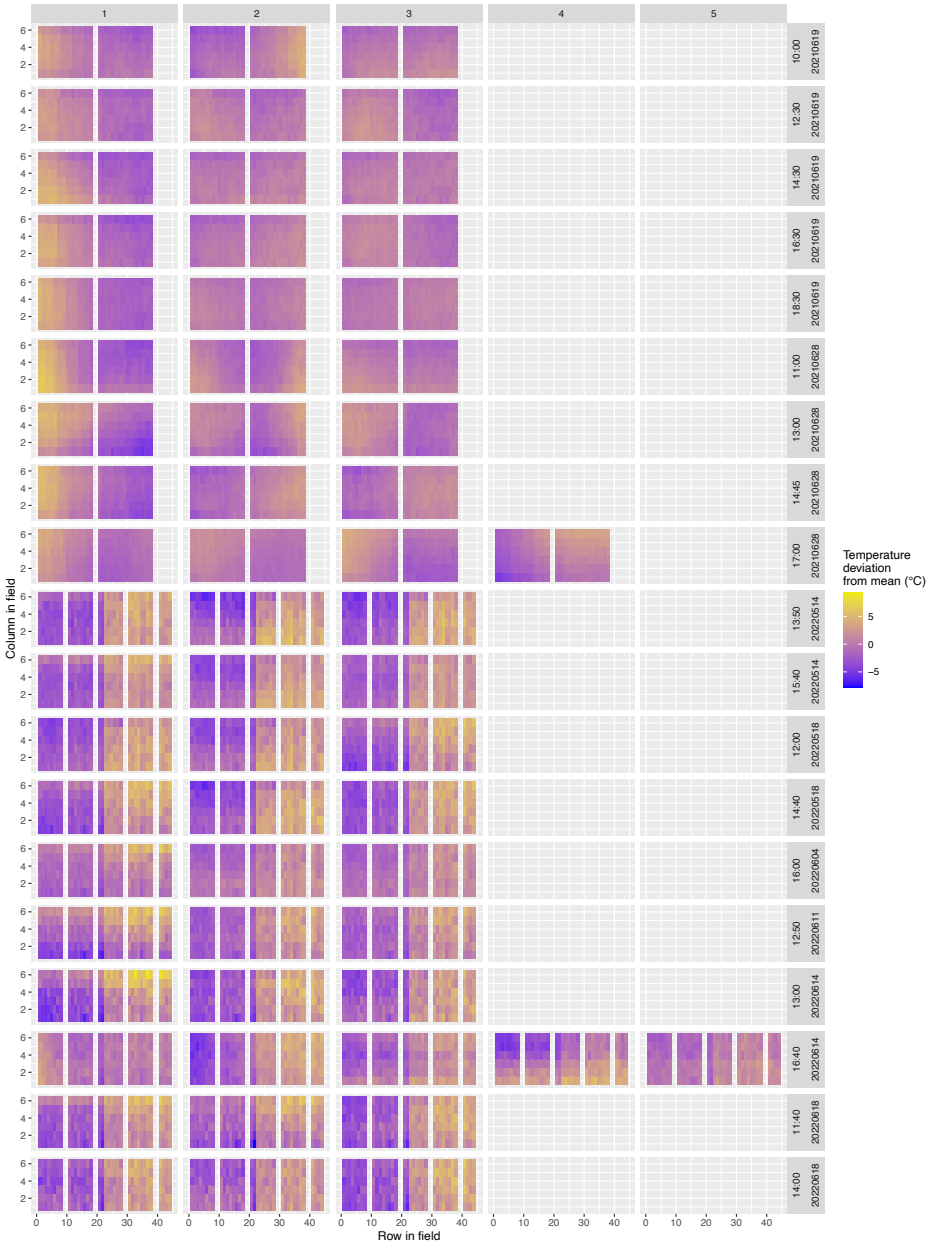


FIGURE S2.15: Unadjusted plot-wise means of SwiVar. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.22.0.2 Temporal trend estimation

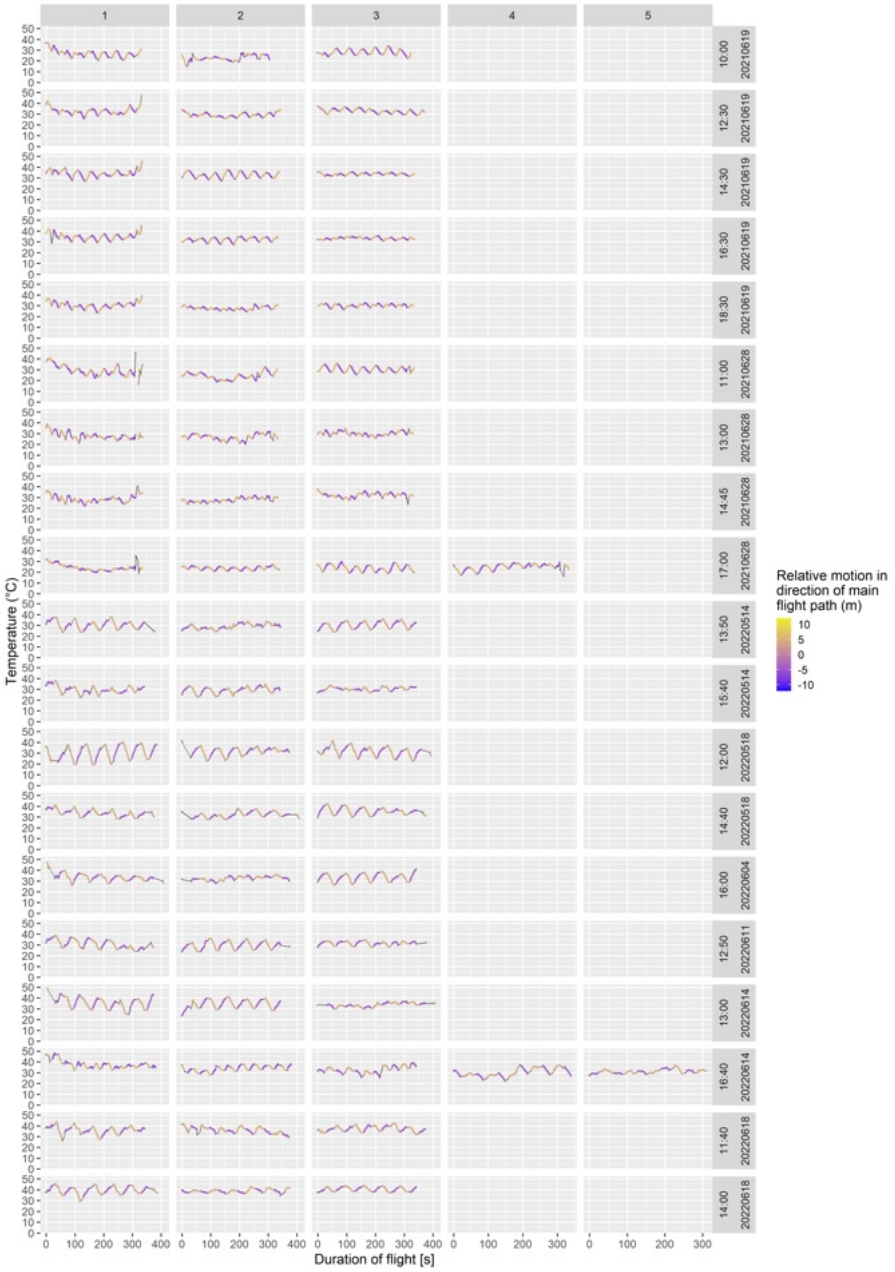


FIGURE S2.16: Estimated thermal drift of TIR measurements throughout the duration of fights for all flights of SwiVar. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. The colors indicate the motion in direction of the main fight path. Purple indicates fights in one direction and yellow in the opposite direction of the flight path grid. For gray points, temporal drift was modeled but there was no corresponding measurement of motion along the main fight path.

S2.22.0.3 Temporal correction applied

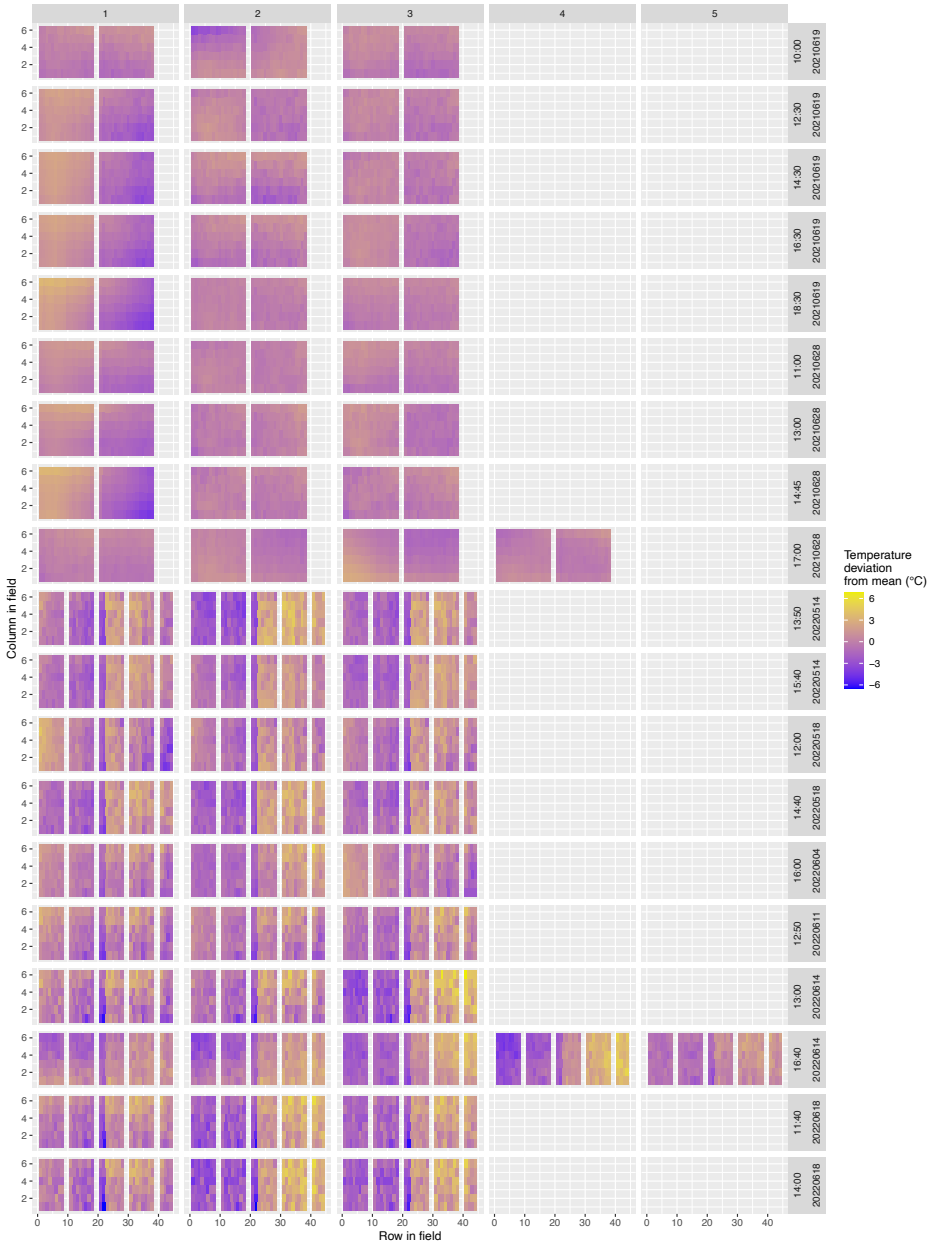


FIGURE S2.17: Adjusted plot-wise means of SwiVar after a temporal correction. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.22.0.4 Temporal and spatial correction applied

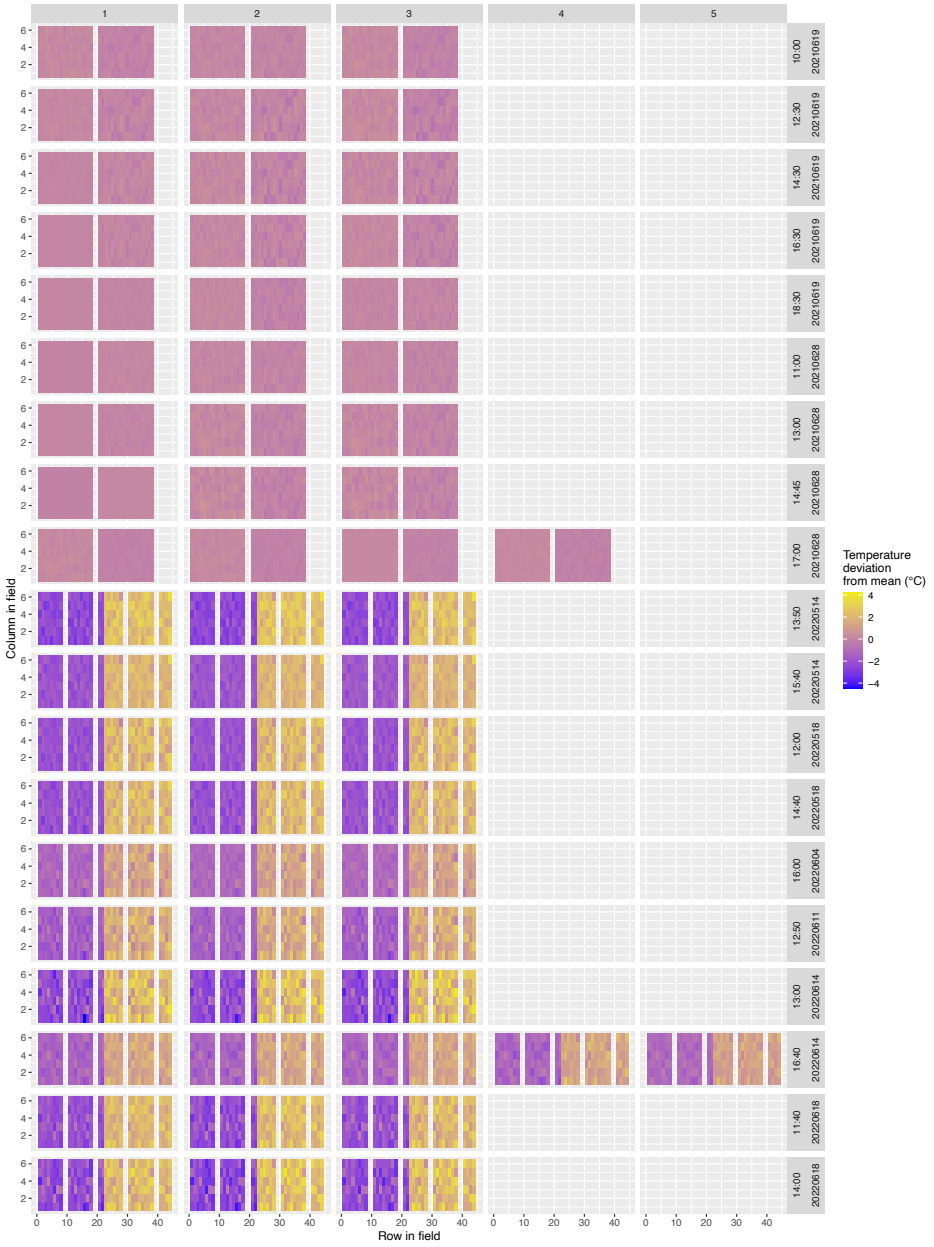


FIGURE S2.18: Adjusted plot-wise means of SwiVar after a temporal and spatial correction. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.22.0.5 Genotypic effect (correction applied for effects of trigger timing, field heterogeneity, plots and treatment regimens)

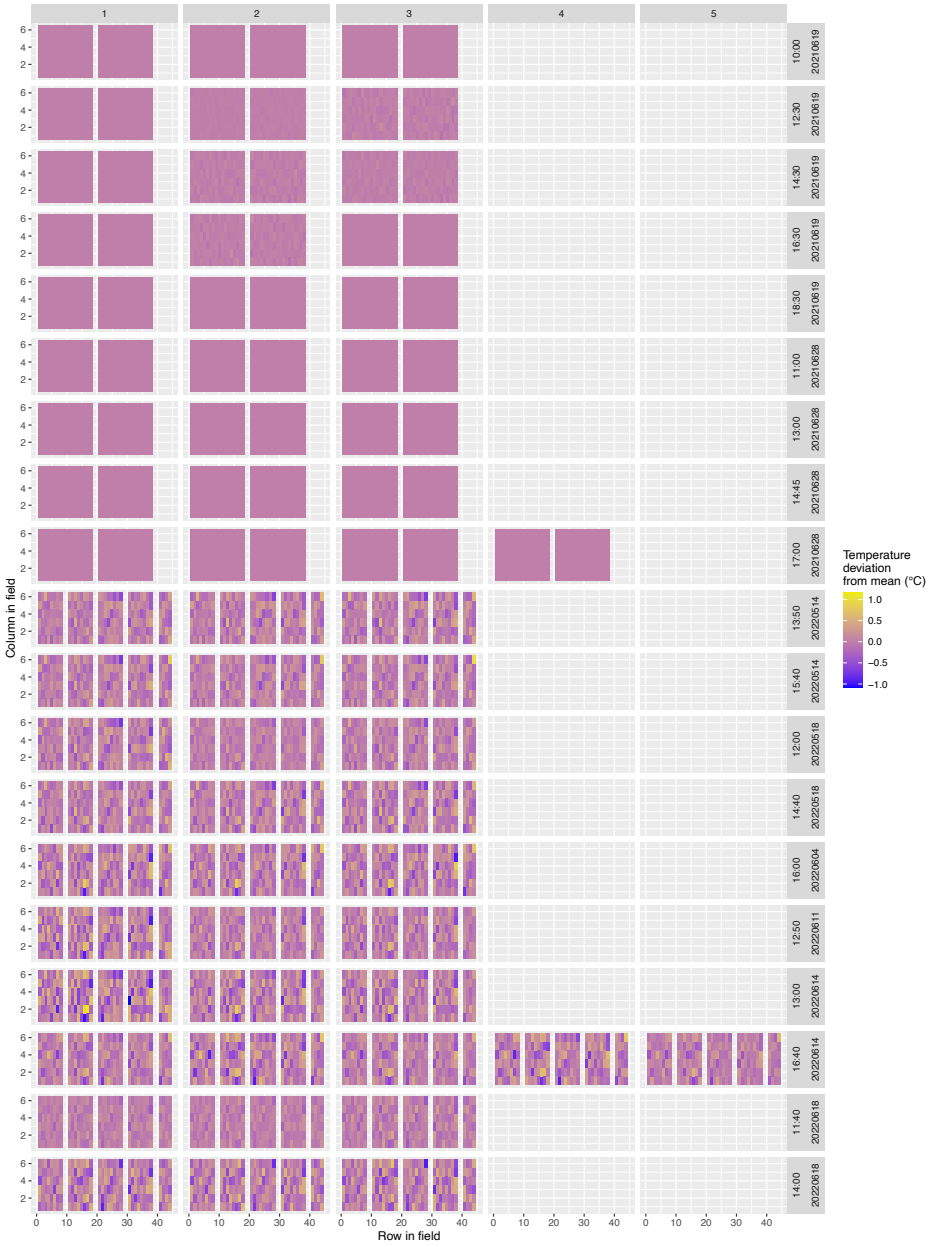


FIGURE S2.19: Estimated effect for the single genotypes and genotype-treatment interactions for all flights flown on SwiVar. Temporal correction, spatial correction and treatment deflation were applied. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.22.0.6 Treatment effect (correction applied for effects of trigger timing, field heterogeneity, plots and genotypes)

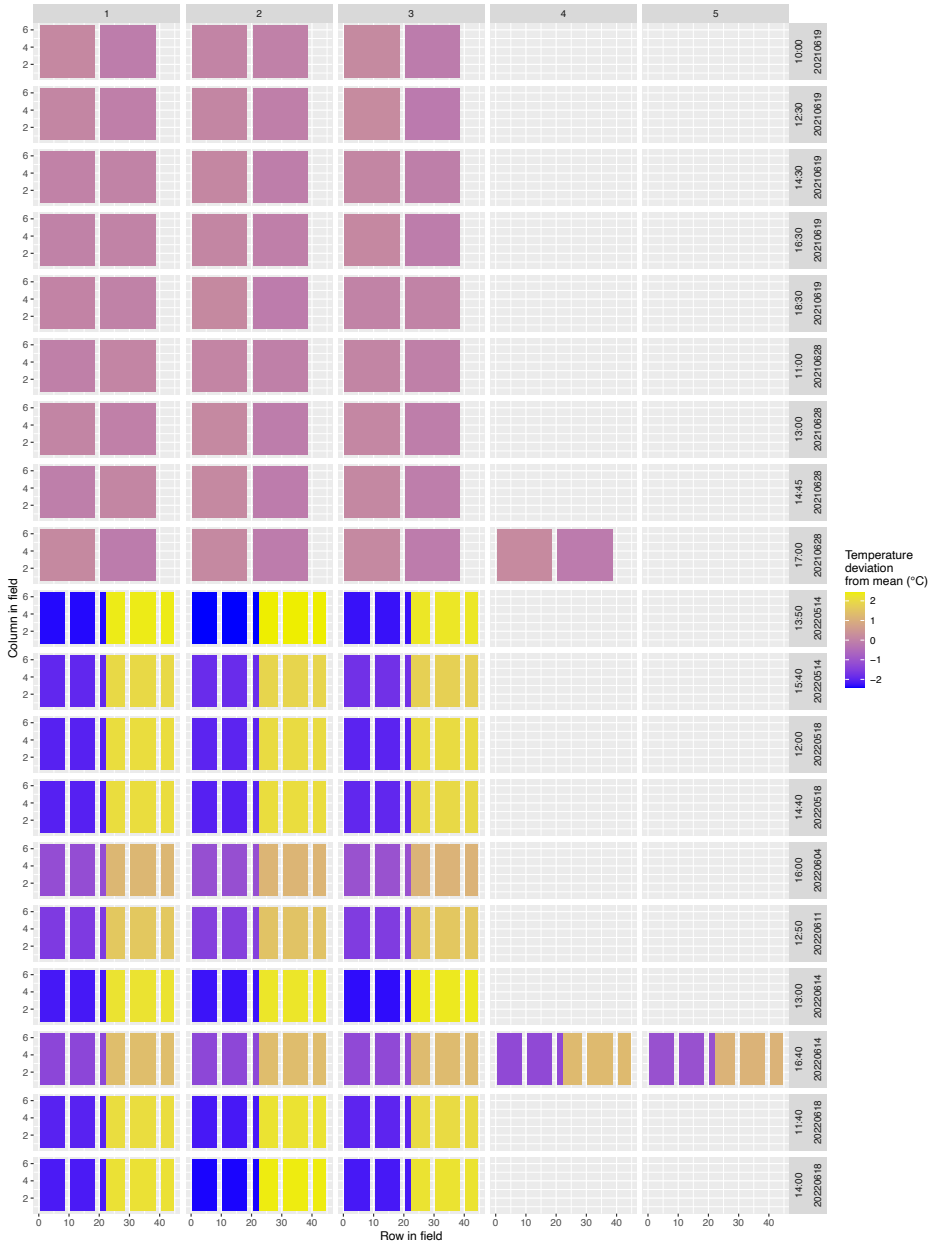


FIGURE S2.20: Estimated effect for the treatment regimens for all flights flown on SwiVar. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.23 Uncorrected phenotypic traits

EuVar21

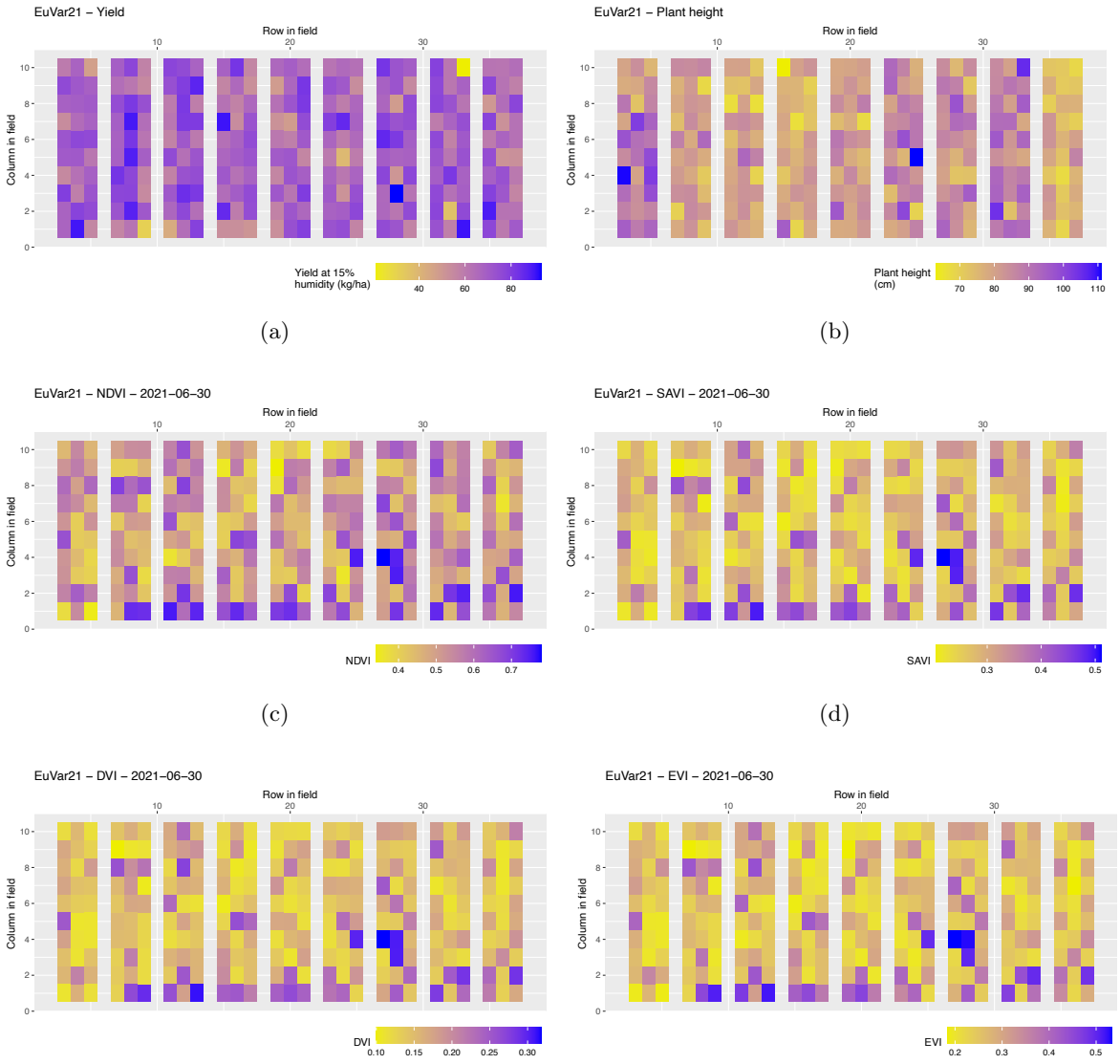
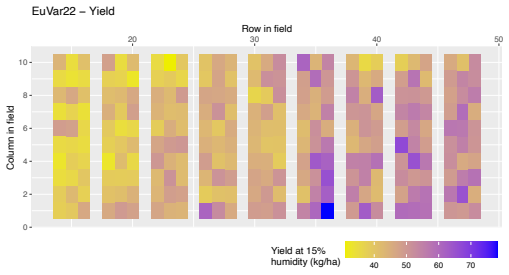
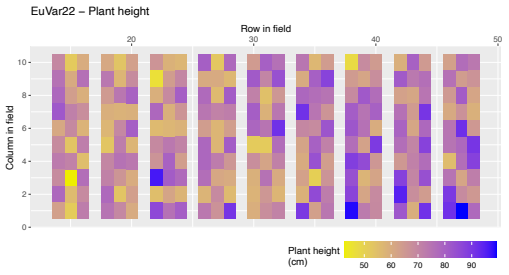


FIGURE S2.21: (a - f) show the uncorrected phenotypic traits of EuVar21. (a): Yield at 15 % water content, (b): Plant height based on five measurements per plot, (c): NDVI multispectral index, (d): SAVI multispectral index, (e): DVI multispectral index, (f): EVI multispectral index.

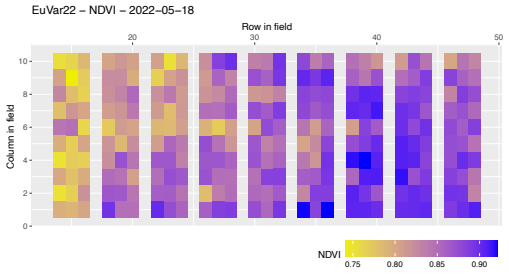
EuVar22



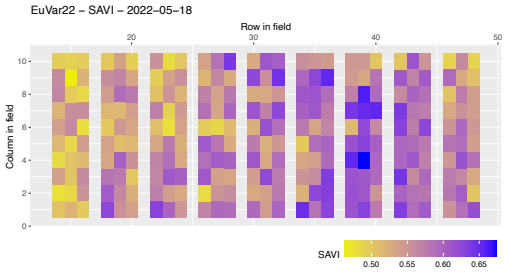
(a)



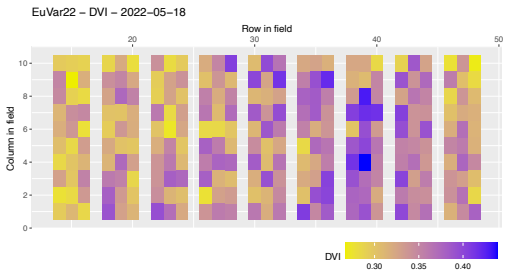
(b)



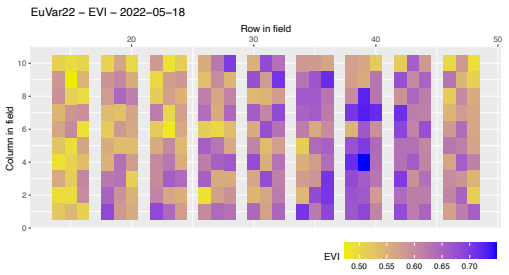
(c)



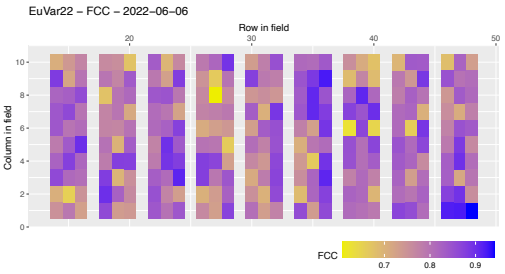
(d)



(e)



(f)



(g)

FIGURE S2.22: (a - g) show the uncorrected phenotypic traits of EuVar22. (a): Yield at 15 % water content, (b): Planth height based on five measurements per plot, (c): NDVI multispectral index, (d): SAVI multispectral index, (e): DVI multispectral index, (f): EVI multispectral index, (g): FCC.

SwiVar21

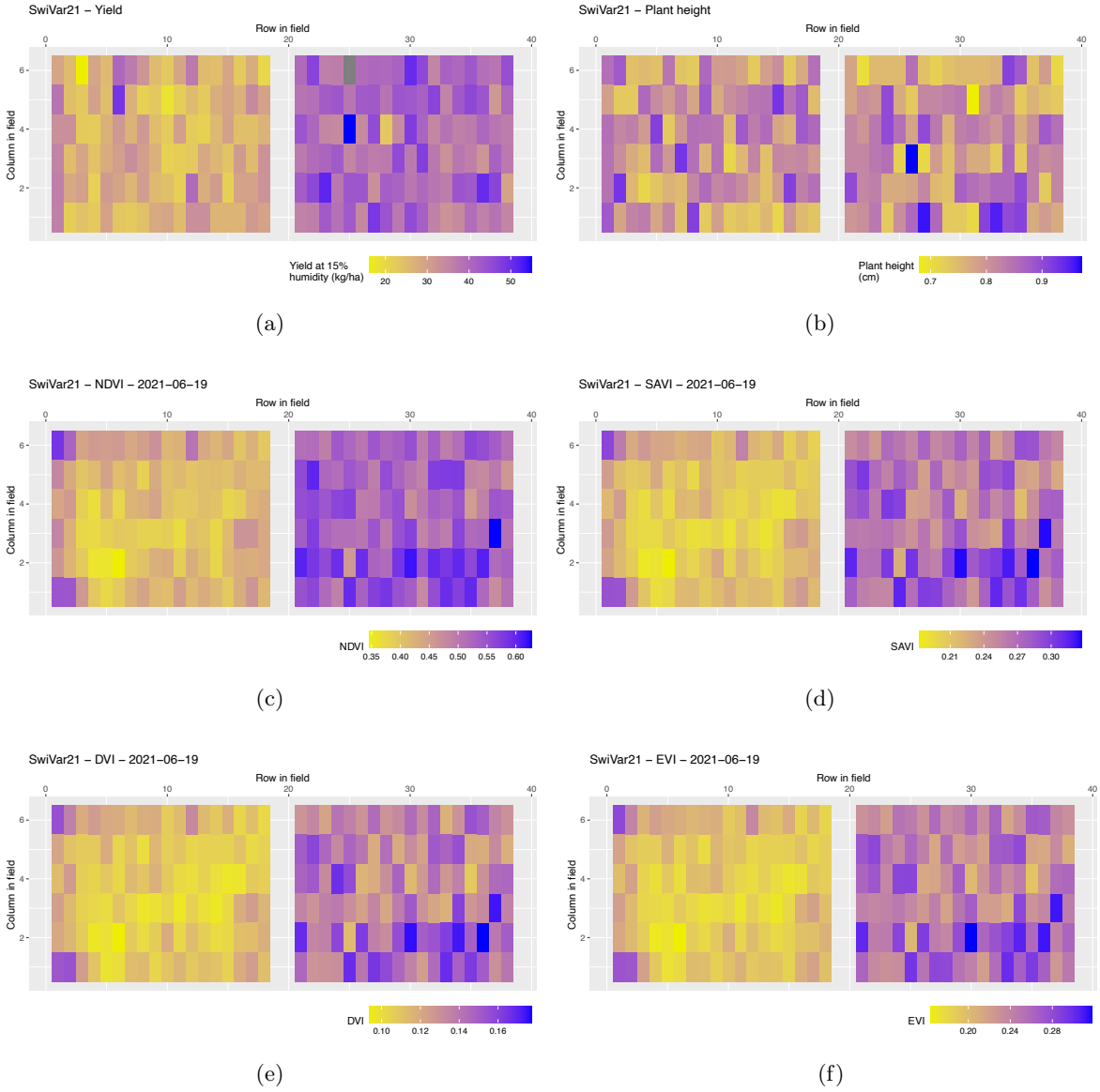
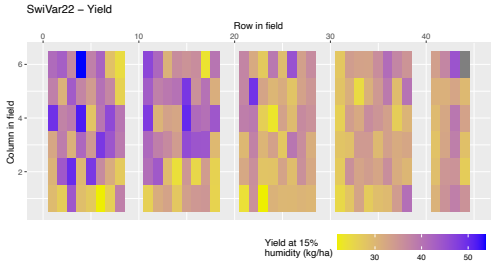
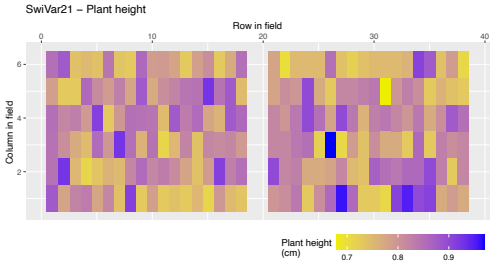


FIGURE S2.23: (a - f) show the uncorrected phenotypic traits of SwiVar21. (a): Yield at 15 % water content, (b): Planth height based on five measurements per plot, (c): NDVI multispectral index, (d): SAVI multispectral index, (e): DVI multispectral index, (f): EVI multispectral index.

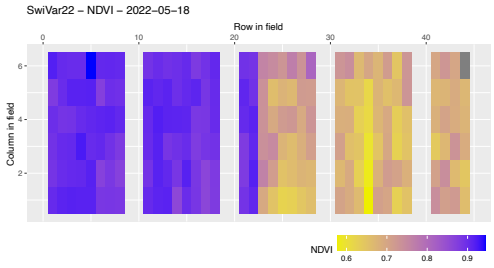
SwiVar22



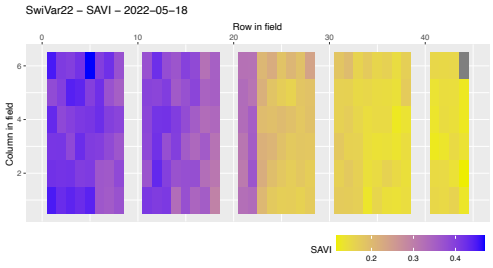
(a)



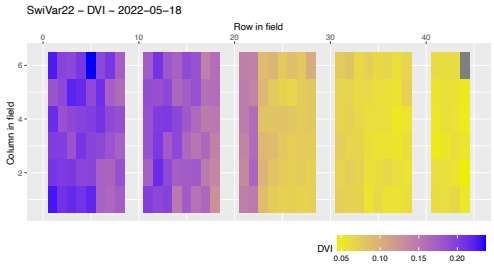
(b)



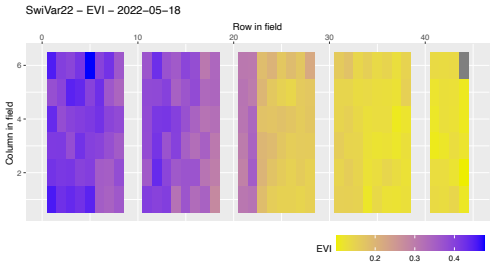
(c)



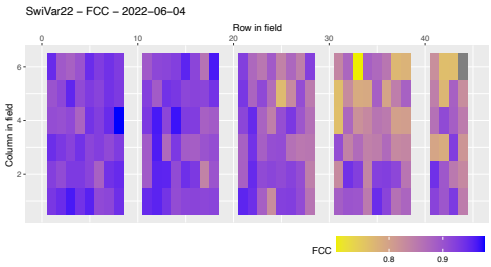
(d)



(e)



(f)



(g)

FIGURE S2.24: (a - g) show the uncorrected phenotypic traits of SwiVar22. (a): Yield at 15 % water content, (b): Planth height based on five measurements per plot, (c): NDVI multispectral index, (d): SAVI multispectral index, (e): DVI multispectral index, (f): EVI multispectral index, (g): FCC. On the plot in grey, a sowing error occurred and the plot was excluded from analysis.

S2.24 Correlation with yield based on campaign-wise CT estimates

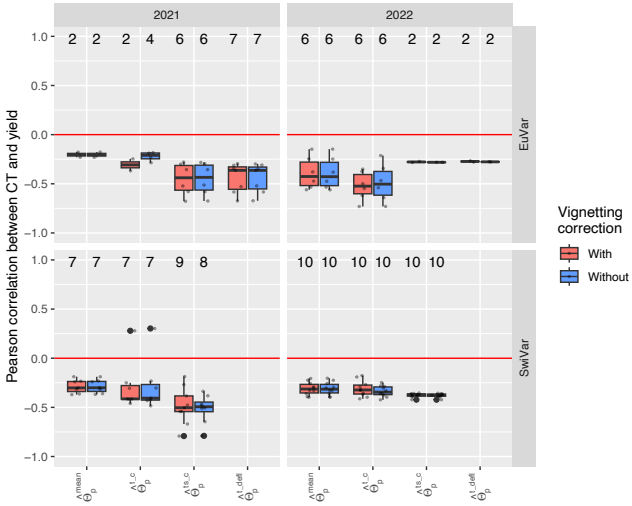


FIGURE S2.25: The CT estimates based on all flights within campaigns without and with vignetting correction were correlated to yield. Just correlations significant at $p \leq 0.01$ are shown. The number above the boxplots indicates the number of campaigns with significant correlations included in the respective box plots.

S2.25 Spatial trend estimates

S2.25.1 Spatial trend estimation EuVar - individual flights

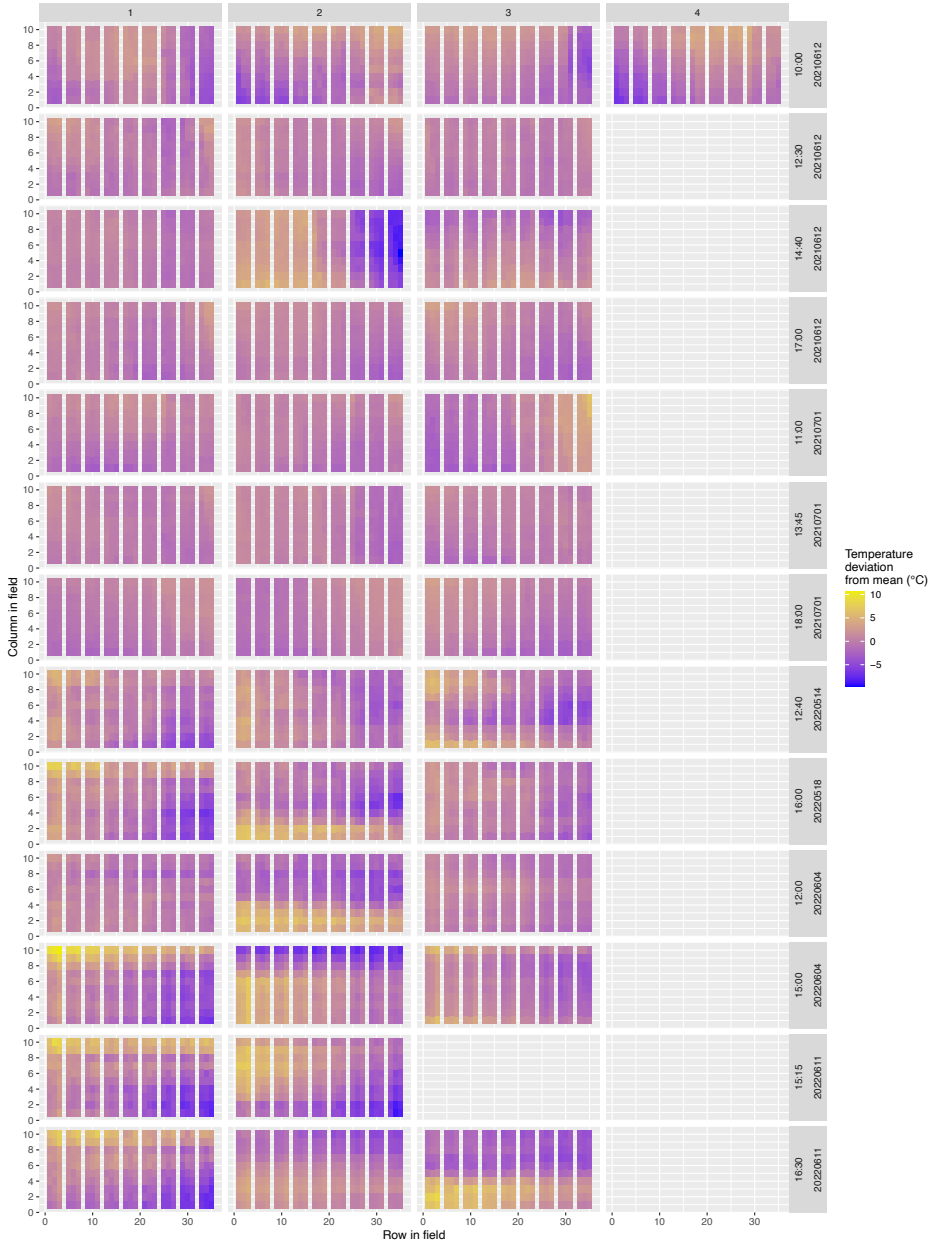


FIGURE S2.26: Spatial trend estimates for single flights based on Eq. 3.10 for EuVar. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.25.2 Spatial trend estimation SwiVar - individual flights

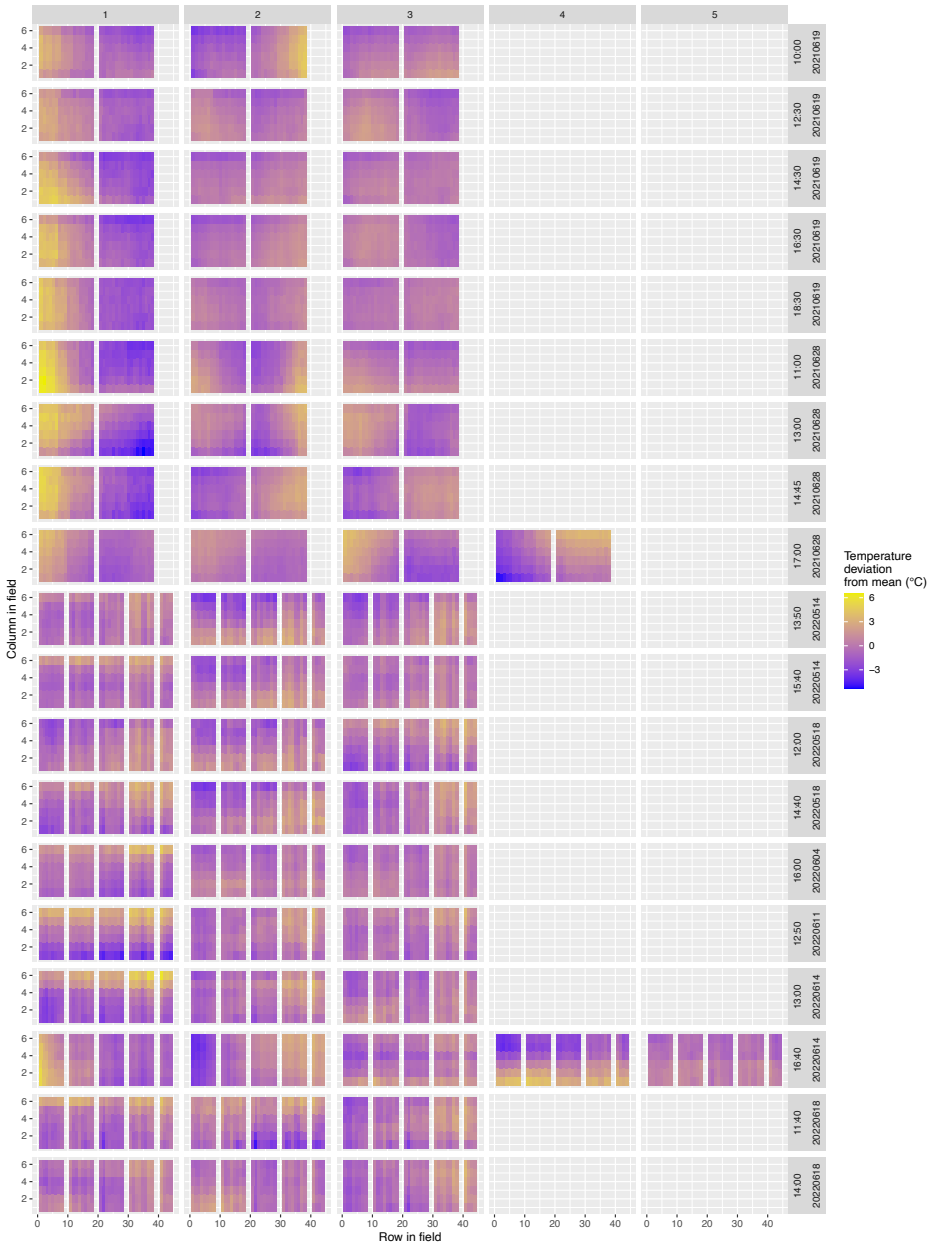


FIGURE S2.27: Spatial trend estimates for single flights based on Eq. 3.10 for SwiVar. Flights are horizontally grouped by dates and flight times. Each row corresponds to a campaign. Columns indicate the flight order within campaigns. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.26 Correlation of plot-wise estimates of spatial field trends for individual flights - EuVar 2021

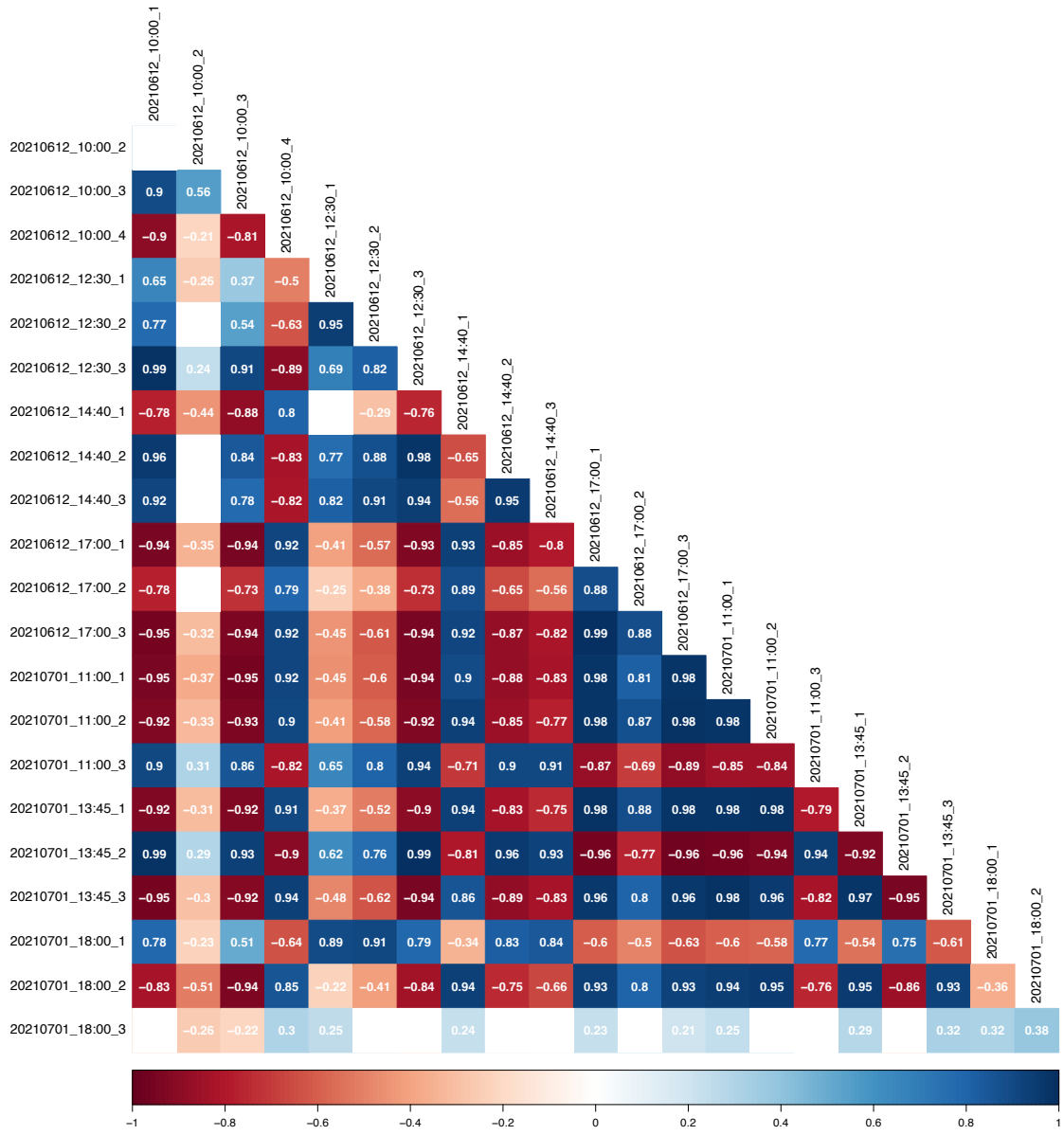


FIGURE S2.28: Pearson correlation of plot-wise estimates of spatial trends for EuVar21 which are shown in Fig. S2.26. Just correlations significant at $p \leq 0.001$ are shown.

S2.27 Correlation of plot-wise estimates of spatial field trends for individual flights - EuVar 2022

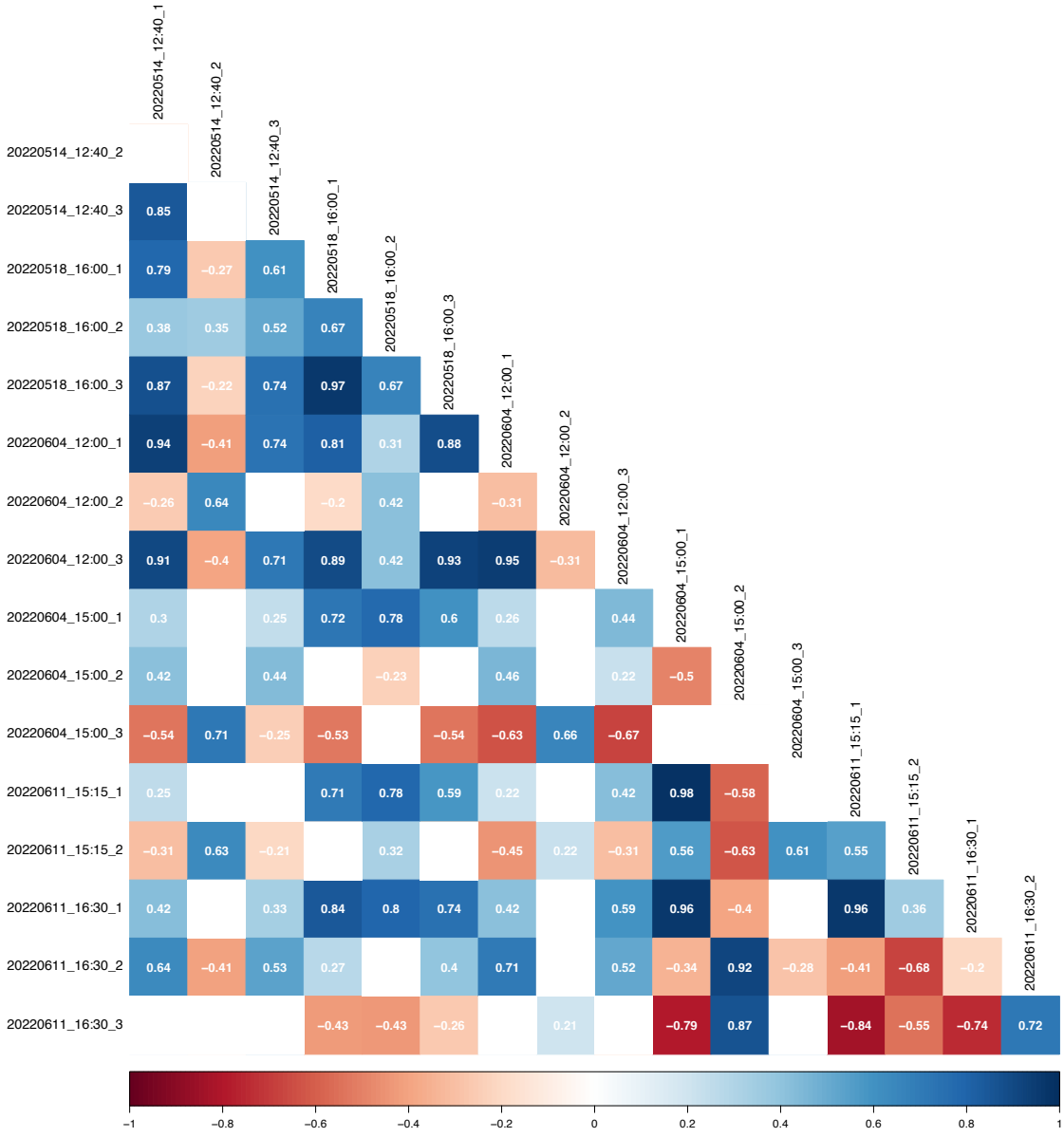


FIGURE S2.29: Pearson correlation of plot-wise estimates of spatial trends for EuVar22 which are shown in Fig. S2.26. Just correlations significant at $p \leq 0.001$ are shown.

S2.28 Correlation of plot-wise estimates of spatial field trends for individual flights - SwiVar 2021

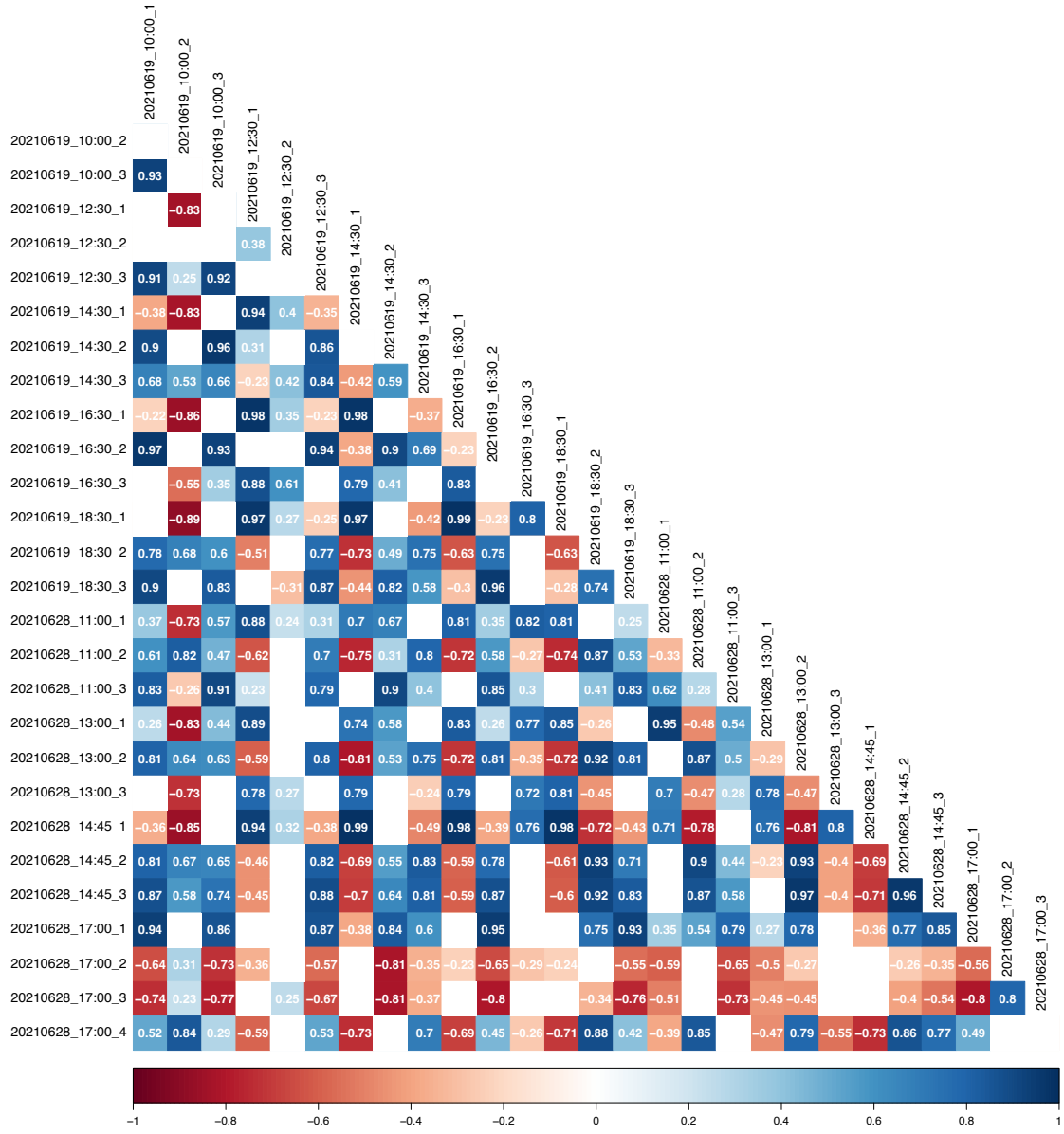


FIGURE S2.30: Pearson correlation of plot-wise estimates of spatial trends for SwiVar21 which are shown in Fig. S2.27. Just correlations significant at $p \leq 0.001$ are shown.

S2.29 Correlation of plot-wise estimates of spatial field trends for individual flights - SwiVar 2022

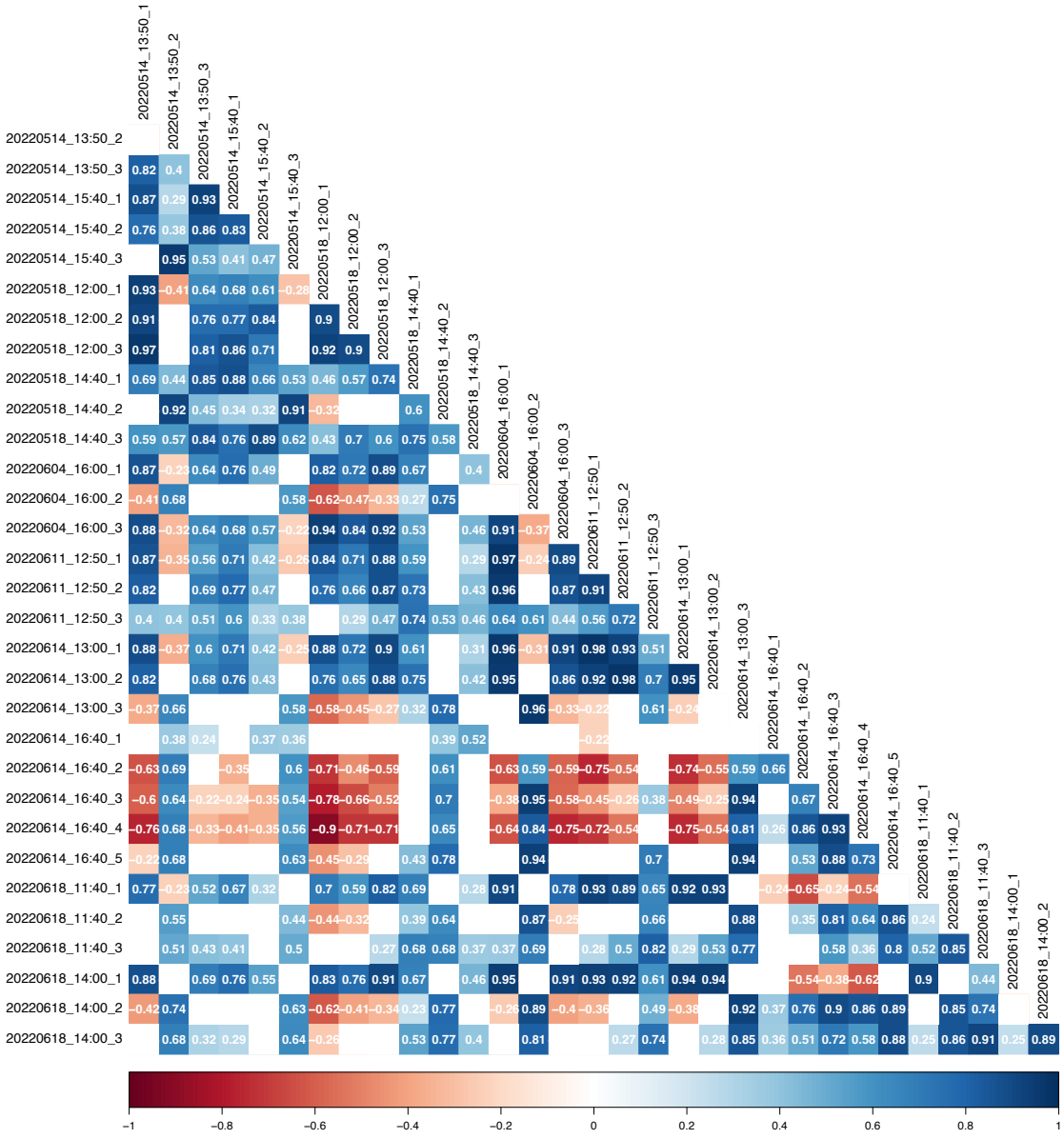


FIGURE S2.31: Pearson correlation of plot-wise estimates of spatial trends for SwiVar22 which are shown in Fig. S2.27. Just correlations significant at $p \leq 0.001$ are shown.

S2.30 CT differences from mean, arranged by flag leaf rolling ratings - EuVar22

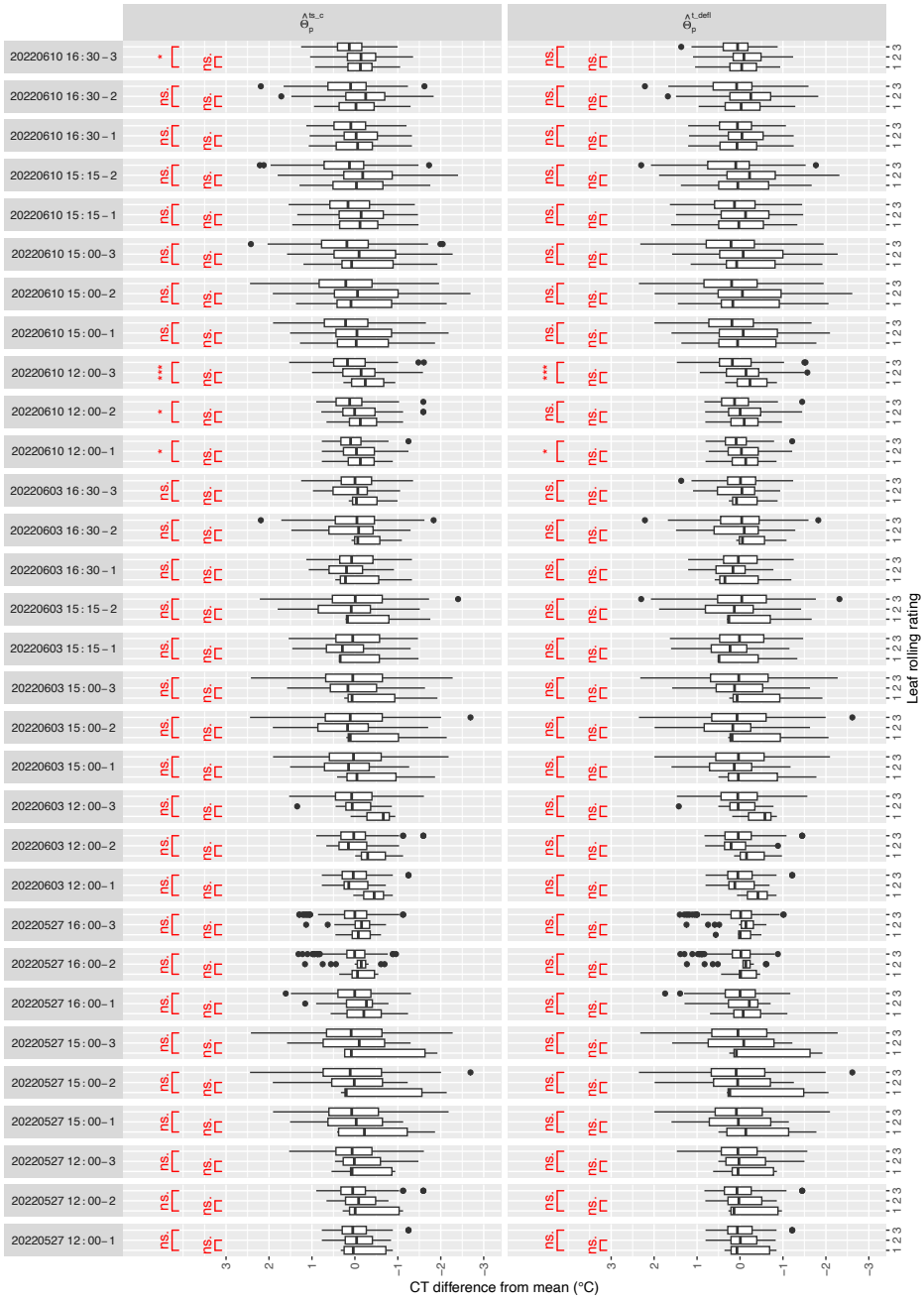


FIGURE S2.32: Corrected CT estimates of were grouped according to their flag leaf rolling score for EuVar22 before ($\hat{\theta}_p^{ts-c}$) and after ($\hat{\theta}_p^{t-defl}$) applying a treatment deflation on CT estimates. Significance levels: ns: $p > 0.05$; *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$.

S2.31 CT differences from mean, arranged by flag leaf rolling ratings - SwiVar22 Part I

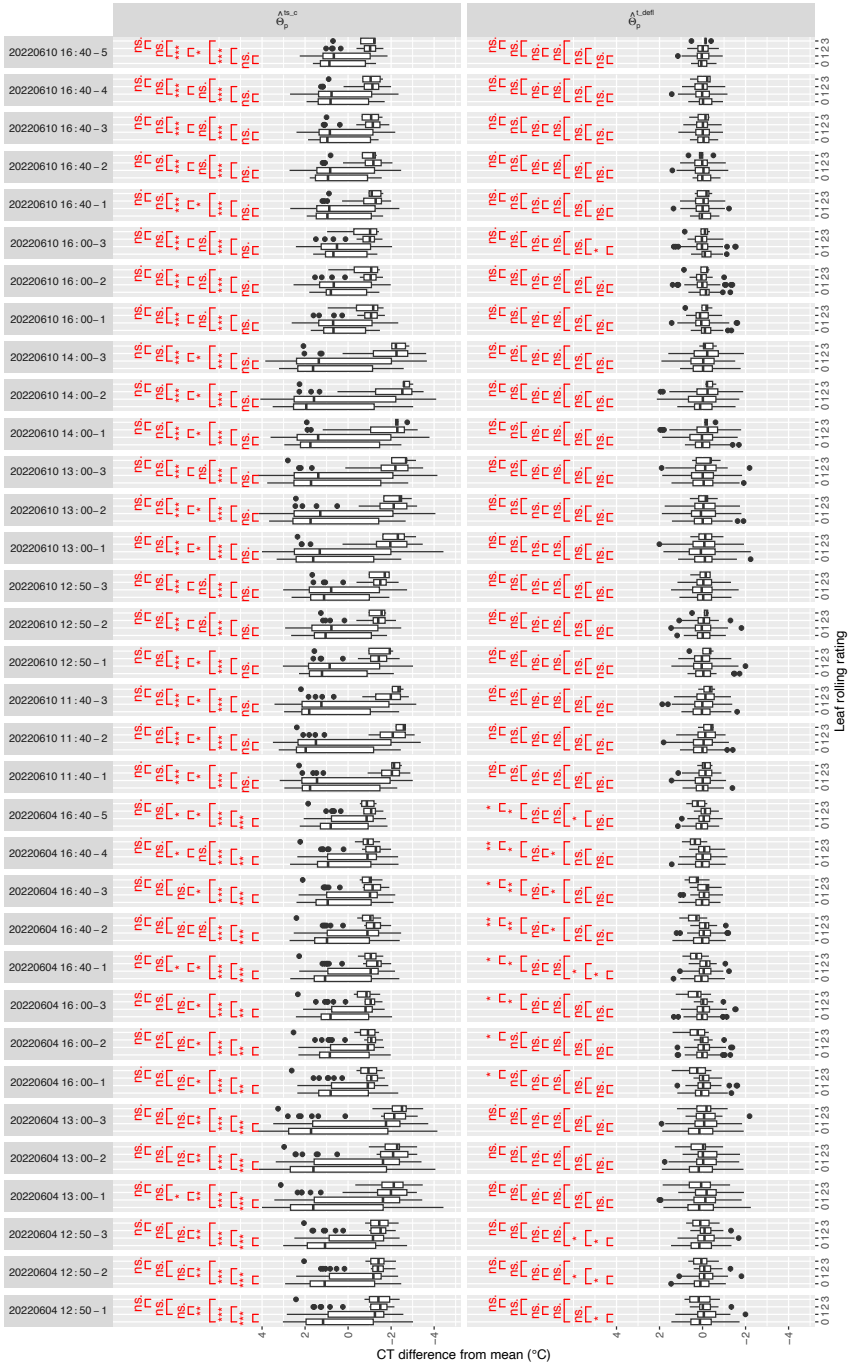


FIGURE S2.33: Corrected CT estimates were grouped according to their flag leaf rolling score for SwiVar22 (2022-06-04 & 2022-06-10) before ($\hat{\theta}_p^{ts-c}$) and after ($\hat{\theta}_p^{t-defl}$) applying a treatment deflation on CT estimates. Significance levels: ns: $p > 0.05$; *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$.

S2.32 CT differences from mean, arranged by flag leaf rolling ratings - SwiVar22 Part II

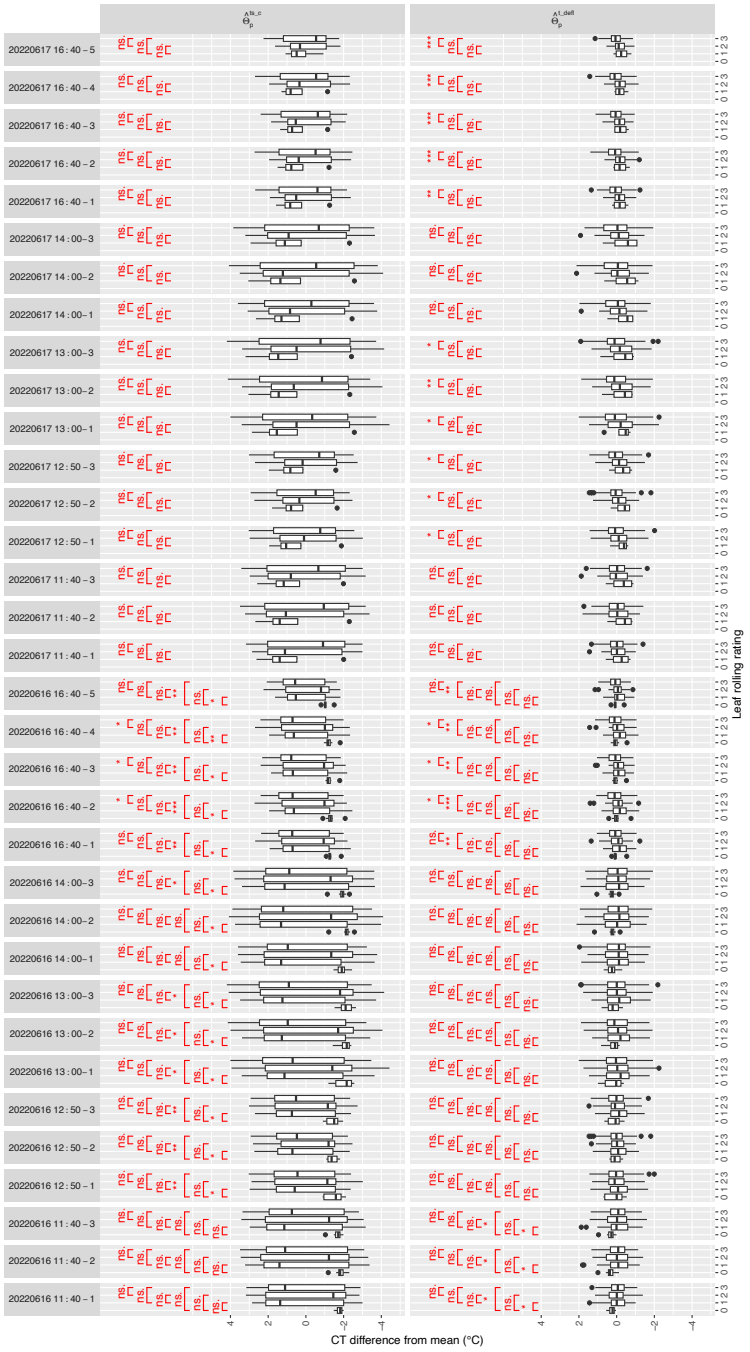


FIGURE S2.34: Corrected CT estimates were grouped according to their flag leaf rolling score for SwiVar22 (2022-06-16 & 2022-06-17) before ($\hat{\theta}_p^{ts-c}$) and after ($\hat{\theta}_p^{t-defl}$) applying a treatment deflation on CT estimates. Significance levels: ns: $p > 0.05$; *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$.

S2.33 Flag leaf rolling ratings of SwiVar22 on 2022-06-10

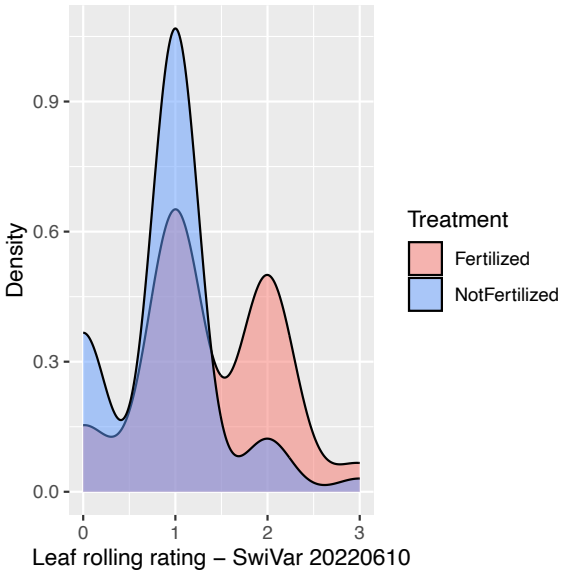
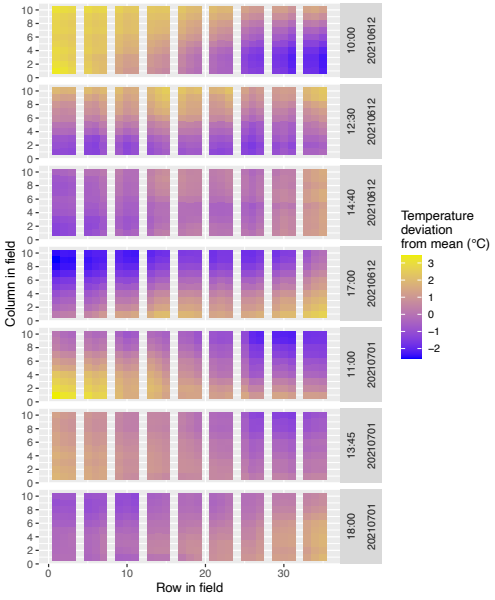
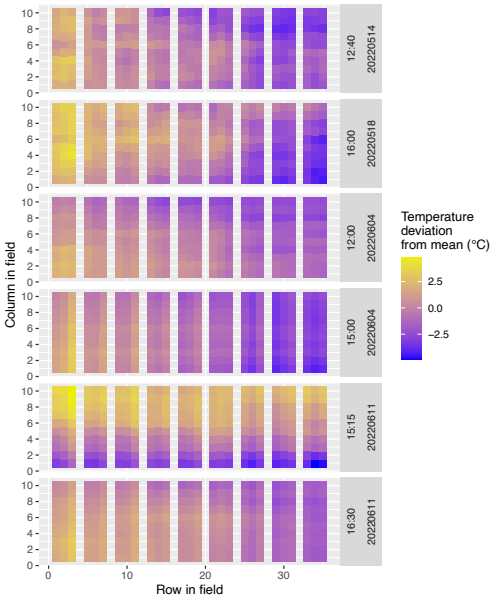


FIGURE S2.35: Flag leaf rating density distribution for SwiVar on 2022-06-10 for the two treatments "Fertilized" and "Not fertilized".

S2.34 Campaign-wise spatial trends EuVar21 & EuVar22



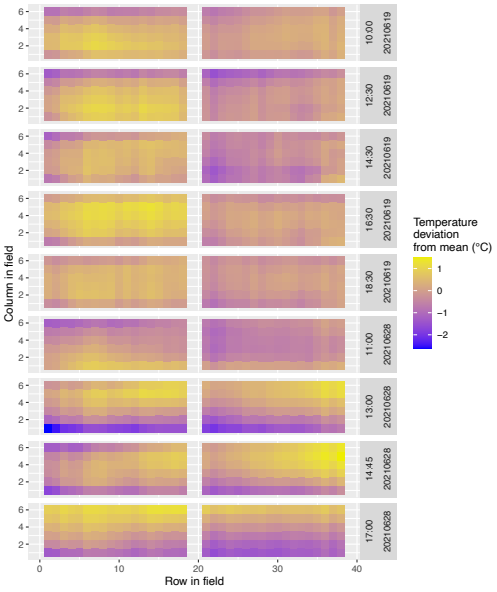
(a) .



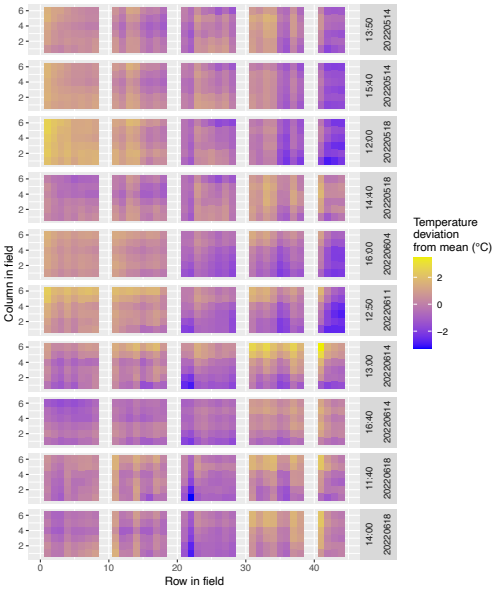
(b)

FIGURE S2.36: Spatial trend estimates for campaigns, based on Eq. 3.10 for EuVar21 (a) and EuVar22 (b) when processing multiple flights of a campaign with the same mixed model. Flights are horizontally grouped by dates and flight times. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.35 Campaign-wise spatial trends SwiVar21 & SwiVar22



(a) .



(b)

FIGURE S2.37: Spatial trend estimates for campaigns, based on Eq. 3.10 for SwiVar21 (a) and SwiVar22 (b) when processing multiple flights of a campaign with the same mixed model. Flights are horizontally grouped by dates and flight times. "Column in field" and "Row in field" indicate the spatial position of the plot in in the field where column increases along the tractor track direction. To allow for a meaningful representation of contrasting temperature ranges, flight-wise temperature deviations from flight-wise mean values are shown.

S2.36 Detailed correlation charts of campaign-wise spatial trends - EuVar21

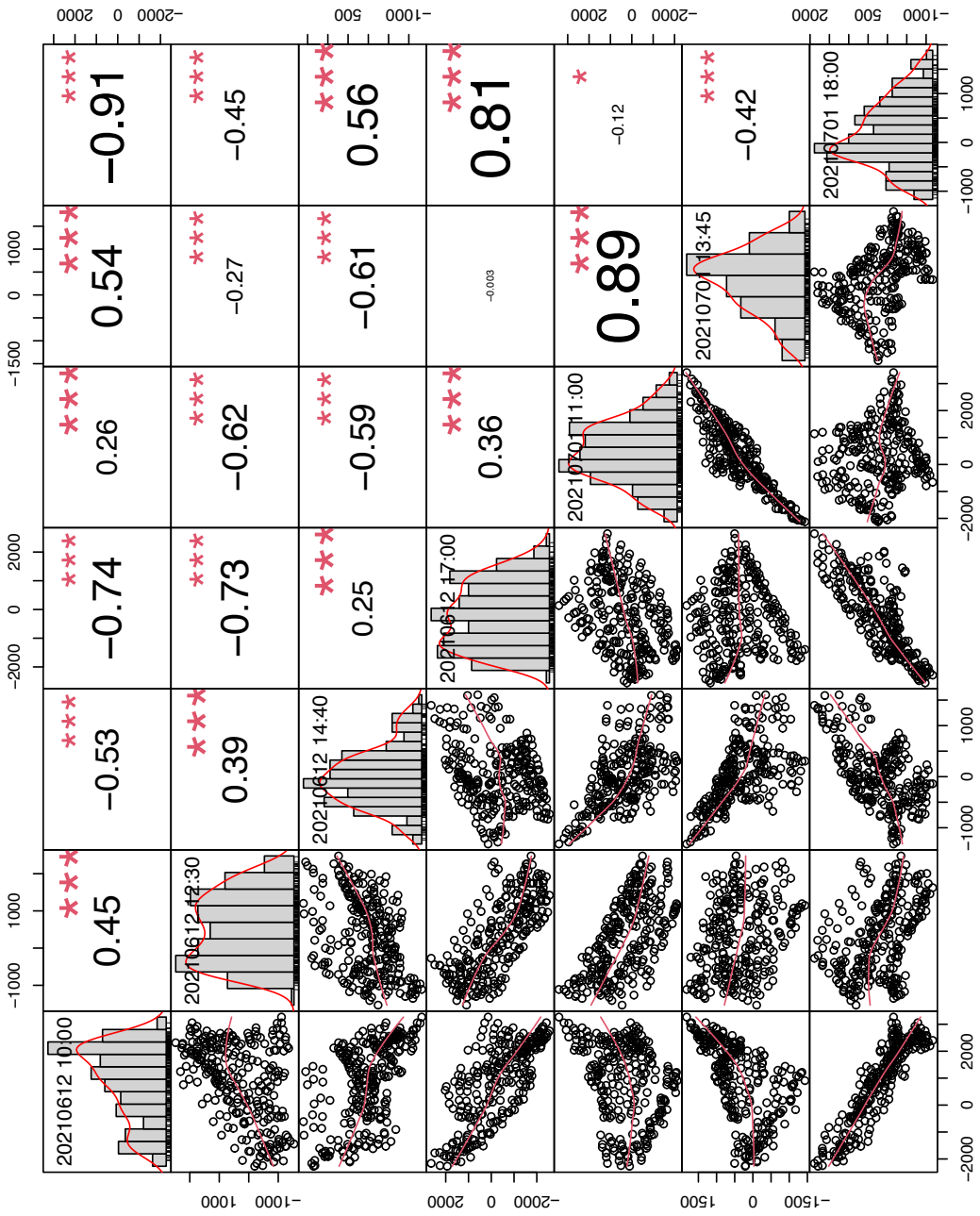


FIGURE S2.38: Pearson correlation of estimates of spatial trends according to Eq. 3.10 between campaigns of EuVar21 (data of Fig. S2.36a).

S2.37 Detailed correlation charts of campaign-wise spatial trends - EuVar22

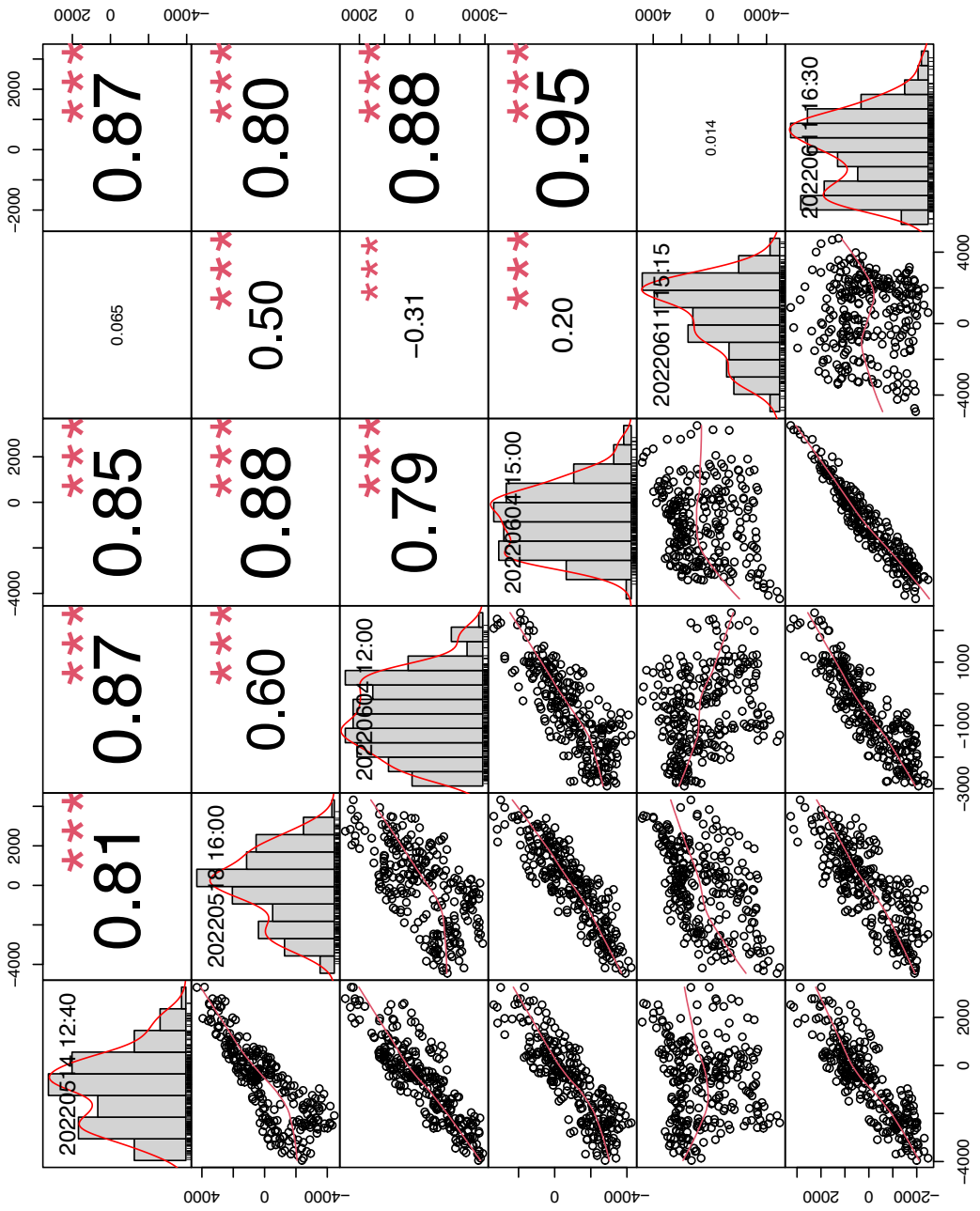


FIGURE S2.39: Pearson correlation of estimates of spatial trends according to Eq. 3.10 between campaigns of EuVar22 (data of Fig. S2.36b).

S2.38 Detailed correlation charts of campaign-wise spatial trends - SwiVar21

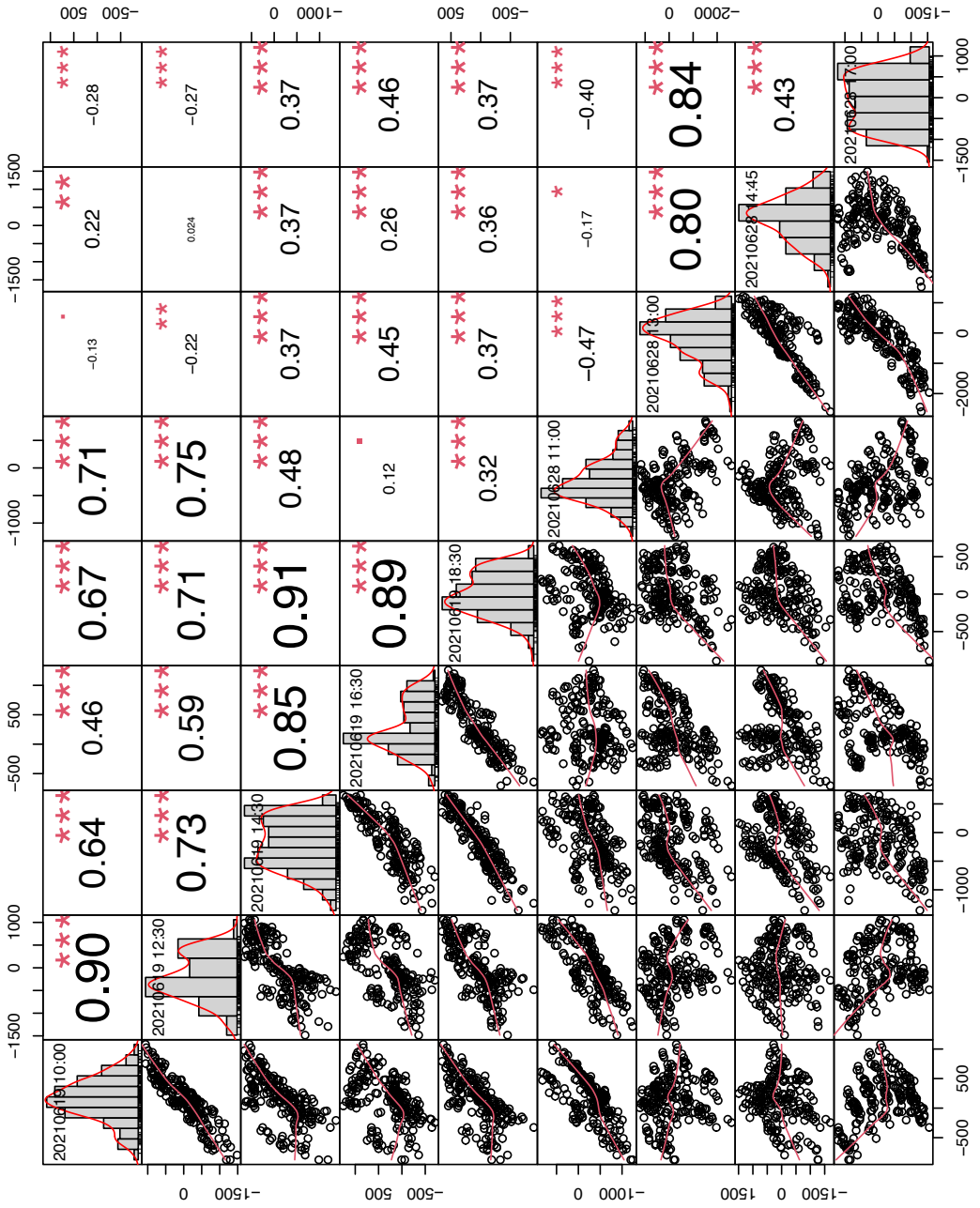


FIGURE S2.40: Pearson correlation of estimates of spatial trends according to Eq. 3.10 between campaigns of SwiVar21 (data of Fig. S2.37a).

S2.39 Detailed correlation charts of campaign-wise spatial trends - SwiVar22

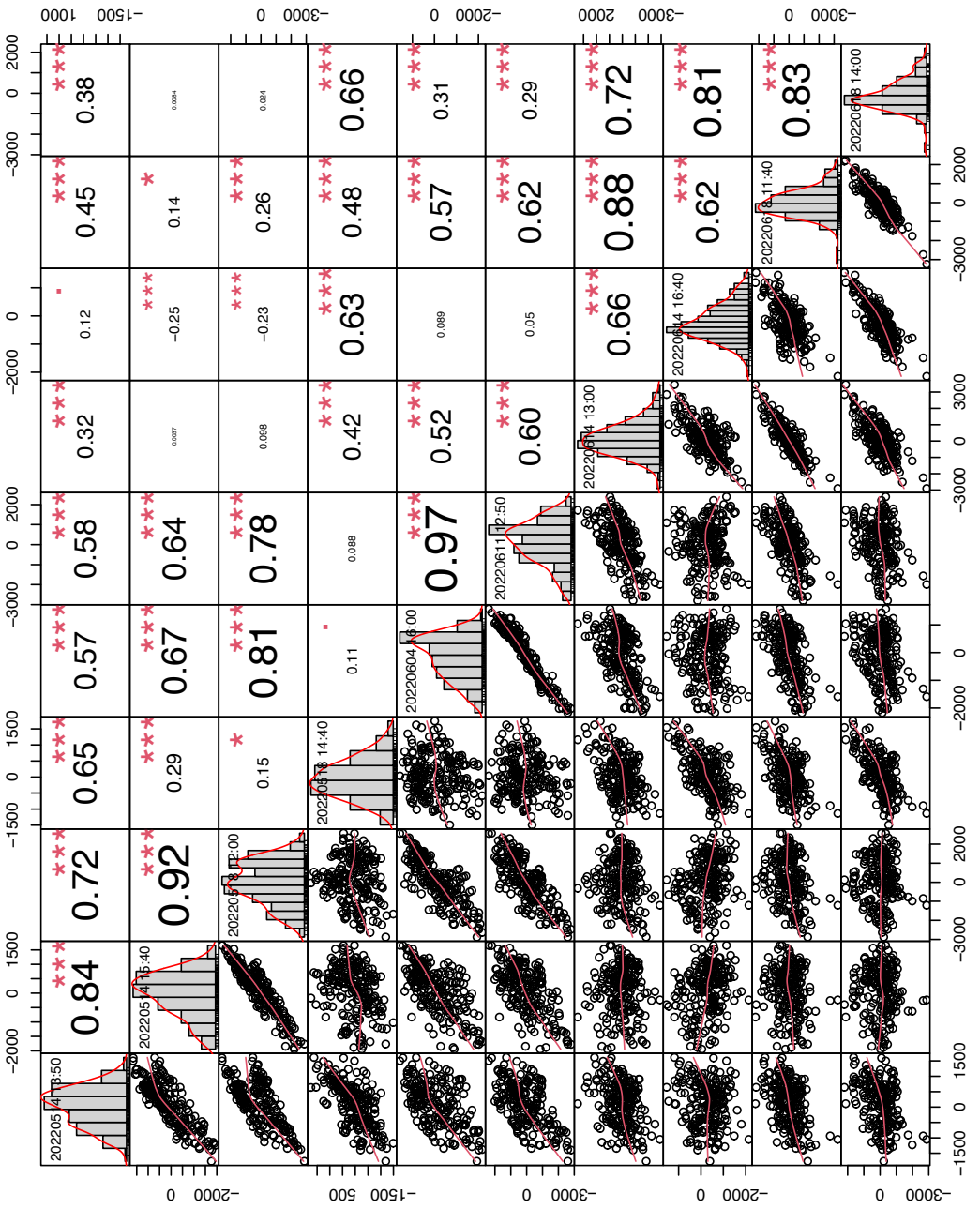


FIGURE S2.41: Pearson correlation of estimates of spatial trends according to Eq. 3.10 between campaigns of SwiVar22 (data of Fig. S2.37b).

S2.40 Flight-wise variance reduction by mixed models and PLSR - EuVar without vignetting correction

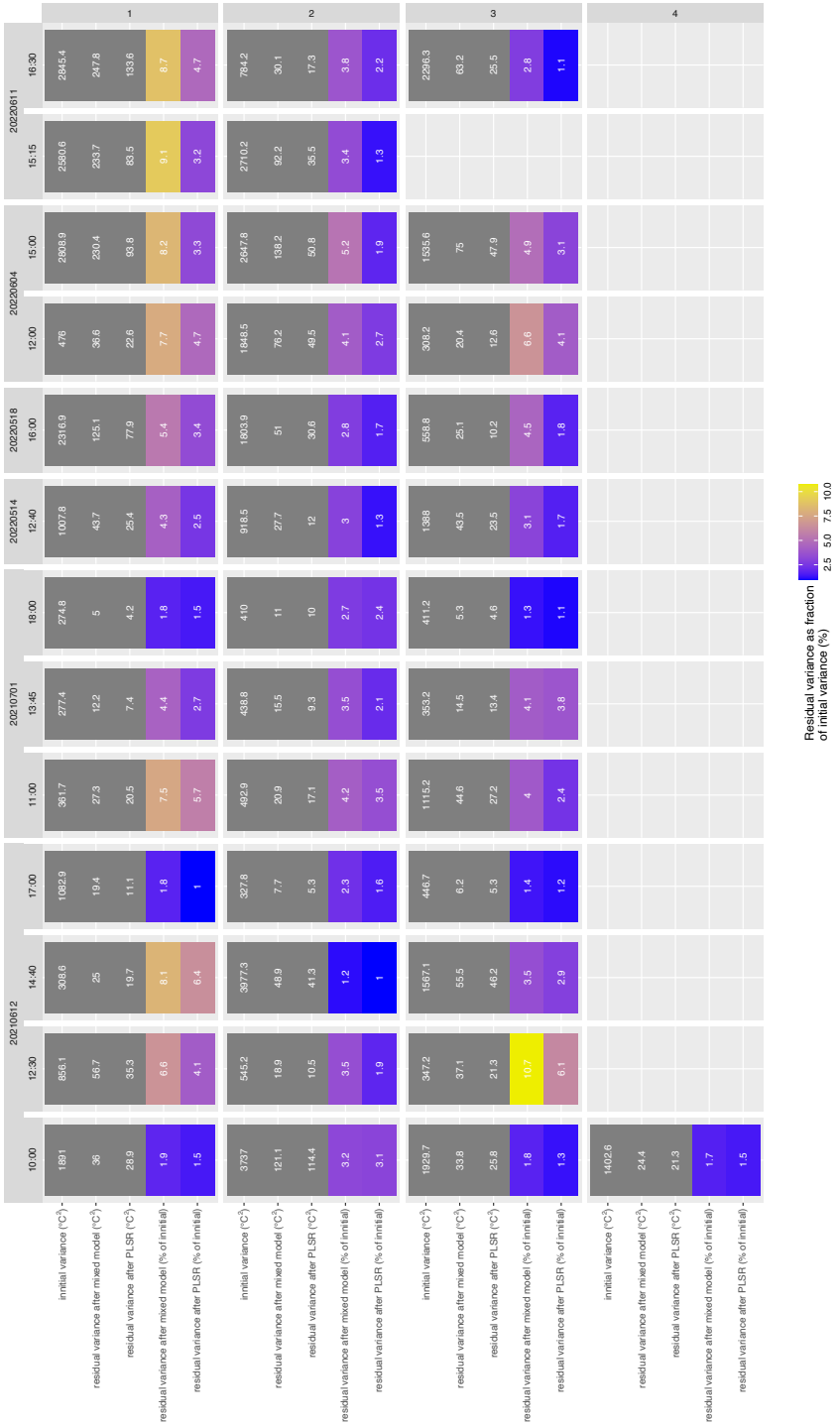


FIGURE S2.42: CT estimate variance reduction by mixed model and PLSR for EuVar without vignetting correction applied. Variance is shown for initial estimates (multiple per plot), after mixed models and after PLSR. Variance after mixed models and PLSR are also indicated as % of initial variance. Individual campaigns are arranged in columns, the rows represent the flights within the campaigns.

S2.41 Flight-wise variance reduction by mixed models and PLSR - EuVar with vignetting correction

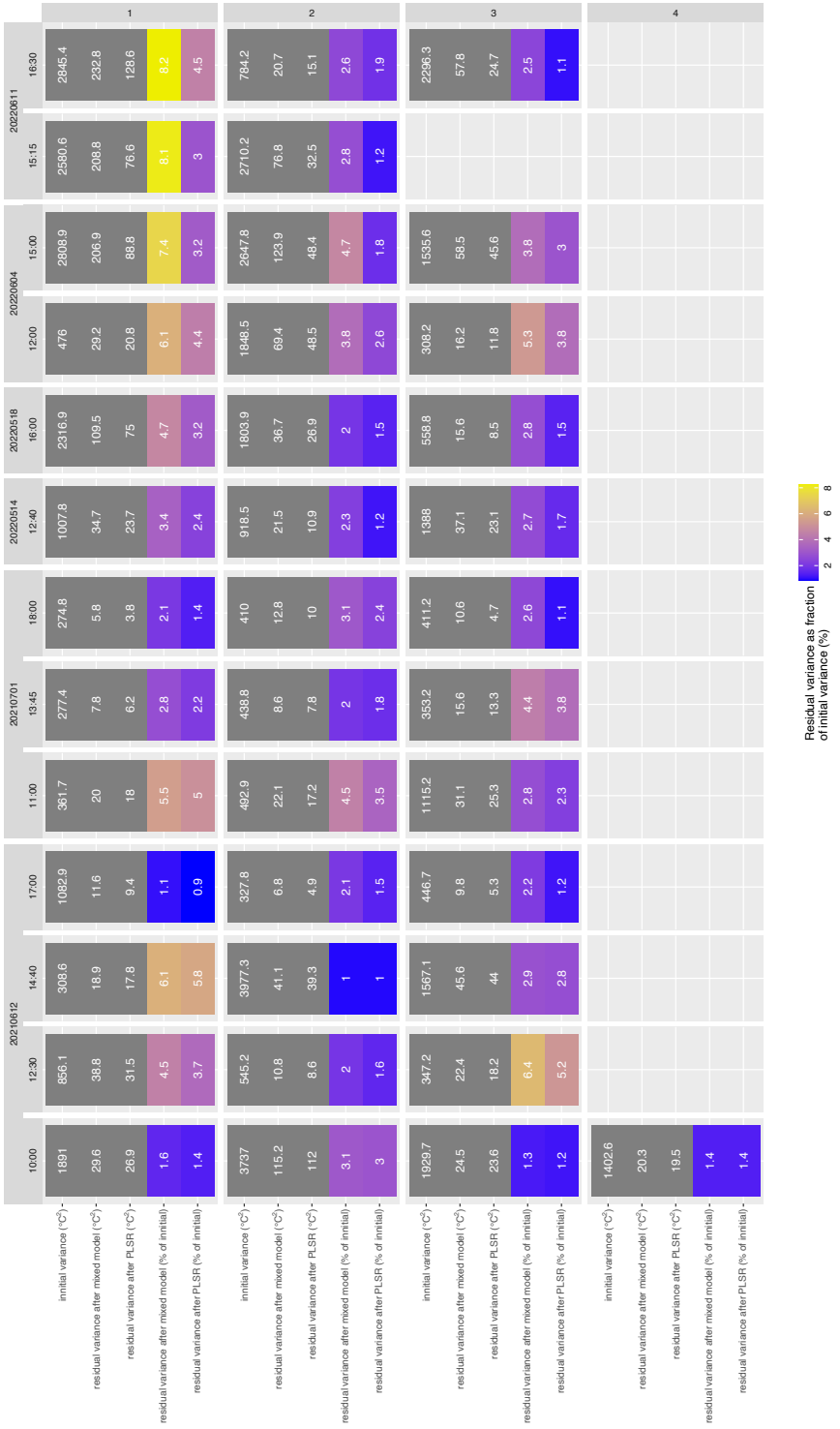


FIGURE S2.43: CT estimate variance reduction by mixed model and PLSR for EuVar with vignetting correction applied. Variance is shown for initial estimates (multiple per plot), after mixed models and after PLSR. Variance after mixed models and PLSR are also indicated as % of initial variance. Individual campaigns are arranged in columns, the rows represent the flights within the campaigns.

S2.42 Flight-wise variance reduction by mixed models and PLSR - SwiVar without vignetting correction

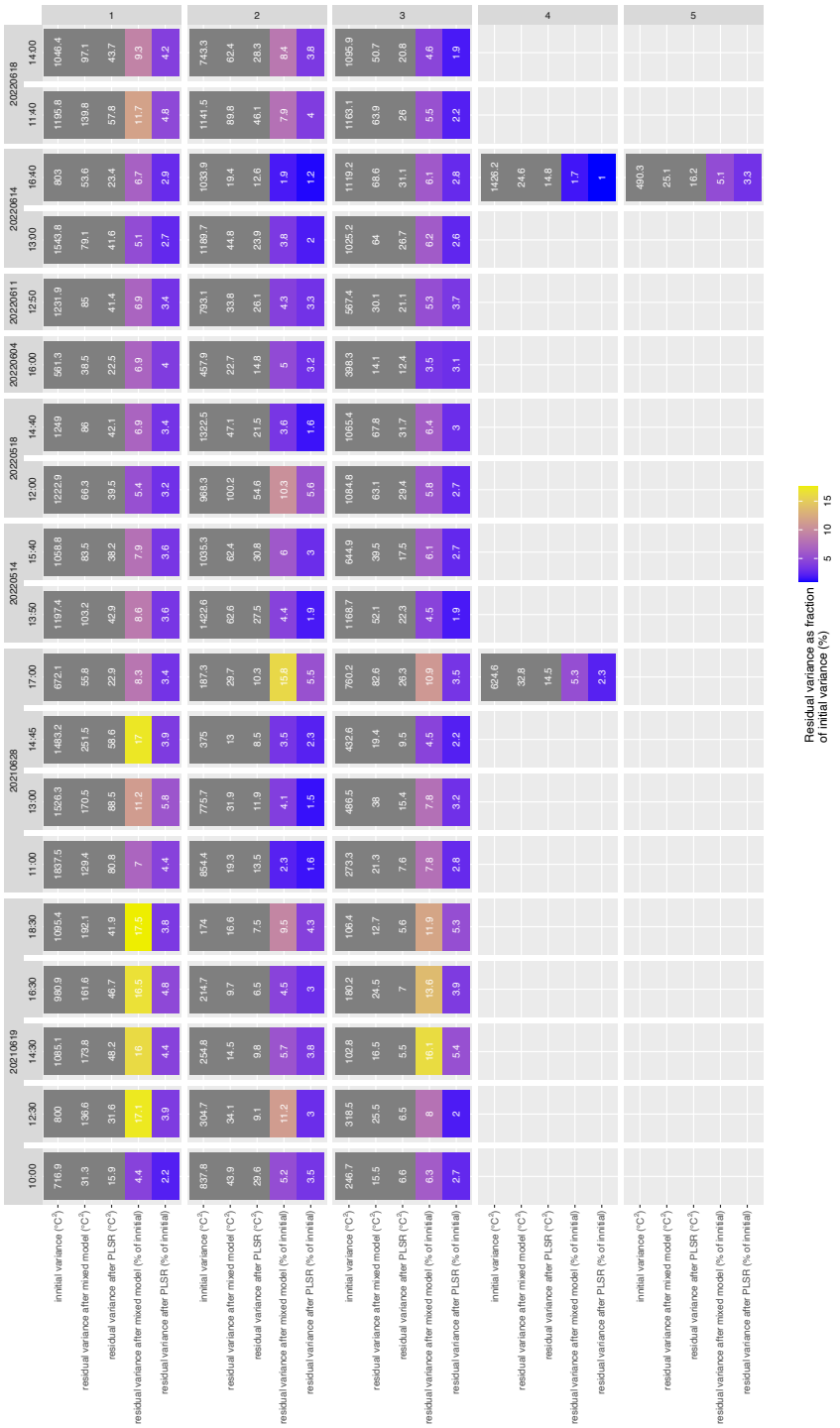


FIGURE S2.44: CT estimate variance reduction by mixed model and PLSR for SwiVar without vignetting correction applied. Variance is shown for initial estimates (multiple per plot), after mixed models and after PLSR. Variance after mixed models and PLSR are also indicated as % of initial variance. Individual campaigns are arranged in columns, the rows represent the flights within the campaigns.

S2.43 Flight-wise variance reduction by mixed models and PLSR - SwiVar with vignetting correction

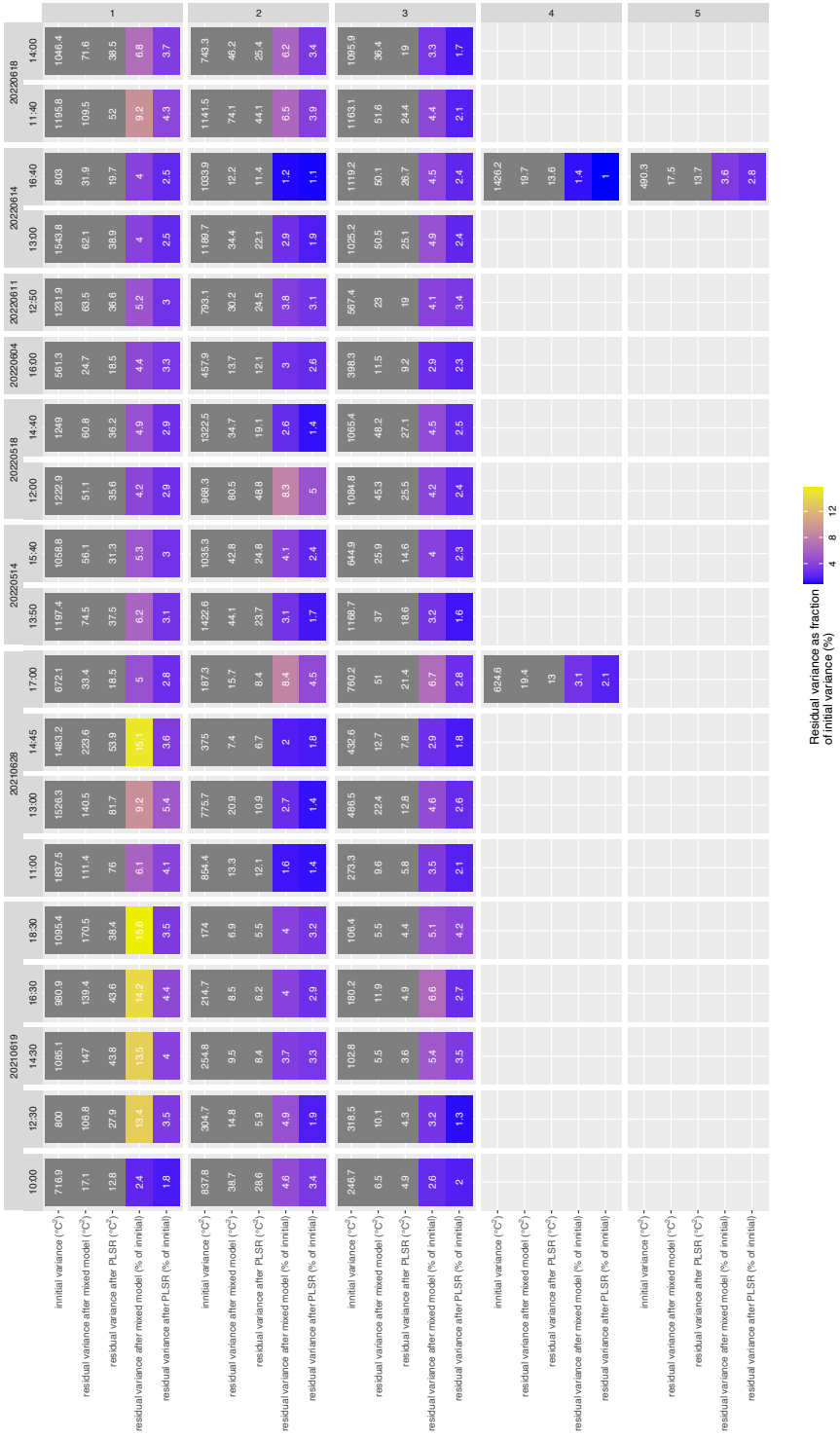


FIGURE S2.45: CT estimate variance reduction by mixed model and PLSR for SwiVar with vignetting correction applied. Variance is shown for initial estimates (multiple per plot), after mixed models and after PLSR. Variance after mixed models and PLSR are also indicated as % of initial variance. Individual campaigns are arranged in columns, the rows represent the flights within the campaigns.

S3 Supplementary Materials - Comparison of PhenoCams and drones for lean phenotyping of phenology and senescence of wheat genotypes in variety testing

S3.1 Experimental design

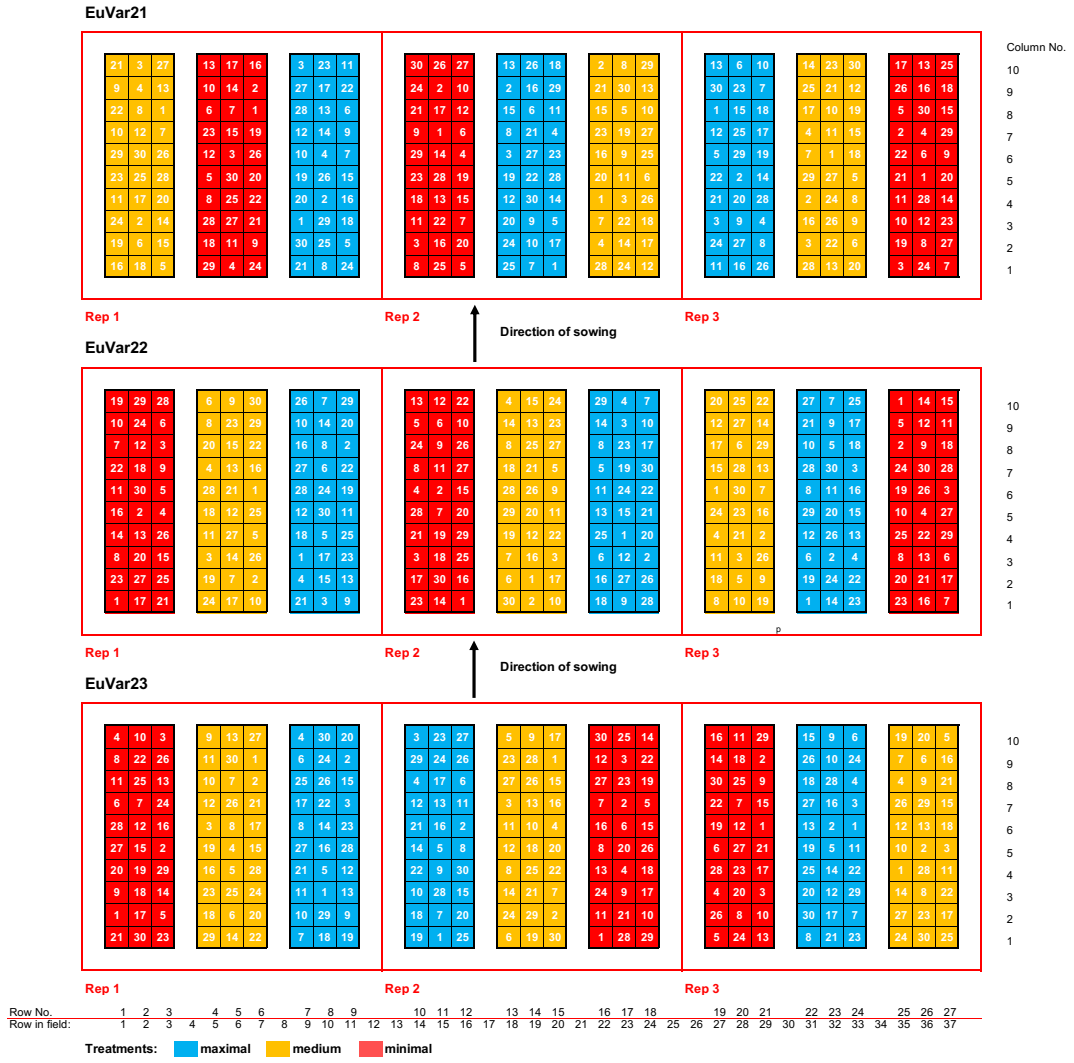


FIGURE S3.1: The experimental design of EuVar for the three seasons. The numbers inside the blocks indicate the genotypes.

S3.2 Details on Field treatments

TABLE S3.1: Overview of trial treatments and most important field interventions for all trials. "too wet" indicates that treatments were intended but could no be applied as conditions were too wet and heavy machinery could not enter the field.

				Herbicides					**Fertilization (kg/ha)**			
Experiment	**Treatment**	**Sowing date**	**Harvest date**	Monocot. 1st	Monocot. 2nd	Dicot.	**Growth regulator**	**Fungicide**	N	CaO	MgO	SO_3
EuVar21	Minimal	2020-10-22	2021-07-20	Archipel®	too wet	too wet	-	-	140	15	32	30
	Medium						Moddus®	-				
	Maximal						Moddus®	Amistar®				
EuVar22	Minimal	2021-10-15	2022-06-30	Archipel®	Othello Star®	Cleave®/ Express Max®	-	-	140	23	37	30
	Medium						Moddus®	-				
	Maximal						Moddus®	Amistar®				
EuVar23	Minimal	2022-10-19	2023-07-13	Herold®	Othello Star®	Herold®	-	-	132	22	12	-
	Medium						Ethephon®	Amistar				
	Maximal						Ethephon®	Proline®				

TABLE S3.2: Chemical compositions of field treatments and quantities applied.

Procuct	**Active ingredients**	**Application rate (g/ha)**	**Producer**
Archipel®	Iodosulfuron-methyl-sodium	9	Syngenta
	Mesosulfuron-methyl	9	
	Mefenpyr-diethyl	27	
Moddus®	Trinexapac-ethyl	125	Syngenta
Amistar®	Azoxystrobin	200	Syngenta
	Cyproconazole	80	
Othello Star®	Iodosulfuron-methyl-sodium	9	Bayer
	Mesosulfuron-methyl	9	
	Mefenpyr-diethyl	27	
	Thiencarbazone-methyl	7.5	
Cleave®	Fluroxypyr	90	Syngenta
	Fluroxypyr-meptyl	130	
	Florasulam	23	
Express Max®	Metsulfuron-methyl	5	Syngenta
	Tribenuron-methyl	5	
Herold®	Flufenacet	120	Bayer
	Diflufenican	120	
Ethephon®	Ethephon	480	Leu & Gygax
Proline®	Prothioconazol	200	Bayer

S3.3 Spectral properties of the Micasense RedEdge-MX Dual Camera System

TABLE S3.3: Specification of the ten bands of the Micasense RedEdge-MX Dual Camera System

Micasense band-name	**Band variable**	**Center wavelength** (nm)	**Band width** (nm)	**Micasence Band Suffix**
Coastal Blue	$Blue_{444}$	444	28	6
Blue	$Blue_{475}$	475	32	1
Green	$Green_{531}$	531	14	7
Green	$Green_{560}$	560	27	2
Red	Red_{650}	650	16	8
Red	Red_{668}	668	14	3
Red Edge	Red_Edge_{705}	705	10	9
Red Edge	Red_Edge_{717}	717	12	5
Red Edge	Red_Edge_{740}	740	18	10
Near IR	NIR_{842}	842	57	4

S3.4 Hardware issues

FIGURE S3.2: A ground screw was used to reinforce anchor pins. The pins are prone to loosening with the mast continuously shaking slightly in the wind.

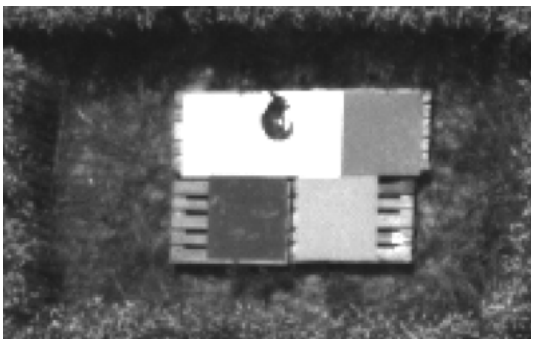
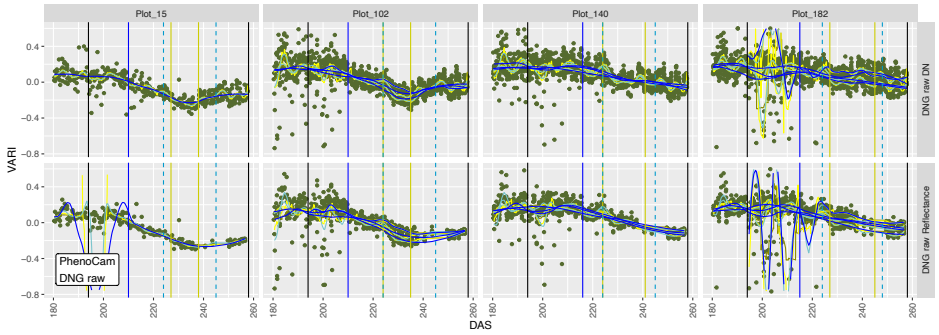
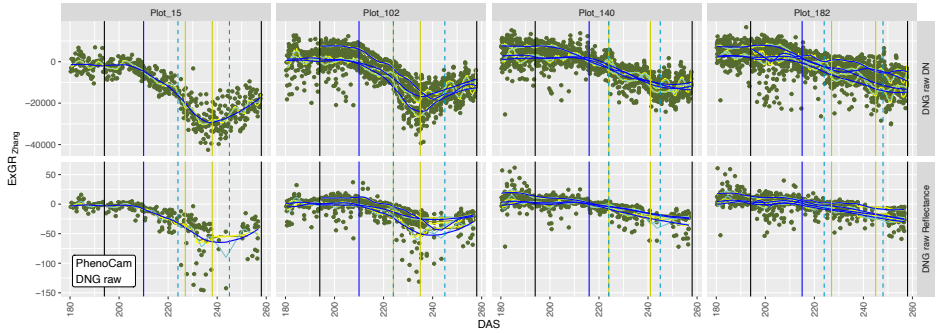


FIGURE S3.3: A fox laying on a calibration panel during a flight with the multispectral sensor. Foxes also left foodprints on the calibration panels, impacting their reflectance.

S3.5 Example VI data based on PhenoCam DNG raw format images



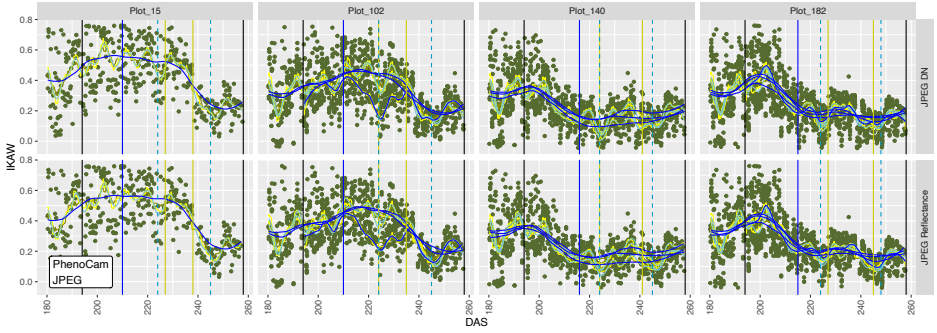
(a) VARI VI values derived from PhenoCam based DNG raw format images



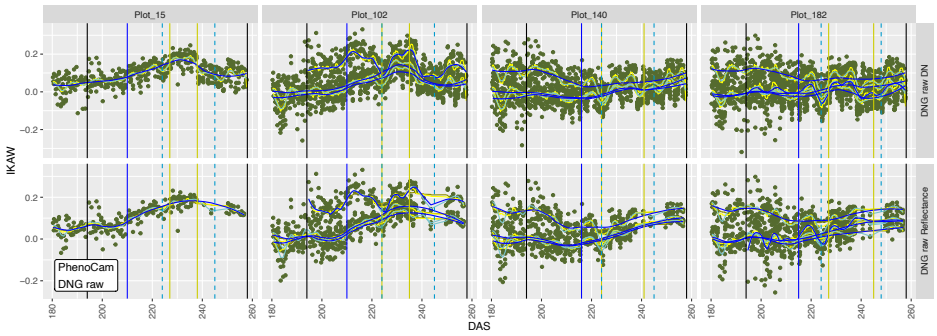
(b) ExGR $_{\text{Zhang}}$ VI values derived from PhenoCam based DNG raw format images

FIGURE S3.4: Example of VI data derived from PhenoCams images in DNG raw format for the two VIs VARI (a) ExGR $_{\text{Zhang}}$ (b) and four plots during the seasons 2022. The temporal axis is in days after sowing (DAS). Greenish points are initial VI values and lines represent smoothed data of different smoothing methods (dark blue: rolling mean; bright blue: loess smoothing; yellow: Savitzky–Golay; dark yellow: spline). In plots where multiple lines of the same color are present, multiple cameras observed the same plot. Data is shown for unprocessed data ("DN") and for calculated reflectance values. The solid blue vertical line indicates the heading date (BBCH 59) as observed on the respective plots, the dashed light blue line indicates plant senescence levels of 10 % and 90 % respectively. The yellow vertical lines correspond to flag leaf senescence at 10 % and 90 %. The black line toward the end mark the harvest date. The first vertical black line shows the date of PhenoCam maintenance.

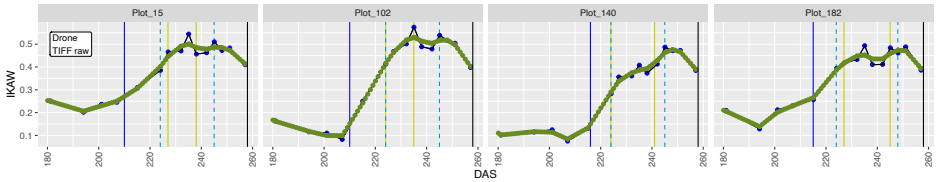
S3.6 Example IKAW VI



(a) IKAW VI values derived from PhenoCam based JPEG format images



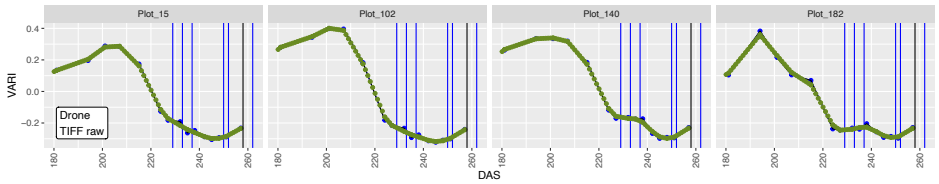
(b) IKAW VI values derived from PhenoCam based DNG raw format images



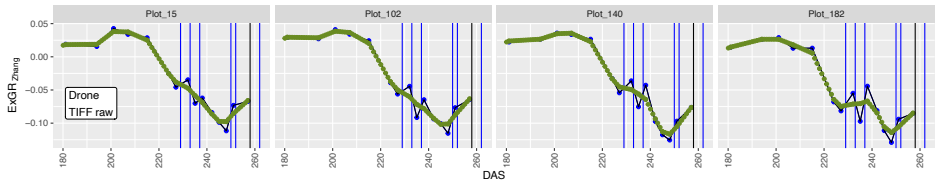
(c) IKAW VI values derived from drone based TIFF raw format images

FIGURE S3.5: Example of VI data derived from PhenoCam images in JPEG format (a) and in DNG raw format (b) and from a drone-based camera (c) for IKAW VI and four plots during the seasons 2022. The temporal axis is in days after sowing (DAS). For the PhenoCam data, greenish points in the PhenoCam image are initial VI values and lines represent smoothed data of different smoothing methods (dark blue: rolling mean; bright blue: loess smoothing; yellow: Savitzky–Golay; dark yellow: spline). In plots where multiple lines of the same color are present, multiple cameras observed the same plot. Data is shown for unprocessed data ("DN") and for calculated reflectance values. For the drone data, the initial VI values are blue dots. Greenish lines represent a smoothed spline interpolated for a daily temporal resolution. The solid blue vertical line indicates the heading date (BBCH 59) as observed on the respective plots, the dashed light blue line indicates plant senescence levels of 10 % and 90 % respectively. The yellow vertical lines correspond to flag leaf senescence at 10 % and 90 %. The black line toward the end mark the harvest date. The first vertical black line for the PhenoCam data shows the date of PhenoCam maintenance.

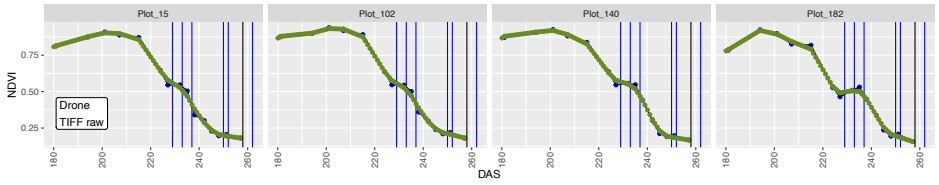
S3.7 Patterns of drone based VIs after rain events in a dry period



(a) VARI



(b) ExGR $_{\text{Zhang}}$



(c) NDVI

FIGURE S3.6: Example of of multispectral VI patterns after significant rainfalls in an otherwise dry summer of 2022. VIs were derived from TIFF raw format for the two VIs VARI (a), ExGR $_{\text{Zhang}}$ (b) and NDVI (c). Some index values are the same as in (Figs. 4.5c & 4.5d) but presented with blue vertical lines, indicating significant rainfalls ($> 5\,\text{mm}\,\text{d}^{-1}$). After rain events, descending trends of VIs often weakened or were even reversed.

S3.8 RMSE of PLSR-predictions

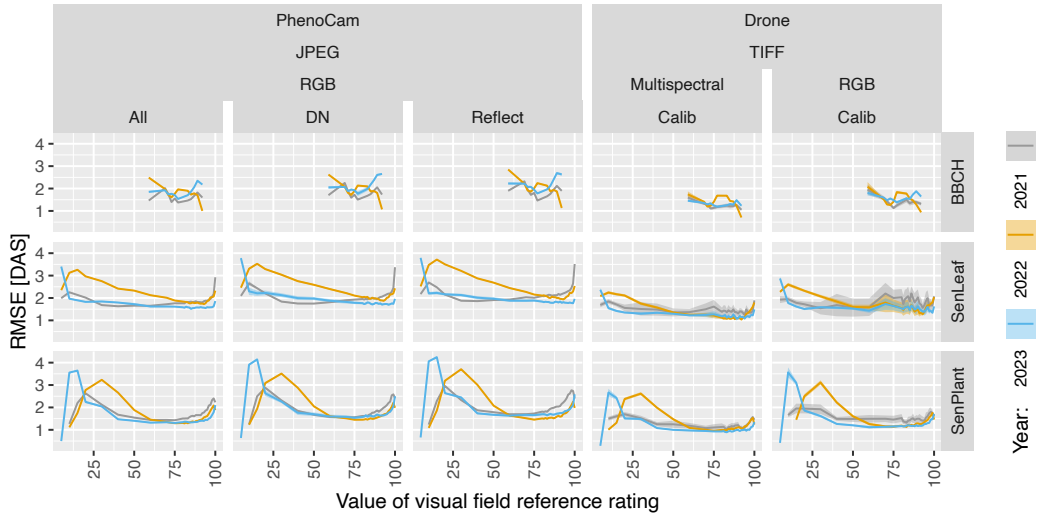
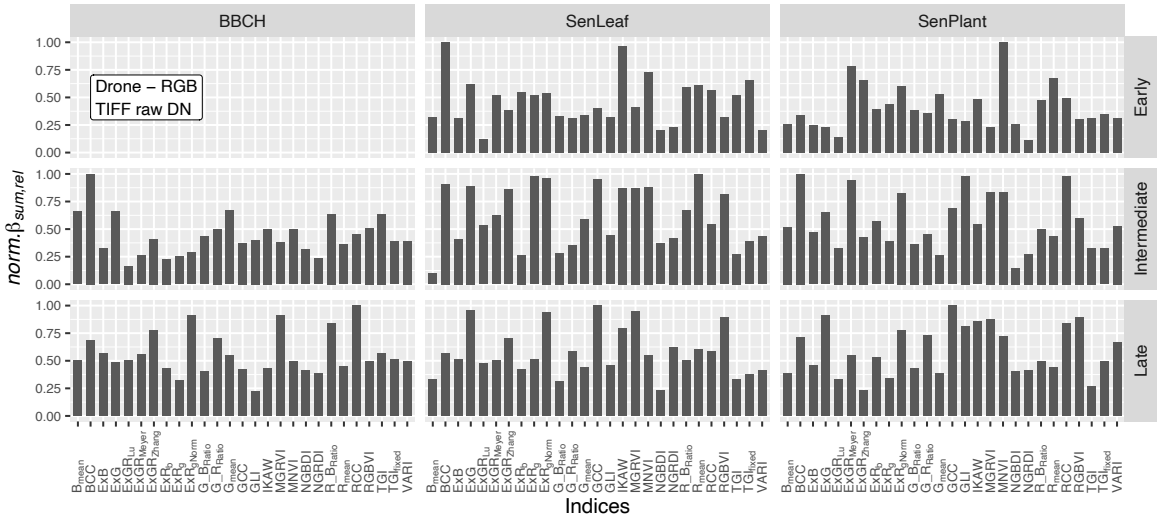
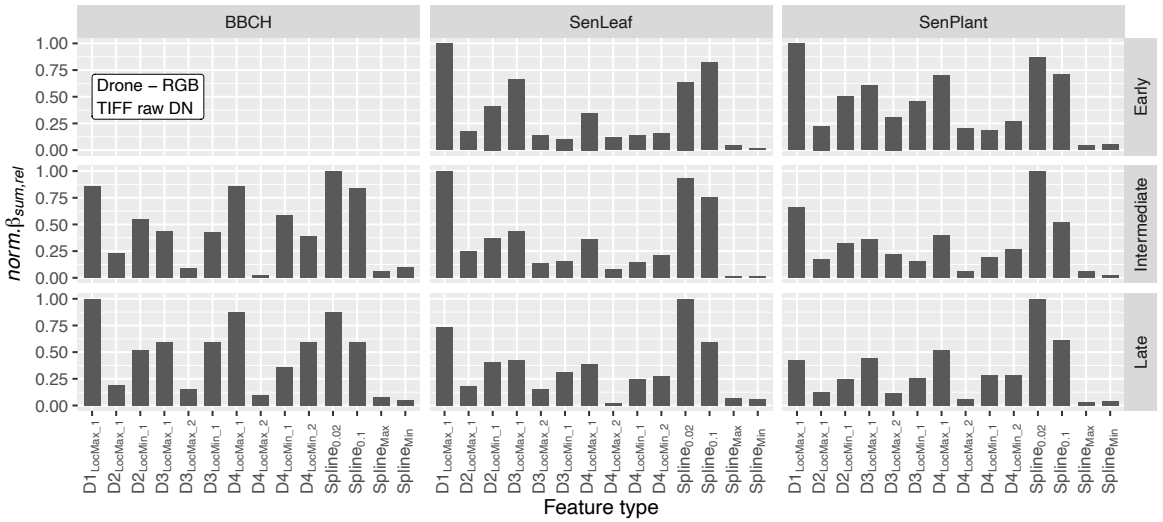


FIGURE S3.7: RMSE of PLSR-based predictions and field reference ratings. Data of the different sensors and processing methods is arranged in columns, the rows represent the three types of field reference measurements. Colors represent the different years. The shaded areas indicated mean ± standard deviation across 100 repetitions of cross validation while the lines are the means.

S3.9 Index and feature type importance - Drone RGB



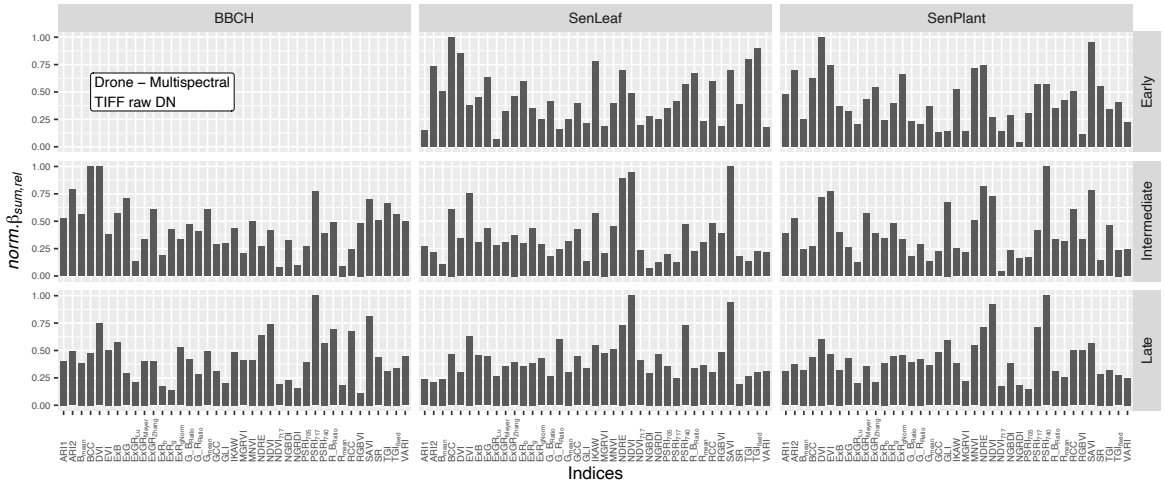
(a)



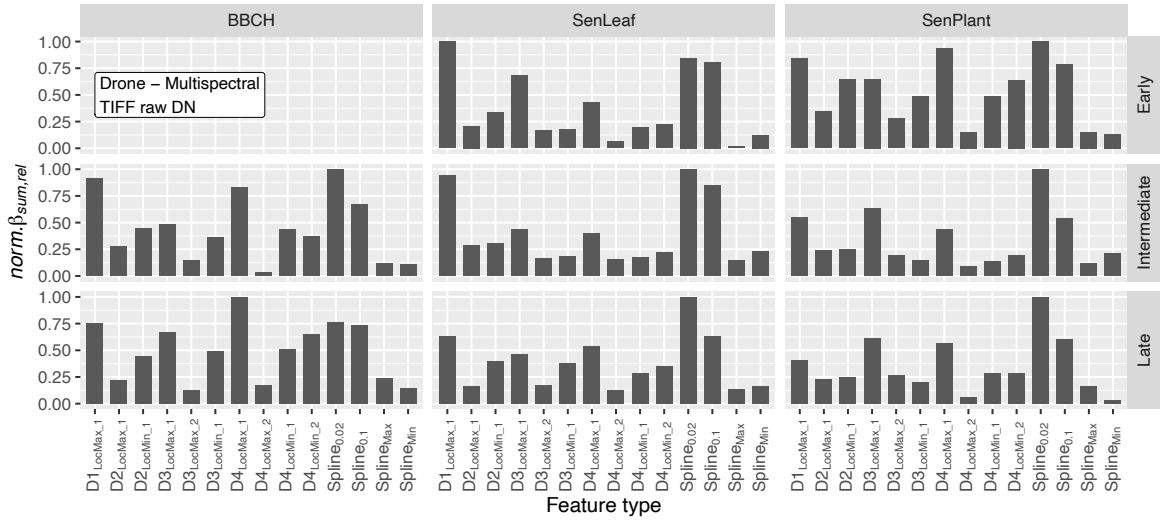
(b)

FIGURE S3.8: Importance of VIs and temporal feature types in PLSR models based on drone derived RGB TIFF raw images. Data is arranged by reference types (Phenology: BBCH; Flag leaf senescence: SenLeaf; Plant senescence: SenPlant), and reference classes (Early, Intermediate, Late). (a) Importance of the different indices as relative PLSR coefficients sum $\beta_{rel,sum}$ over the 100 repetitions of cross validation. (b) Importance of the different feature types as $norm.\beta_{rel,sum}$ over the 100 repetitions of cross validation. D1 - D4 indicate the first four derivatives of the Gompertz function. LocMax and LocMin refer to local maxima and minima and the number after LocMax and LocMin the order along increasing DAS, when there were multiple local maxima/minima of the same type. The remaining features correspond to non-parametric temporal features of the smoothing types loess, rolling mean, Savitzky-Golay and spline.

S3.10 Index and feature type importance - Drone Multispectral



(a)



(b)

FIGURE S3.9: Importance of VIs and temporal feature types in PLSR models based on drone derived mutlispectral TIFF raw images. Data is arranged by reference types (Phenology: BBCH; Flag leaf senescence: SenLeaf; Plant senescence: SenPlant), and reference classes (Early, Intermediate, Late). (a) Importance of the different indices as relative PLSR coefficients sum $\beta_{rel,sum}$ over the 100 repetitions of cross validation. (b) Importance of the different feature types as $norm.\beta_{rel,sum}$ over the 100 repetitions of cross validation. D1 - D4 indicate the first four derivatives of the Gompertz function. LocMax and LocMin refer to local maxima and minima and the number after LocMax and LocMin the order along increasing DAS, when there were multiple local maxima/minima of the same type. The remaining features correspond to non-parametric temporal features of the smoothing types loess, rolling mean, Savitzky-Golay and spline.

S3.11 Importance of mean and multiple percentiles as data aggregation methods

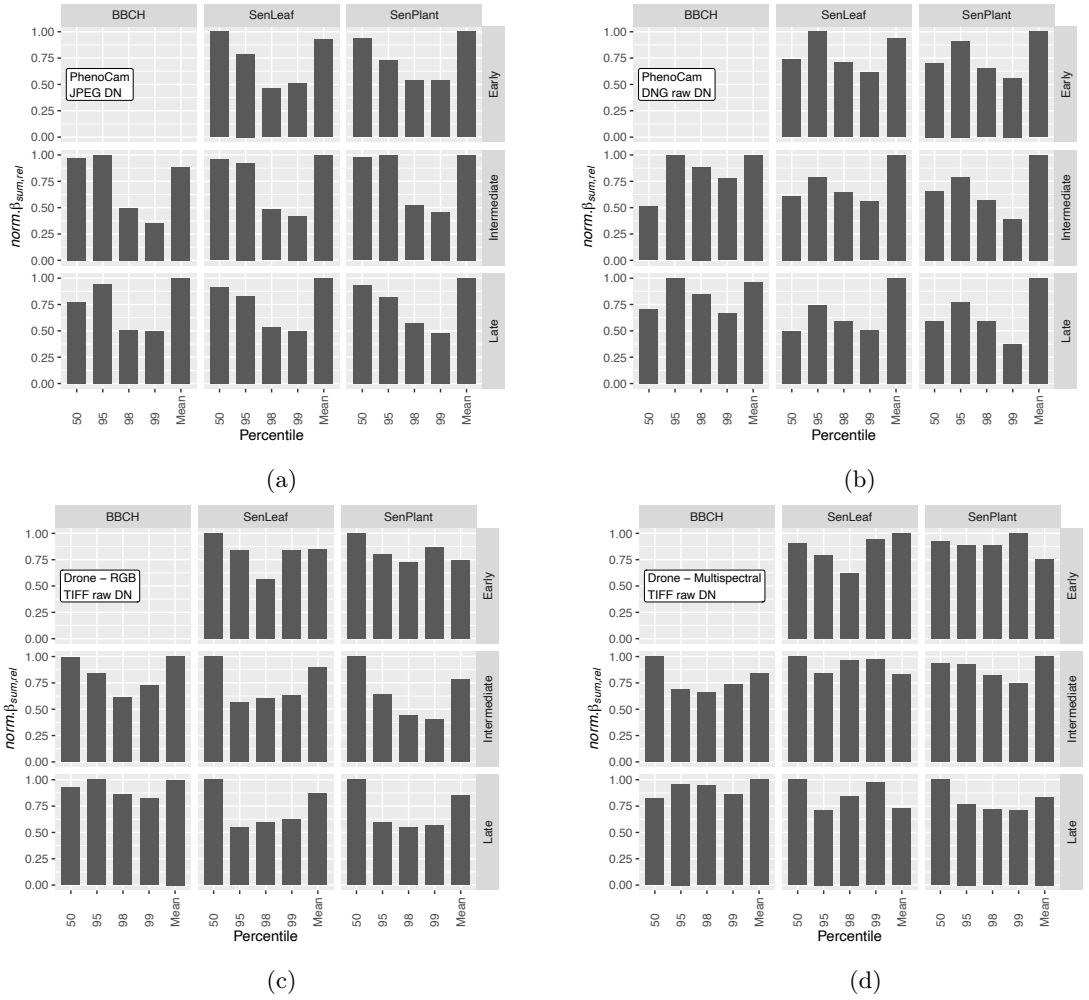


FIGURE S3.10: Index values were derived from image data aggregated per plot with multiple percentiles and the mean across all pixel values within a plot. Data is arranged by reference types (Phenology: BBCH; Flag leaf senescence: SenLeaf; Plant senescence: SenPlant), and reference classes (Early, Intermediate, Late). The importance of the different percentiles and the mean is presented as $norm.\beta_{rel,sum}$ over the 100 repetitions of cross validation for (a) PhenoCam data in JPEG format, (b) PhenoCam data in DNG raw format, (c) TIFF format drone data in RGB color space and (d) in multispectral color space.

S3.12 Impact of number of temporal features on performance in PLSR predictions

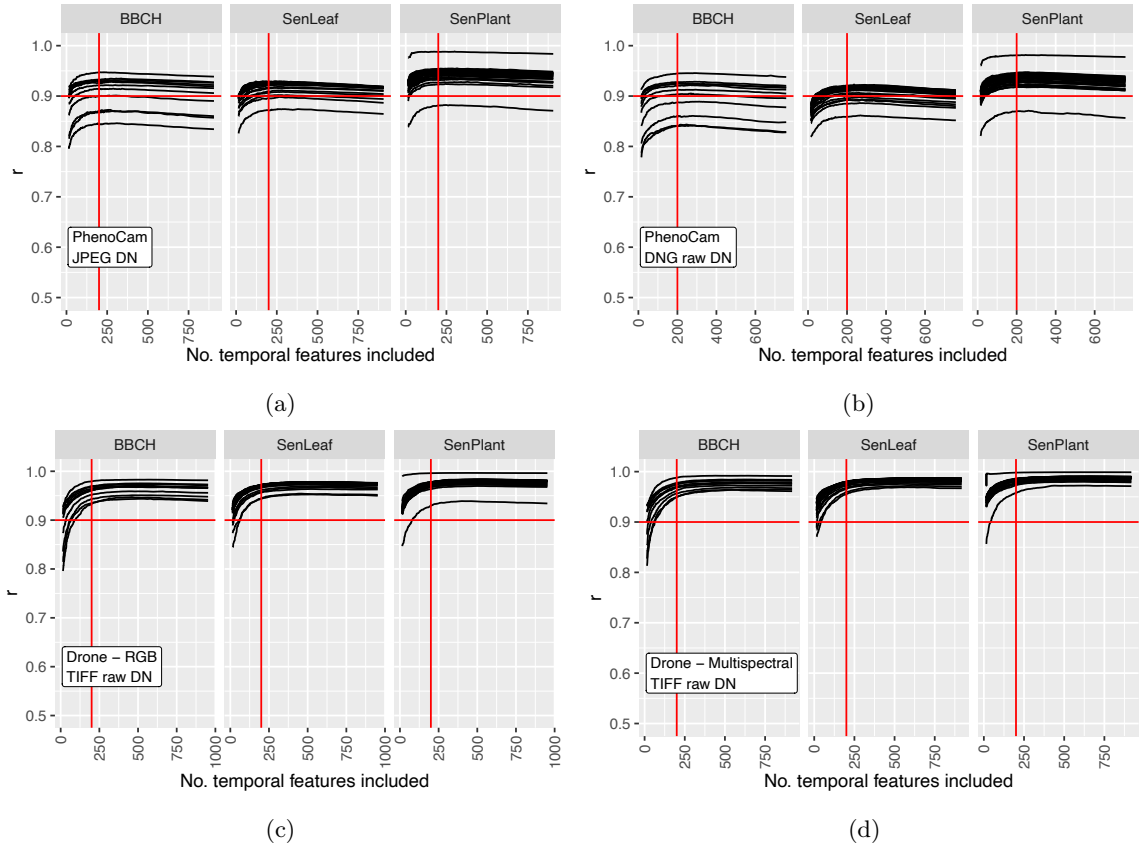
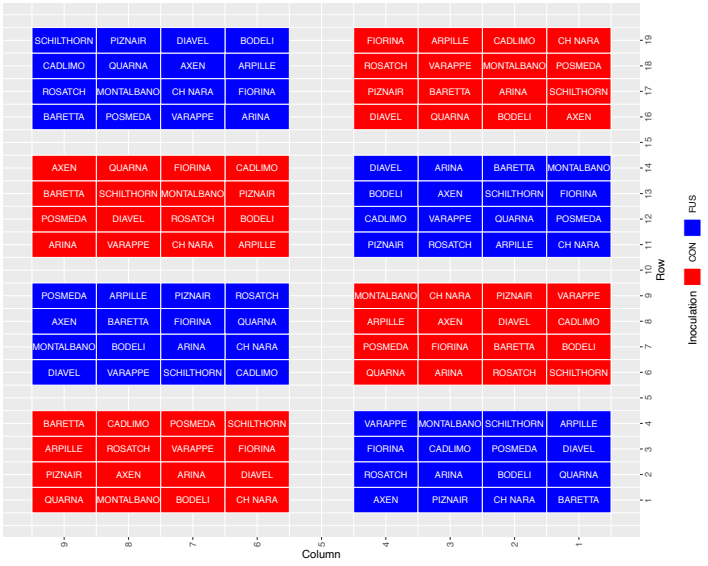


FIGURE S3.11: Pearson's correlation between PLSR predictions and visual field reference ratings in dependence of number of temporal features included in PLSR modeling. (a) PhenoCam data in JPEG format. (b) PhenoCam data in DNG raw format. (c) Drone data in DNG in RGB color space. (d) Drone data in multispectral color space. The vertical red line indicates 200 features and the horizontal red line a correlation coefficient of 0.9. Data is arranged in columns by reference types (Phenology: BBCH; Flag leaf senescence: SenLeaf; Plant senescence: SenPlant).

S4 Supplementary Materials - Evaluating the potential of chlorophyll fluorescence to detect and rate *Fusarium* head blight on field experiments for winter wheat variety testing

S4.1 Experimental design Changins and Cadenazzo 2021

(a)

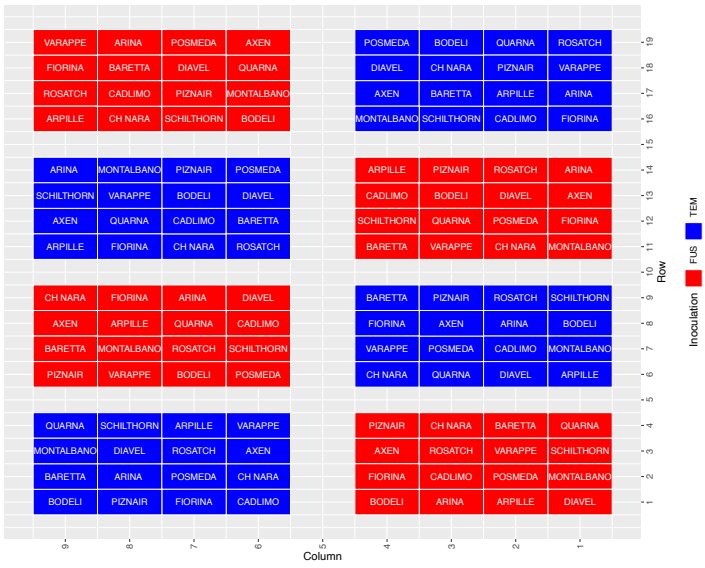


(b)

FIGURE S4.1: The arrangement of the 16 genotypes (a) in four replications for the two treatments (b) in the field for the 2021 season. The design was identical for Changins and Cadenazzo.

S4.2 Experimental design Changins 2022

(a)



(b)

FIGURE S4.2: The arrangement of the 16 genotypes (a) in four replications for the two treatments (b) in the field for the 2022 season at Changins.

S4.3 Wheat varieties

Table S4.1: Wheat varieties used in the experiment. 13 winter wheat varieties and 3 spring wheat varieties were used. The table lists the *Fusraium* tolerance and wheat types. Tolerance encoding: "–": weak; "-": weak to fair; "Ø": fair; "+": fair to good; "+": good (Agroscope, 2004; Agroscope, 2017; Agroscope, 2020a; Agroscope, 2020b; Strebel, Levy Häner, Mattin, et al., 2022; FiBL, 2022; Strebel, Levy Häner, Watroba, et al., 2024).

Variety	**Tolerance**	**Wheat type**
ARINA	++	
AXEN	-	
BARETTA	Ø	
BODELI	Ø	
CADLIMO	+	
CH-NARA	–	
DIAVEL	Ø	Winter
MONTALBANO	++	
PIZNAIR	-	
POSMEDA	-	
ROSATCH	+	
SCHILTHORN	Ø	
VARAPPE	-	
ARPILLE	Ø	
FIORINA	Ø	Summer
QUARNA	++	

S4.4 Details on Field treatments

Table S4.2: Overview of trial treatments and most important field interventions for all trials. "too wet" indicates that treatments were intended but could not be applied as conditions were too wet and heavy machinery could not enter the field.

				Herbicides			**Fertilization [kg/ha]**					
Experiment	**Treatment**	**Sowing date**	**Harvest date**	Monocot. 1st	Monocot. 2nd	Dicot.	N	P_2O_5	K_2O	CaO	MgO	SO_3
CHA21	Minimal	2020-11-07	2021-07-20	Archipel®	too wet	too wet	140	-	-	15	9	-
CHA22	Minimal	2021-10-15	2022-07-08	Archipel®	Othello Star®	Cleave®/Express Max®	140	-	-	23	36	30
CAD21	Minimal	2020-10-29	2021-07-12	-	-	-	136	46	74	$\sim 15^*$	$\sim 9^*$	-

*exact amount unknown

Table S4.3: Chemical compositions of field treatments and quantities applied.

Procuct	**Active ingredient(s)**	**Application rate [g/ha]**	**Producer**
Archipel®	Iodosulfuron-methyl-sodium	9	Syngenta
	Mesosulfuron-methyl	9	
	Mefenpyr-diethyl	27	
Othello Star®	Iodosulfuron-methyl-sodium	9	Bayer
	Mesosulfuron-methyl	9	
	Mefenpyr-diethyl	27	
	Thiencarbazone-methyl	7.5	
Cleave®	Fluroxypyr	90	Syngenta
	Fluroxypyr-meptyl	130	
	Florasulam	23	
Express Max®	Metsulfuron-methyl	5	Syngenta
	Tribenuron-methyl	5	

S4.5 Greenhouse trial

S4.5.1 Correlation between visual ratings and F_v/F_m

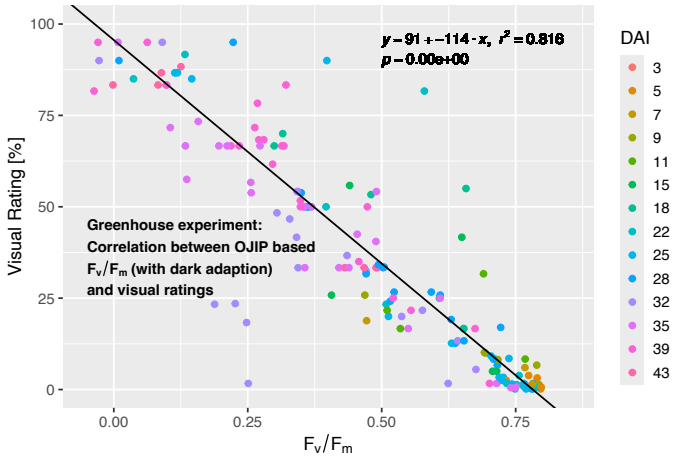


FIGURE S4.3: Correlation between visual ratings and F_v/F_m values for the greenhouse experiment. Values were grouped by variety, treatment and measurement event and group-wise means were correlated with each other. Values with a visual rating of either 0 % or 100 % were excluded from the data to avoid clustered data around the extremes do be the main drive of the strong correlation. DAI: Days after inoculation.

S4.6 Field trials

S4.6.1 Visual ratings: Changins 2021

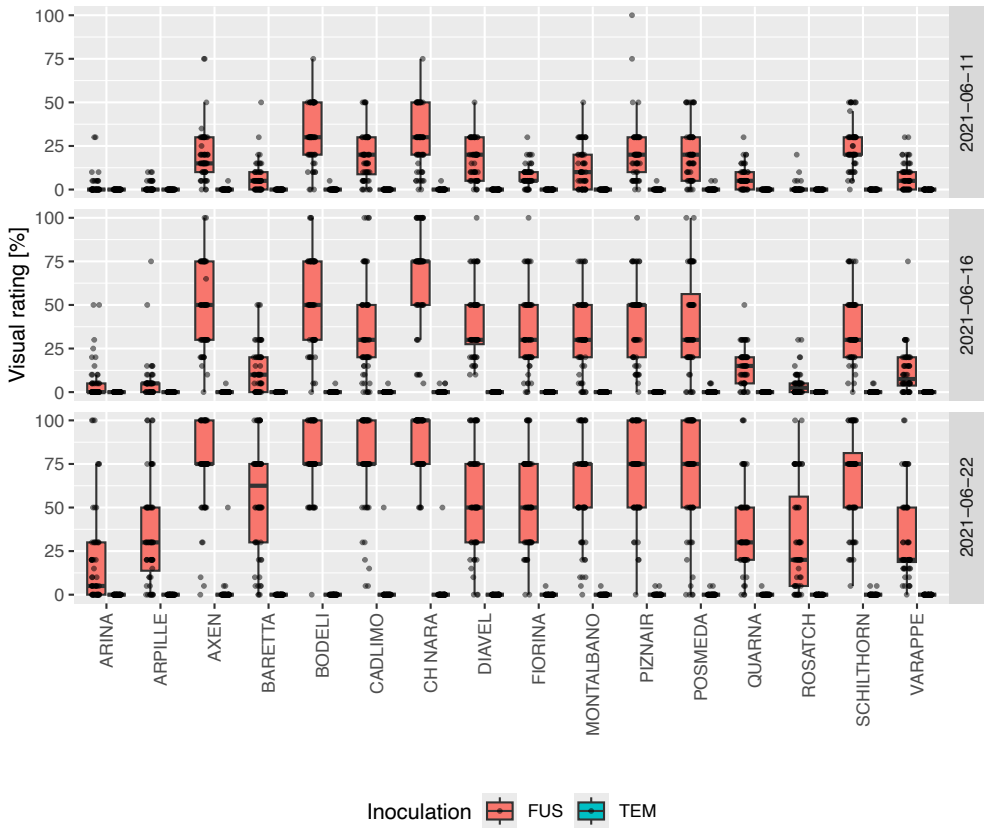


FIGURE S4.4: Visual ratings of *Fusarium* infestation, Changins 2021. Rating was according to Moll et al. (2000). 0 %: no visible symptoms; 100 % whole spike was infested. 15 spikes were rated per plot, resulting in 60 spikes per variety and treatment (1'920 spikes in total for each date of measurement).

S4.6.2 Visual ratings: Changins 2022

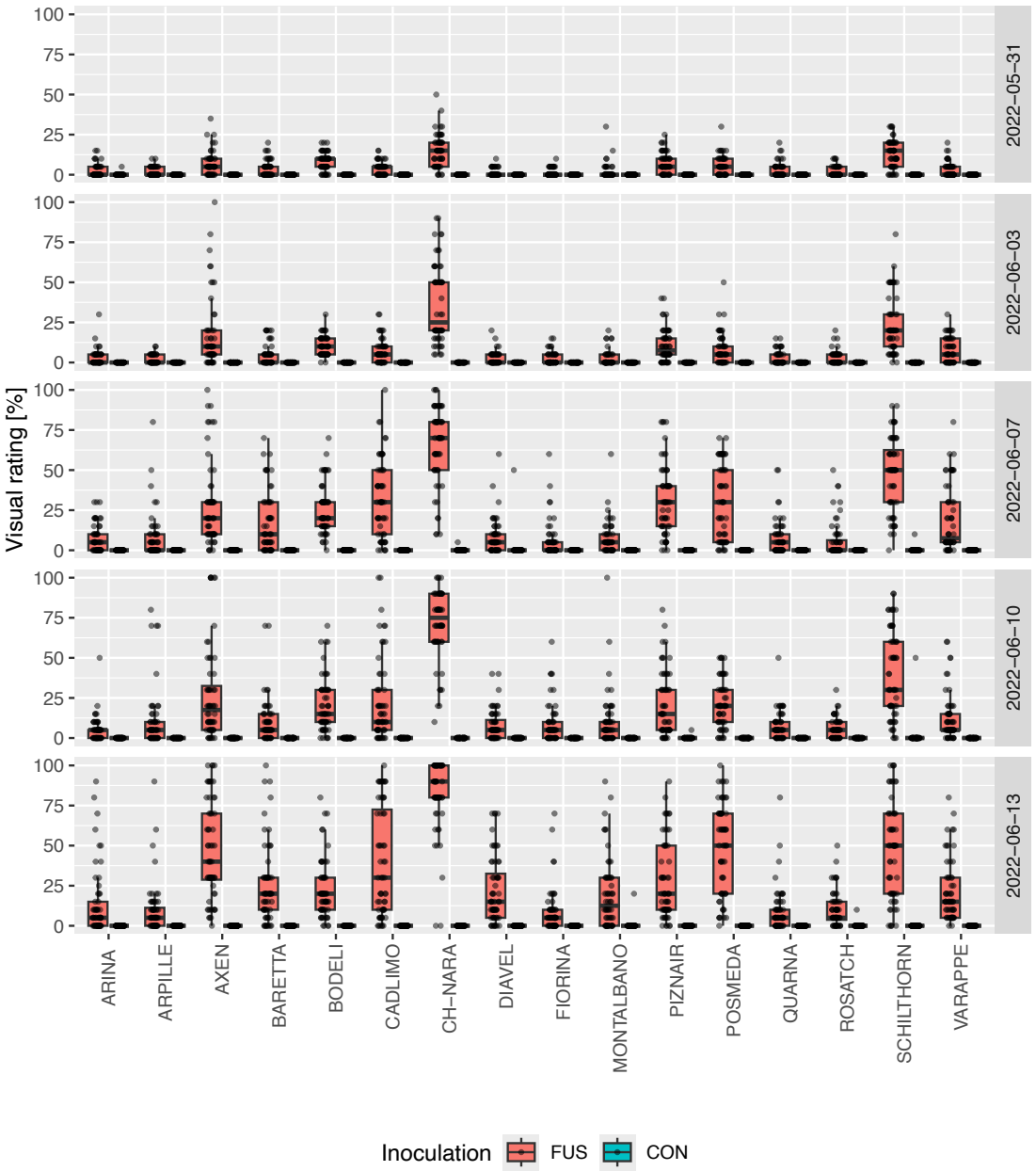


FIGURE S4.5: Visual ratings of *Fusarium* infestation, Changins 2022. Rating was according to Moll et al. (2000). 0 %: no visible symptoms; 100 % whole spike was infested. 15 spikes were rated per plot, resulting in 60 spikes per variety and treatment (1'920 spikes in total for each date of measurement).

S4.6.3 Visual ratings: Cadenazzo 2021

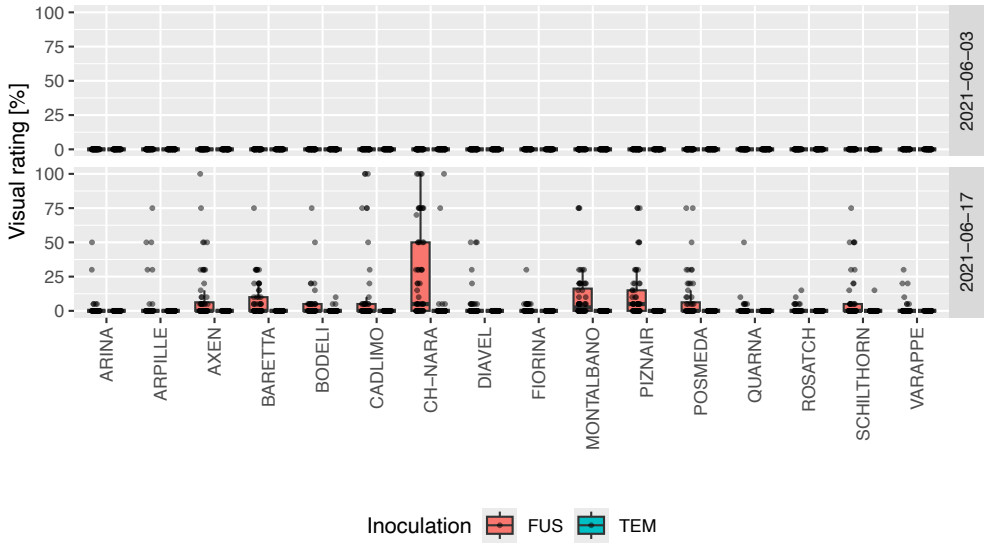


FIGURE S4.6: Visual ratings of *Fusarium* infestation, Cadenazzo 2021. Rating was according to Moll et al. (2000). 0 %: no visible symptoms; 100 % whole spike was infested. 15 spikes were rated per plot, resulting in 60 spikes per variety and treatment (1'920 spikes in total for each date of measurement).

S4.6.4 p-values of ANOVA on OJIP parameters

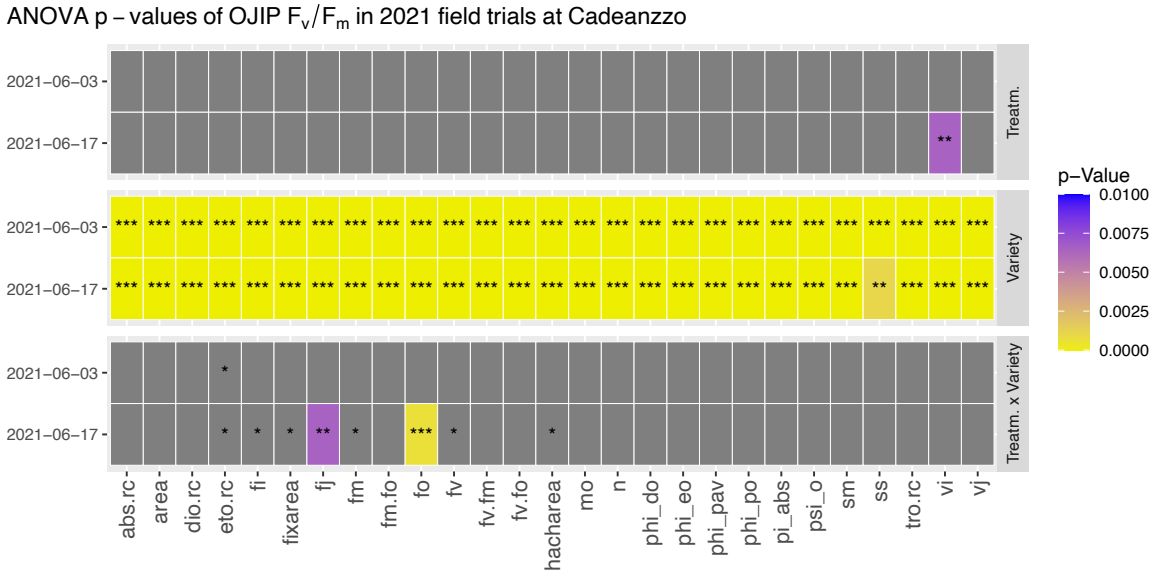


FIGURE S4.7: p-values of ANOVA on OJIP parameters of dark adapted measurements on field samples at Cadenazzo in 2021 for the factors "Treatm.", "Variety" and their interaction. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: p <, and colors indicate significance at $p < 0.01$.

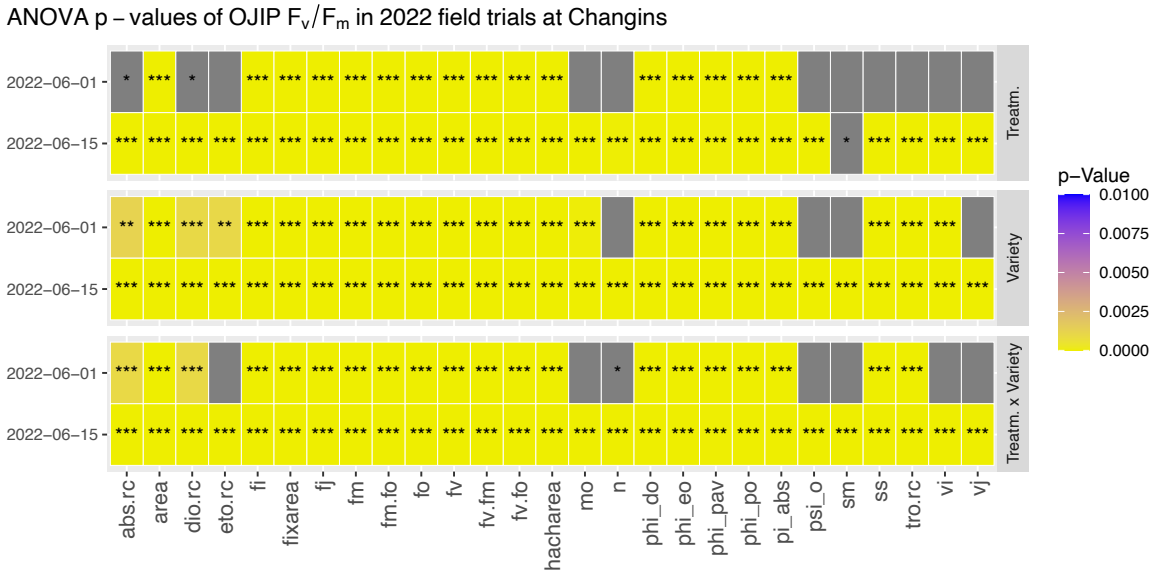


FIGURE S4.8: p-values of ANOVA on OJIP parameters of dark adapted measurements on field samples at Changins in 2022 for the factors "Treatm.", "Variety" and their interaction. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: p <, and colors indicate significance at $p < 0.01$.

S4.6.5 F_v/F_m parameter of OJIP protocol, Changins 2022

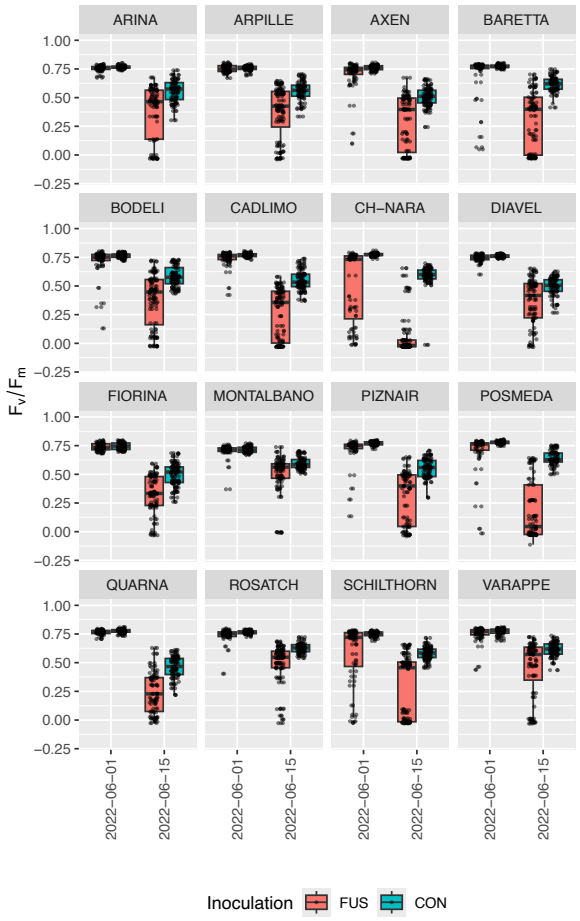


FIGURE S4.9: F_v/F_m parameter of OJIP data, Changins 2023. The 16 tiles represent the 16 wheat varieties tested over time. Dates are the individual measurement events. Inoculation treatments are indicated by color. 10 spikes were measured on spikelets on a central spikelet for each plot, resulting in n = 40 measurements per variety and inoculation treatment for each date of measurements (1'280 measurements for each date in total).

S4.6.6 p-values of ANOVA on rapid F_v'/F_m parameter, Changins 2021

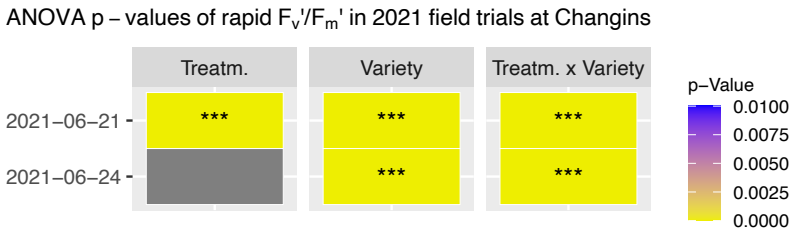


FIGURE S4.10: p-values of ANOVA on rapid F_v'/F_m' parameter without dark adaption at Changins 2021 for the factors "Treatm.", "Variety" and their interaction. Significance levels: NS: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p <$, and colors indicate significance at $p < 0.01$.

S4.6.7 Rapid F_v'/F_m' parameter without dark adaption, Changins 2021

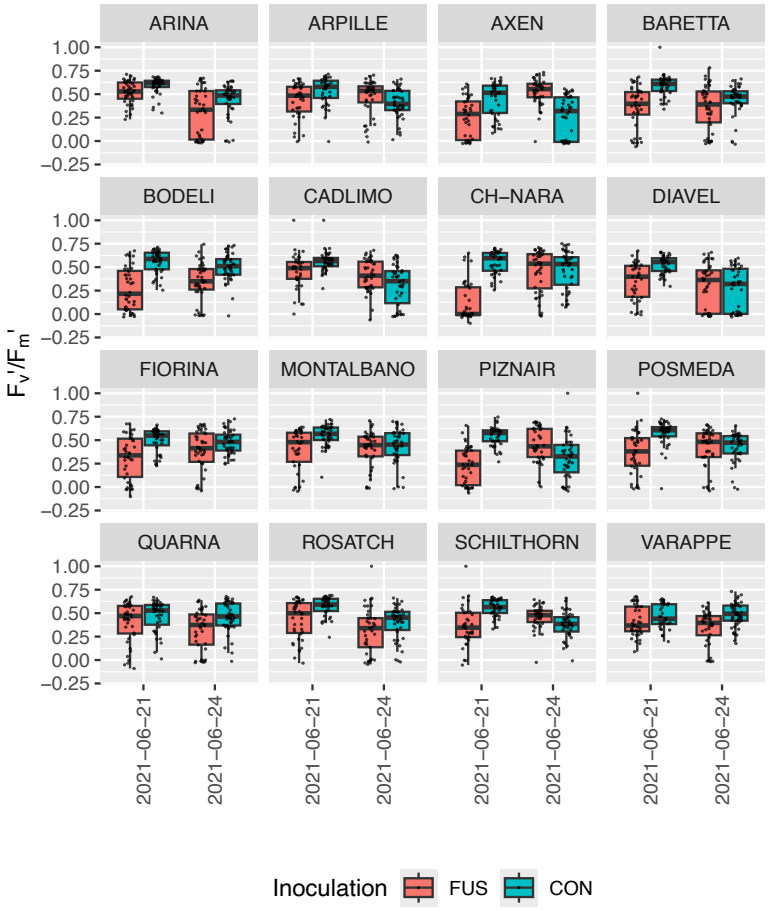


FIGURE S4.11: Rapid F_v'/F_m' parameter without dark adaption, Changins 2022. The 16 tiles represent the 16 wheat varieties tested over time. Dates are the individual measurement events. Inoculation treatments are indicated by color. 10 spikes were measured on a spikelet from the central spike for each plot, resulting in n = 40 measurements from four replication per variety and inoculation treatment for each date of measurements (1'280 measurements for each date in total).

S4.6.8 Infestation example

FIGURE S4.12: Example of the severity of *Fusarium* infestation for the variety MONTALBANO, taken on 2021-06-25. Spikes toward to opper part of the canopy are completely infested with the fungus.

List of Abbreviations

AIC	**A**kaike **I**nformation **C**riterion
ANOVA	**An**alysis **o**f **Va**riance
AUC	**A**rea **U**nder the **C**urve
AUDPC	**A**rea **U**nder the **D**isease-**P**rogress **C**urve
BBCH	**B**iologische Bundesanstalt, **B**undessortenamt und **CH**emische Industrie
BIC	**B**aysian **I**nformation **C**riterion
BRDF	**B**idirectional **R**eflectance **D**istribution **F**unction
CCD	**C**harge-**C**oupled **D**evice
CF	**C**hlorophyll **F**luorescence
CRS	**C**oordinate **R**eference **S**ystem
CSV	**C**omma **S**eparated **V**alues
CT	**C**anopy **T**emperature
CVEP	**C**rop **V**ariety **E**valuation **P**rograms
CV	**C**ross-**V**alidation
CWSI	**C**rop **W**ater **S**tress **I**ndex
DAI	**D**ays **A**fter **I**noculation
DAS	**D**ays **A**fter **S**owing
DEM	**D**igital **E**levation **M**odels
DN	**D**igital **Numbers**
DON	**D**e**o**xy**n**ivalenol
DUS	**D**istinctness, **U**niformity, and **S**tability
ELISA	**E**nzyme-**L**inked **I**mmuno**s**orbent **A**ssay
EU	**E**uropean **U**nion
FCC	**F**ractional **C**anopy **C**over
FDK	**F**usarium **D**amaged **K**ernels
FHB	**F**usarium **H**ead **B**light
FOV	**F**ield **O**f **V**iew
FPA	**F**ocal **P**lane **A**rray
GCP	**G**round **C**ontrol **P**oints
GIS	**G**eographic **I**nformation **S**ystem
GPS	**G**lobal **P**ositioning **S**ystem
GRDC	**G**rains **R**esearch and **D**evelopment **C**orporation
GSD	**G**round **S**ampling **D**istance
HTFP	**H**eigth **T**hroughput **F**ield **P**henotpying
HTPP	**H**eigth **T**hroughput **P**lant **P**henotpying
JPEG	**J**oint **P**hotographic **E**xperts **G**roup
LAI	**L**eaf **A**rea **I**ndex
LM	**L**inear **M**odel
Loess	**Lo**cally **E**stimated **S**catter Plot **S**moothing
MET	**M**ulti **E**nvironment **T**rial
MM	**M**ixed **M**odel
NVT	**N**ational **V**ariety **T**rials

PAM	**P**ulse-**A**mplitude-**M**odulation
PEP	**P**roof of **E**colotocal **P**erformance
PLSR	**P**artial **L**east **S**quares **R**egression
PSII	**P**hoto **S**ystem **II**
RCBD	**R**andomized **C**omplete **B**lock **D**esign
RFE	**R**ecursive **F**eature **E**limination
RGB	**R**ed **G**reen **B**lue (color space)
ROI	**R**egions **o**f **I**nterest
RSME	**R**oot **M**ean **S**quare **E**rror
RTK	**R**eal-**T**ime **K**inematic
SCTI	**S**tandardized **C**anopy **T**emperature **I**ndex
SD	**S**ecure **D**igital
sd	**S**tandard **D**eviation
SIFT	**S**cale **I**nvariant **F**eature **T**ransformation
TIFF	**T**agged **I**mage **F**ile **F**ormat
TIR	**T**hermal **I**nfra**r**ed
VCU	**V**alue for **C**ultivation and **U**se
VPD	**V**apor **P**ressure **D**efficit
WGS	**W**orld **G**eodetic **S**ystem
GIS	**G**eographical **I**nformation **S**ystem

Acknowledgments

The duration of a doctorate can be expressed in many ways. During my doctorate, I stood in the field in four seasons for plant ratings, flying drones, and different measurements. I was once barked at by a closeby fox during night-time measurements (he was probably just as scared of me as I was of him). I circled the globe more than three times in the train, at least measured by distance. I drank about 5'000 coffees to handle smaller daily crisis, undertook larger hikes to overcome more pronounced lows, and visited four very inspiring conferences to finally write the four chapters of a thesis.

During this journey, I experienced the support of many people without whom this thesis would not have been possible. First, I'd like to thank Dr. Juan M. Herrera, who entrusted me with this project and enabled me to spend five exciting and instructive years at Agroscope and ETH. I was able to benefit from his profound knowledge of agronomy and his rich experience in science. In countless exchanges on, physiology, chains of argumentation, writing, etc., which very often started with an unannounced appearance of mine in his office, he took the time to listen and very often showed me a trace to resolve my questions. He gave me a great deal of freedom, but at the right time he made sure that I had little directional adjustments. I am very grateful to him for this opportunity and his persistent support. I also owe my gratitude to the head of the ETH research group in which I conducted my doctorate, Prof. Dr. Achim Walter. I appreciated his clear and immediate feedback, his positive and motivating communication, and his great contribution to the very pleasant atmosphere in the research group. Furthermore, Dr. Lukas Roth contributed invaluably to the success of this work with his patience in dealing with the vast amount of my questions, his keen understanding of the topic, and especially the statistical issues that came with it. His input poured into this work through many exchanges, but especially through his outstanding and conscientious feedback on the writing. In addition, he acted as a mediator when putting reviewer feedback into terms that were acceptable to me, which was very beneficial to my inner balance. I also thank our former group leader at Agroscope, Dr. Didier Pellet, who was always very supportive of my endeavors and helped to overcome numerous organizational hurdles.

During my thesis I was part of two teams, at Agroscope and ETH and I would like to thank all present and former members of the "Production Technology & Cropping Systems Group" of Agroscpoe and the Crop Science group at ETH. A very particular thank you goes to Dr. Mahnaz Katouzi and Dr. Xavier Bousselin with whom I spent many pleasant hours and had exciting exchanges. I also would like to highlight the senior scientists PD Dr. Andreas Hund and Dr. Lilia Levy Häner, who provided me with advice and support at numerous occasions. A special thanks goes to Marianne Wettstein for helping with many administrative matters at ETH but mainly for always having some welcoming words. Finally, I extend my gratitude to Dr. Norbert Kirchgessner for many exciting and enjoyable conversations. Even if he may not remember, a small side note in such a conversation inspired important aspects of this work.

I thank Margot Visse-Mansiaux for the great teamwork when setting up the experiments, during trial management, and in nightly harvest processing sessions. Nicolas Vuille-dit-Bille was a great help for drone flights, but was also always open to a fruitful exchange about the collection and processing of drone data, which made him an indispensable sparring partner to my projects.

Whenever I needed advice on a particularly intricate aspect of breeding or wheat physiology, I could count on instructive discussions with Dr. Dario Fossati. These were not only extremely enlightening but in many cases also very motivating.

Good experimental field data is essential to any agricultural research, which comes with a tremendous amount of work. I therefore owe sincere thanks to Johanna Antretter, Matthias Schmid and Fernanda Arelmann Steinbrecher, who gave great support to the field work and measurements, who worked very conscientiously, and who did not lose their motivation even under the blazing sun of Changins. It is due to such people that an experiment can finally lead to meaningful and interpretable data. In addition to providing great work, they always spread a friendly and pleasant atmosphere. I also acknowledge Ulysse Schaller and Julien Vaudroz for their support in field ratings, and Matthieu Nussbaum for finally constructing an effective bird protection for the PhenoCams. Nicolas Widmer shared a lot of practical knowledge with me, and a big thanks goes to him and his farm team as well as Yann Imhoff for taking care of the field management. I am very grateful to Dr. Flavio Foiada and the whole team at Delley Samen und Pflanzen (DSP) AG for repeatedly providing me with seeds in a very uncomplicated way. I also thank Dr. Fabio Mascher and his former plant pathology team with Stefan Kellenberger, Alain Handley-Cornillet and Amandine Fasel for their advice and practical support in many aspects of the *Fusarium* experiments.

Although all of the help mentioned so far was unquestionably necessary for the success of this work, it was not sufficient, and without great support from family and friends, this doctorate would not have been possible. Therefore, the last section of this thesis is dedicated to them. A distinguished thanks goes first and foremost to Deborah Treier-Gerber, who was by my side throughout the doctorate. I will always remember the walks we took through the countless beautiful places in Fribourg. Sometimes, they helped to regain motivation, often, they were small celebrations of little successes. In many tricky situations, Deborah helped me choose a well-considered strategy over a hasty reaction. These years came with privations, and yet I always knew that she had my back! I would also like to thank our daughter Anna. It will be quite a while before she speaks her first words, nevertheless she has helped me to focus on the more relevant aspects of finishing a thesis ever since we knew about her. Quite some time ago, my brother Andreas Treier got the ball rolling, which led to my studies in agronomy and ultimately to this thesis. I am very thankful to him and my other brother Michael Treier for their positive and encouraging attitude towards my studies. To my mother Susanne Treier, I owe the intrinsic motivation and persistence to take this long but exciting and, therefore, rewarding path. My father Remy Treier has bolstered me along important stages that led to this doctorate in the most straightforward manner. I am so grateful to them from the bottom of my heart and wish they were still around so I could let them know.

Curriculum vitae

Simon Philip Treier

Date of birth	1985-12-11
Nationality	Swiss
Citizen of	Wölflinswil AG

Education

06/2020 - 05/2025	**Doctoral studies in Agricultural Sciences** at ETH Zürich, Switzerland
11/2021 - 05/2022	**Certificate of advanced studies in Geographic Information Systems** at ETH Zürich, Switzerland
09/2017 - 02/2020	**Master of Science in Agricultural Sciences** at ETH Zürich, Switzerland
09/2014 - 09/2017	**Bachelor of Science in Agricultural Sciences** at ETH Zürich, Switzerland
02/2007 - 07/2010	**General qualification for University entrance** at interstaatliche maturitätsschule für Erwachsene, St. Gallen, Switzerland
08/2001 - 08/2005	Professional training as **Polymechanic** (high-precision production and assembly allrounder) at Saurer Hamel AG, Arbon, Switzerland

Professional experience

11/2019 - 05/2020	**Scientific assistant**, crops science group at ETH Zürich, Switzerland
06/2018 - 03/2019	**Internship wheat breeding and remote sensing in variety testing** at Agroscope (Cultivation Techniques and Varieties in Arable Farming Group and Wheat breeding team), Nyon, Switzerland
08/2016 - 08/2018	**Teaching assistant** at Sustainable Agroecosystems Group, ETH Zürich, Switzerland
10/2012 - 08/2014	**Prototyping and manufacturing mechanic** at Mirad Microwave AG, Wittenbach, Switzerland
02/2012 - 09/2012	**Prototyping and assembly mechanic** at Pantec GS Systems, Kradolf, Switzerland
04/2006 - 04/2011	**Prototyping and manufacturing mechanic** at Mirad Microwave AG, Wittenbach, Switzerland

First author publication

Simon Treier, Juan M. Herrera, Andreas Hund, Norbert Kirchgessner, Helge Aasen, Achim Walter, and Lukas Roth (Dec. 2024). Improving

drone-based uncalibrated estimates of wheat canopy temperature in plot experiments by accounting for confounding factors in a multi-view analysis. ISPRS Journal of Photogrammetry and Remote Sensing, 218, pp. 721–741. https://doi.org/10.1016/j.isprsjprs.2024.09.015

Simon Treier, Lukas Roth, Andreas Hund, Helge Aasen, Lilia Levy Häner, Nicolas Vuille-dit-Bille, Achim Walter, Juan M. Herrera (Apr. 2025). Analysis of variance and its sources in UAV-based multi-view thermal imaging of wheat plots. Plant Phenomics, 7.2, pp. 100046. https://doi.org/10.1016/j.plaphe.2025.100046

Simon Treier, Nicolas Vuille-dit-Bille, Margot Visse-Mansiaux, Frank Liebisch, Helge Aasen, Lukas Roth, Achim Walter and Juan M. Herrera (Jul. 2025). Comparison of PhenoCams and drones for lean phenotyping of phenology and senescence of wheat genotypes in variety testing. The Plant Phenome Journal, 8, pp. e70039. https://doi.org/10.1002/ppj2.70039

Contributions to publications

Jonas Anderegg, Flavian Tschurr, Norbert Kirchgessner, **Simon Treier**, Manuel Schmucki, Bernhard Streit, Achim Walter. On-farm evaluation of UAV-based aerial imagery for season-long weed monitoring under contrasting management and pedoclimatic conditions in wheat. Computers and Electronics in Agriculture 204, 107558 (2023). https://doi.org/10.1016/j.compag.2022.107558

Jonas Anderegg, Flavian Tschurr, Norbert Kirchgessner, **Simon Treier**, Lukas Valentin Graf, Manuel Schmucki, Nicolin Caflisch, Camille Minguely, Bernhard Streit & Achim Walter. Pixel to practice: multi-scale image data for calibrating remote-sensing-based winter wheat monitoring methods. Scientific Data 11, 1033 (2024). https://doi.org/10.1038/s41597-024-03842-8

Olten, 2025-09-17