ISPRS Journal of Photogrammetry and Remote Sensing 218 (2024) 721–741

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

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

Simon Treier [a,b,*], Juan M. Herrera [a], Andreas Hund [b], Norbert Kirchgessner [b], Helge Aasen [c], Achim Walter [b], Lukas Roth [b]

[a] Cultivation Techniques and Varieties in Arable Farming Group, Agroscope, Route de Duillier 50, Nyon, 1260, Switzerland
[b] ETH Zürich, Institute of Agricultural Sciences, Universitätstrasse 2, Zürich, 8092, Switzerland
[c] Earth Observation of Agroecosystems Team, Agroecology and Environment Division, Agroscope, Reckenholzstrasse 191, Zürich, 8046, Switzerland

ARTICLE INFO

Dataset link: https://github.com/TreAgron/ThermalMultiviewExample.git

Keywords:
Plant phenotyping
Aerial thermography
Thermal drift
Drift correction
High throughput field phenotyping
Viewing geometry

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.

1. Introduction

Canopy temperature (CT) of wheat (*Triticum aestivum* L.) is an integrative trait "being associated with yield in a range of conditions" (Reynolds et al., 2012). "It is indicative of the relative fitness of a genotype to the environment" (Reynolds 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 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; Reynolds et al., 2012). Therefore, CT can be used as an indirect selection criterion for yield (*e.g.* Das et al., 2021a).

* Corresponding author at: Cultivation Techniques and Varieties in Arable Farming Group, Agroscope, Route de Duillier 50, Nyon, 1260, Switzerland.
E-mail addresses: simon.treier@agroscope.admin.ch (S. Treier), juan.herrera@agroscope.admin.ch (J.M. Herrera), andreas.hund@usys.ethz.ch (A. Hund), norbert.kirchgessner@usys.ethz.ch (N. Kirchgessner), helge.aasen@agroscope.admin.ch (H. Aasen), achim.walter@usys.ethz.ch (A. Walter), lukas.roth@usys.ethz.ch (L. Roth).

https://doi.org/10.1016/j.isprsjprs.2024.09.015
Received 22 December 2023; Received in revised form 15 August 2024; Accepted 14 September 2024
Available online 11 October 2024

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 et al., 2016; Pask et al., 2012). Main sources of short-term variability in environmental conditions include wind, sunlight, clouds, and air temperature (Reynolds et al., 2012). Thus, genotypes should be measured within a short period, *e.g.* within 30 min (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 (*i.e.* a higher 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 lightweight drone (Deery 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 (Jones 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; 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 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 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 (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 et al., 2016) or orthomosaics, *i.e.* large composite images of a series of images with large overlap (*e.g.* Das et al., 2021b; Francesconi et al., 2021; Malbéteau et al., 2021; Messina and Modica, 2020; Perich et al., 2020; 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 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 variation 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 variation, 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; Piepho and Williams, 2010; 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 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 et al. (2021a) 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. Methods

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; de Cárcer et al., 2019).

Air temperature, rainfall, radiation, wind speed, wind direction, relative humidity and vapor pressure deficit (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. A2). 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 × 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. A1).

The two experiment-year combinations are further referred to as EuVar21 and EuVar22 according to year of harvest. Table A1 gives an overview on the different treatments and the most important field interventions and Table A2 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 Appendix section A5).

A description of the equipment and the settings used and of the flight planning can be found in Appendix section A6. 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. Appendix section A7 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 JPG 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 Appendix section A8).

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. 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. 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. 1). TIR data was then treated by different statistical approaches (blue section) and the approaches were compared to each other (violet section).

2.3. TIR image pre-processing

Radiometric JPG format contains an 8-bit gray scale JPG 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 JPGs were transformed to TIFF files representing temperature in °C × 1000 by using a Python 3.8

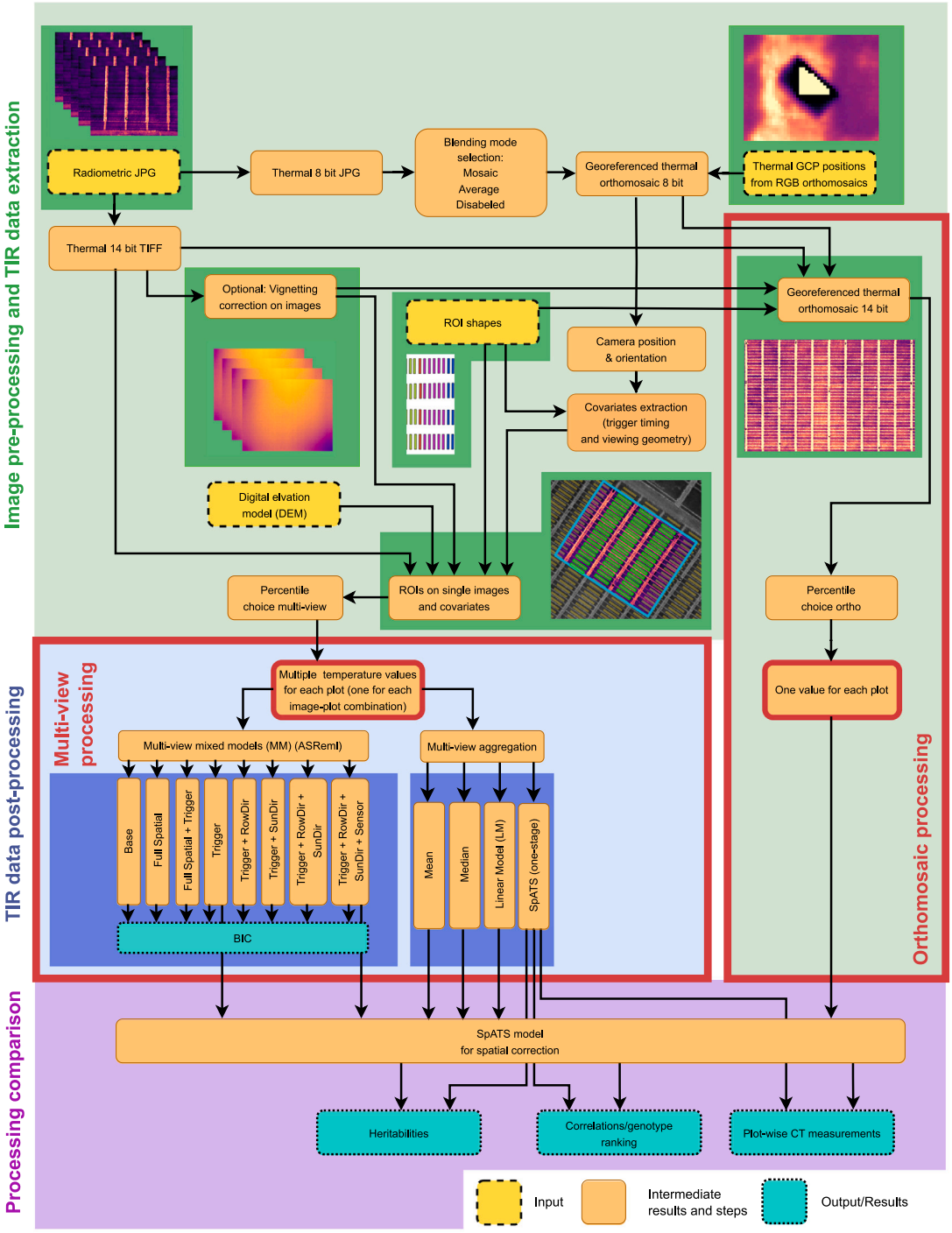


Fig. 1. Overview on the different steps of TIR data processing methods that were compared in this study. Orthomosaics were composed by different blending modes. After image pre-processing (green section), TIR information was analyzed on orthomosaics or with different multi-view approaches (blue section). 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

script (van Rossum and Drake, 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 Appendix section A9). 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. (10), Eq. A1, Eq. A2) of different percentiles was calculated for each flight. The 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 et al., 2020; Deery 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 Appendix section A10). The image corresponding to a mean estimated vignetting effect was subtracted from all the TIR images to get vignetting-corrected images (*e.g.* Figs. A3 & A4). Subsequent analysis was conducted on images with and without vignetting correction for both, the orthomosaic and multi-view methods.

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). 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.

2.5. 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.

2.6. 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) by ray tracing as described in Roth et al. (2018) and Roth 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 JPG 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 1 and illustrated in Fig. 3.

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. 1). In the following, the different processing steps of the different methods are described. The presentation of single steps follows the structure of Fig. 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.

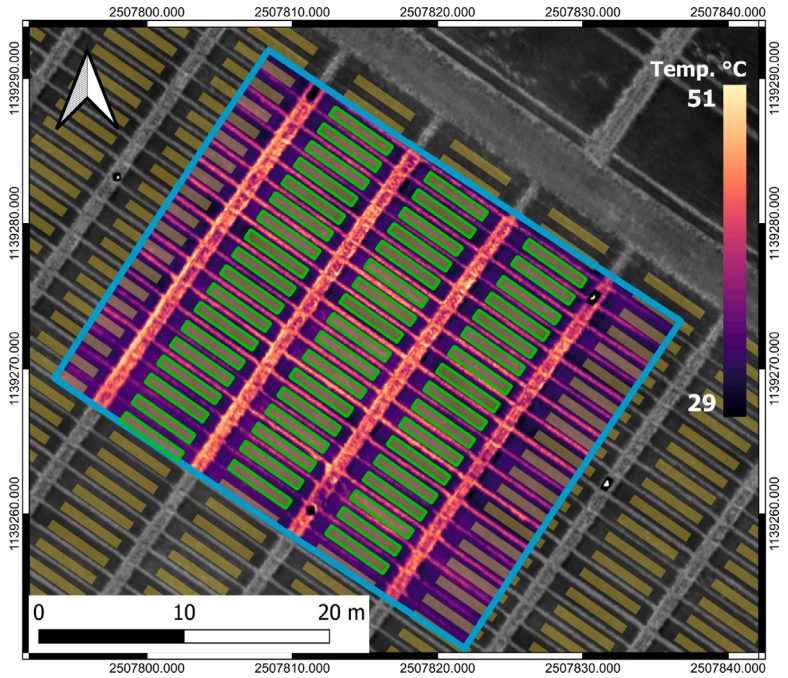
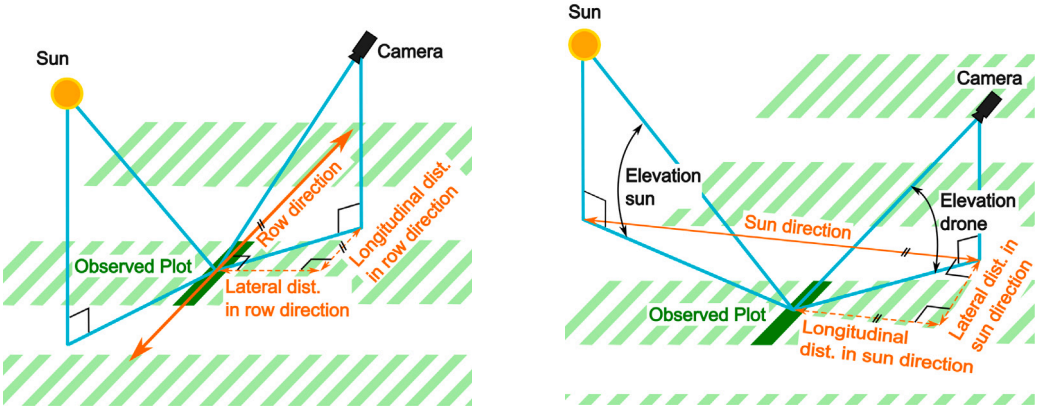


Fig. 2. Example of a TIR image, projected on a DEM. The DEM (gray-scale, in the background) defined the surface on which the TIR image (blue margin) was projected on. Plots were defined for the whole field (shaded in yellow). Plot shapes that fell entirely within the extent of the TIR image (green margins) were projected to image coordinates and all plot-wise TIR percentiles were extracted. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 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



(a) Plot position in row direction (b) Plot position in sun direction

Fig. 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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. 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),

$$\hat{\theta}_{p\_mean} = \text{mean}(\theta_{jp}). \quad (1)$$

$$\hat{\theta}_{p\_median} = \text{median}(\theta_{jp}). \quad (2)$$

A more complex way is to correct for the effect of trigger timing with a simple linear model simultaneously for all images (Eq. (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,

$$\theta_{jp} = \nu_j + \phi_p + e_{jp}. \quad (3)$$

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,

$$\hat{\theta}_{p\_LM} = \phi_p. \quad (4)$$

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.

2.7.2. 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 Piepho, 2007; Wu, 2010).

Mixed models of varying complexity were fitted in ASReml-R (Butler, 2019). The parameters of the mixed model (Eq. (5)) are explained in Table 2. The terms were grouped by types of terms (“Design-Factors”, “Spatial-Autoregression”, “Spatial-Smoothing-Spline”, etc.). Note that not all models included all term types. Table 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 1). This led to nested models where simpler models were fully included in more complex models, culminating in the most complex model,

$$\begin{aligned}
\theta_{i(j)knp} = {} & \theta_i + \tau_k + \phi_p + r_n + (\tau r)_{kn} + && \text{(Design-Factors)}\\
& (\alpha\beta)_{c(p)r(p)} + \alpha_{c(p)} + \beta_{r(p)} + && \text{(Spatial-Autoregression)}\\
& f_{\text{spl}\times\text{spl}}\left(c(p), r(p)\right) + f_{\text{spl}}(c(p)) + f_{\text{spl}}(r(p)) + && \text{(Spatial-Smoothing-Spline)}\\
& f_{\text{spl}}(j) + && \text{(Temporal-Trend)}\\
& f_{\text{spl}\times\text{spl}}(\lambda_{\text{lon,Row,jp}}, \lambda_{\text{lat,Row,jp}}) + && \text{(Row-Direction-Trend)}\\
& f_{\text{spl}\times\text{spl}}(\lambda_{\text{lon,Sun,jp}}, \lambda_{\text{lat,Sun,jp}}) + && \text{(Sun-Direction-Trend)}\\
& f_{\text{spl}\times\text{spl}}(s_{x,jp}, s_{y,jp}) + && \text{(Sensor-Plane-Trend)}\\
& e_{i(j)knp} && \text{(Residuals)}
\end{aligned} \quad (5)$$

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. 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. (5)), plot-wise CT values were estimated in a similar approach as for the other, simpler models (Eq. (1), (2) & (4)). Specifically, $\hat{\theta}_{p\_MM}$ were estimated as sum of genotype effects

Table 2
Terms of the mixed models (Eq. (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
	$\alpha_{c(p)}$	One-dimensional autocorrelation model for columns in the field (orthogonal to tractor track direction)	Random
	$\beta_{r(p)}$	One-dimensional autocorrelation model for rows in the field (in tractor track direction)	Random
Spatial-Smoothing-Spline:	$f_{\text{spl×spl}}(c(p), r(p))$	Two-dimensional spatial smoothing spline model based on row and column position in the field	Random
	$f_{\text{spl}}(c(p))$	One-dimensional smoothing spline model for columns in the field (orthogonal to tractor track direction)	Random
	$f_{\text{spl}}(r(p))$	One-dimensional smoothing spline model for rows in the field (in tractor track direction)	Random
Temporal-Trend:	$f_{\text{spl}}(j)$	Trigger timing smoothing spline along the j sequential trigger events	Random
Row-Direction-Trend:	$f_{\text{spl×spl}}(\lambda_{\text{lon,Row,jp}}, \lambda_{\text{lat,Row,jp}})$	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:	$f_{\text{spl×spl}}(\lambda_{\text{lon,Sun,jp}}, \lambda_{\text{lat,Sun,jp}})$	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:	$f_{\text{spl×spl}}(s_{x,jp}, s_{y,jp})$	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:	$e_{i(j)knp}$	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

(θ_i), treatment effects (τ_k), plot effects (ϕ_p), and replication effects (r_n),

$$\hat{\theta}_{p\_MM} = \theta_i + \tau_k + \phi_p + r_n \ . \tag{6}$$

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.

2.8. 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. 1, violet section).

The plot values of the orthomosaic method, the aggregated plot-wise multi-view values (Eq. (1), (2)), the multi-view values estimated with the LM (Eq. (4)) and the plot-wise results from the CT estimations with the mixed models (Eq. (6)) were first fitted with a spatial model (Eq. (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 (Jones 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 (Jones et al., 2009; Prashar and Jones, 2014; Das et al., 2021c).

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.

$$\begin{aligned}\hat{\theta}_p = \hat{\theta}_{iknp} = \theta_i + \tau_k + \left(\theta_n\tau_n\right)_{ik} + \phi_p + r_n + & \quad \text{(base model)}\\ \tau r_{kn} + & \quad \text{(repl. × treat. (just EuVar))}\\ f\left(c(p), r(p)\right) + \psi_{c(p)} + \psi_{r(p)} + & \quad \text{(spatial model)}\\ e_{iknp} & \quad \text{(error term)}\end{aligned} \tag{7}$$

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. (5) and can be looked up in Table 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. (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

Table 3
Term type combinations used in plot-wise CT estimation with mixed models (Eq. (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. The prefix "MM" 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 autocorrelated interaction term and a bi-variate 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.

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,

$$\hat{\theta}_{p\_SpATS} = \theta_i + \tau_k + e_{iknp} \ . \tag{8}$$

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; Jones 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:

$$\sigma_{gen\_r} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \ , \tag{9}$$

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. (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:

$$H_s^2 = \frac{\sigma_g^2}{(\sigma_g^2 + \frac{\sigma_e^2}{r})}. \tag{10}$$

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. (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. (10) showcases the principles of heritability, for the interested reader, the formula framework of generalized heritability is provided in Eq. A1 & Eq. A2. 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 et al., 2021).

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. (7) for the effect of the j^{th} trigger event (Eq. (11)).

$$\begin{aligned}\hat{\theta}_{ijknp} = {} & \theta_i + \tau_k + \left(\theta_n \tau_n\right)_{ik} + \phi_p + r_n + && \text{(base model)} \\ & \tau r_{kn} + && \text{(repl. × treat.)} \\ & f\left(c(p), r(p)\right) + \psi_{c(p)} + \psi_{r(p)} + && \text{(spatial model)} \\ & \nu_j + && \text{(trigger timing)} \\ & e_{ijknp} && \text{(error term)}\end{aligned} \tag{11}$$

As with the other approaches, plot-wise CT values were then predicted with Eq. (8) for comparison.

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 3) and SpATS according to Eq. (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. (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.

3. Results

3.1. TIR data processing and processing comparison

3.1.1. Example of selected correction steps

Fig. 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 baseline 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. 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 $\hat{\theta}_{p\_Ortho}$ and $\hat{\theta}_{p\_mean}$ show similar general trends of ranking between the two methods but not compared to $\hat{\theta}_{p\_MM}$.

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. A5b & Fig. A6b) and multi-view method (Fig. A7).

The median (*i.e.* the 50th 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. A5b & Fig. A6b) and multi-view (Fig. A7) methods were small. Hence, the 50th 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.

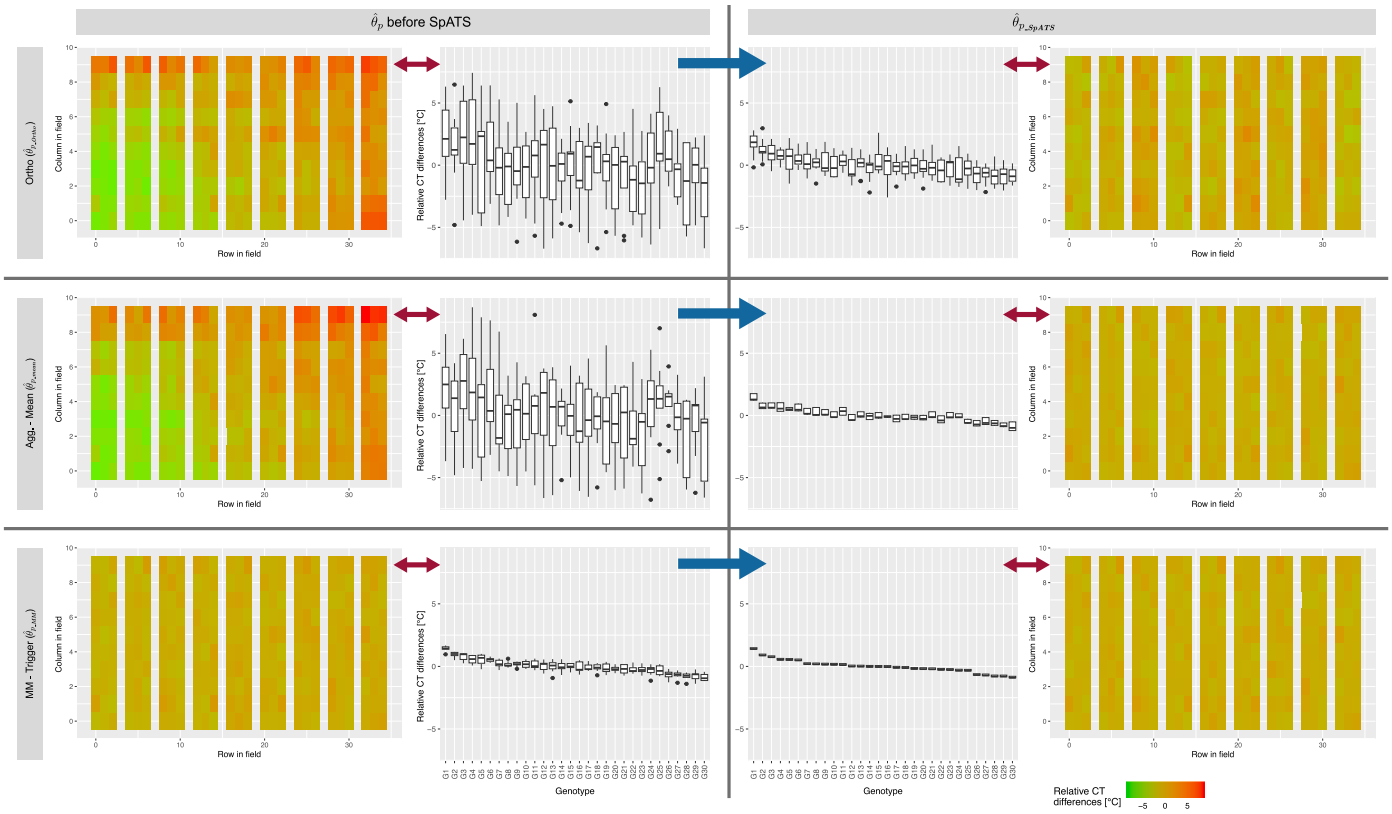


Fig. 4. Data visualization of CT differences from mean CT within the first flight of the campaign on 2022-05-18 at 16.00. CT was estimated with the orthomosaic method ($\hat{\theta}_{p\_Ortho}$), the "Agg.-Mean" aggregation method ($\hat{\theta}_{p\_mean}$) and the "MM Trigger" mixed model ($\hat{\theta}_{p\_MM}$). Values are shown before spatial correction ($\hat{\theta}_p$ before SpATS) and after SpATS ($\hat{\theta}_{p\_SpATS}$). 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.

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. 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.

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. 6). Analyzing the estimated temperature drift with time ($f_{\mathrm{spl}}(j)$ in Eq. (5)) with the "MM Trigger" mixed model in relation to relative movements along the main flight direction (Fig. 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.

3.1.5. Consistency of plot-wise CT estimates and genotype CT ranking

As a metric of consistency, correlations of plot-wise values $\hat{\theta}_{p\_SpATS}$ 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 $\hat{\theta}_{p\_SpATS}$ 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 were taken at times further apart. These patterns were consistent over both years (Figs. 7 & A8).

Mean plot-wise correlations of CT measurements over all dates were calculated for different CT pre-processing and post-processing methods (similar to Fig. 7) and correlations aggregated in box plots for comparison (Fig. 8(a)). 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. 7, A8). Vignetting correction almost did not change

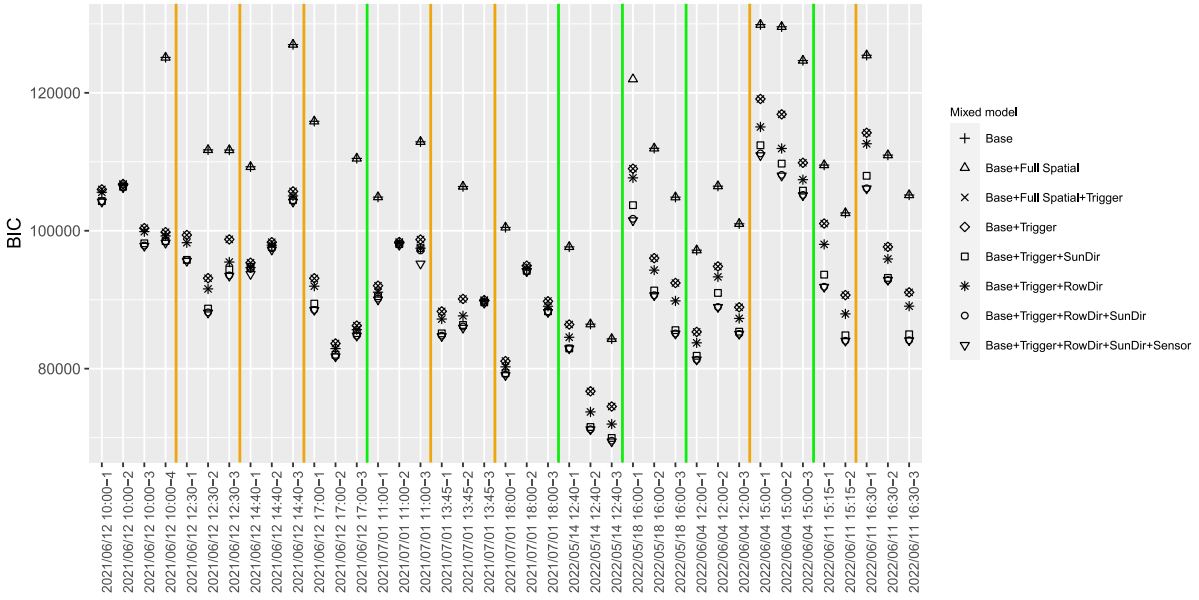


Fig. 5. The Bayesian information criterion (BIC) for all flights. With BIC, the quality of the model fit was compared. Lower BIC values indicate preferable models. Green lines separate different measurement days, orange lines different flight campaigns within the same day. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

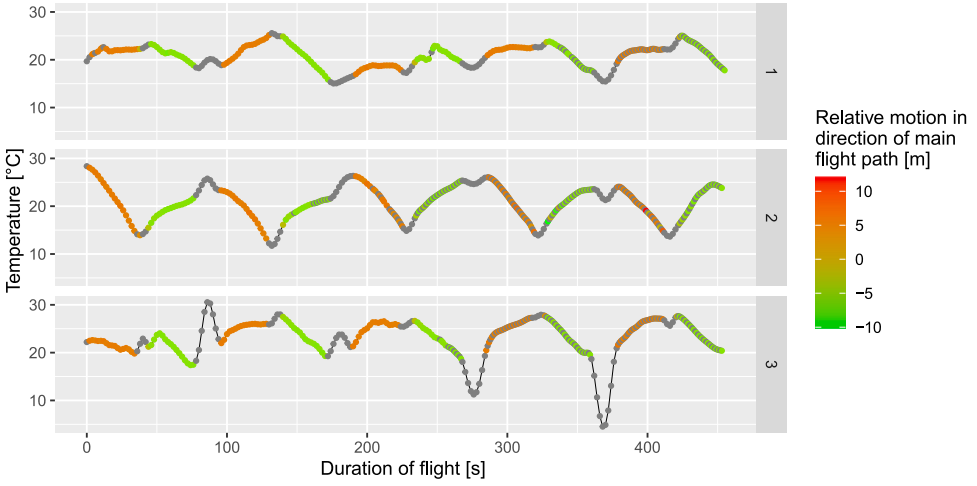


Fig. 6. Estimated thermal drift of TIR measurements throughout the duration of flights for the three flights of the 13.45 campaign on 2021-07-01 on EuVar21. Rows 1 to 3 represent the three different flights of the same campaign. Flight plan and sensor orientation were identical for the three flights 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the values, and the values mentioned are those without vignetting correction.

The sd of genotype ranking within campaigns σ_{gen_r} (Eq. (9)) was calculated for all processing methods (Fig. 8(b)). 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 A4.

3.1.6. Genotypic specificity of apparent CT

Heritabilities were generally high to very high (Fig. 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

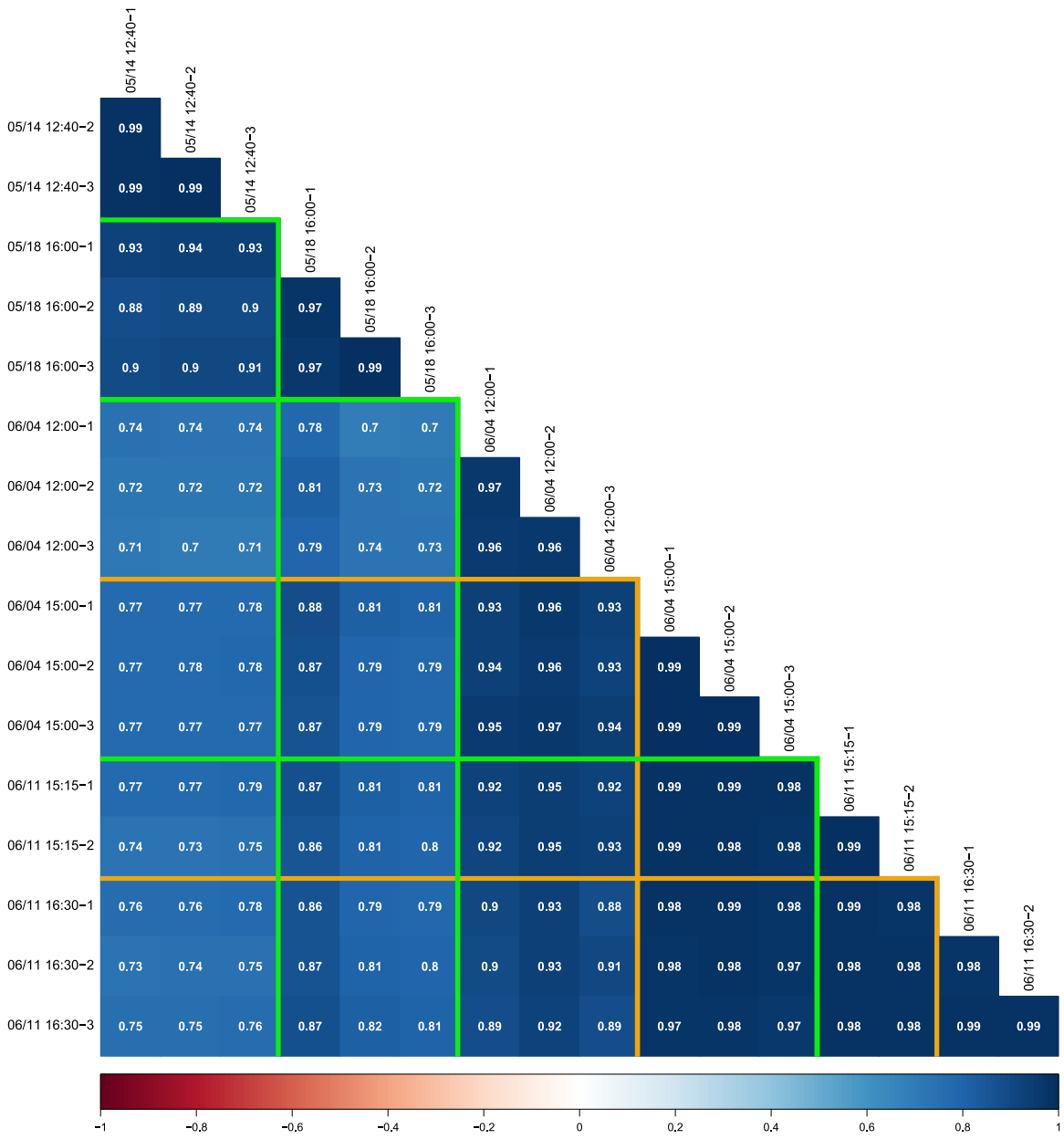


Fig. 7. Pearson's correlations of EuVar22 plot values $\hat{\theta}_{P_{SpATS}}$ after correction for spatial and temporal covariates ("MM Trigger + RowDir + SunDir + Sensor" and subsequent fitting with SpATS) and removing dominant treatment effects. Green lines separate different measurement days, orange lines different flight campaigns within the same day. All correlations are significant at $P < 0.001$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.

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

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. 10(a)). The correlation between the flights within one swath width of nadir-view exclusion got weaker in general with increasing swath width (Fig. 10(b)).

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. (11)). When comparing the resulting correlations and heritabilities in box plots, they consistently increased with increased number of observations for both years (Fig. 11(a)) 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. 11(b)).

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

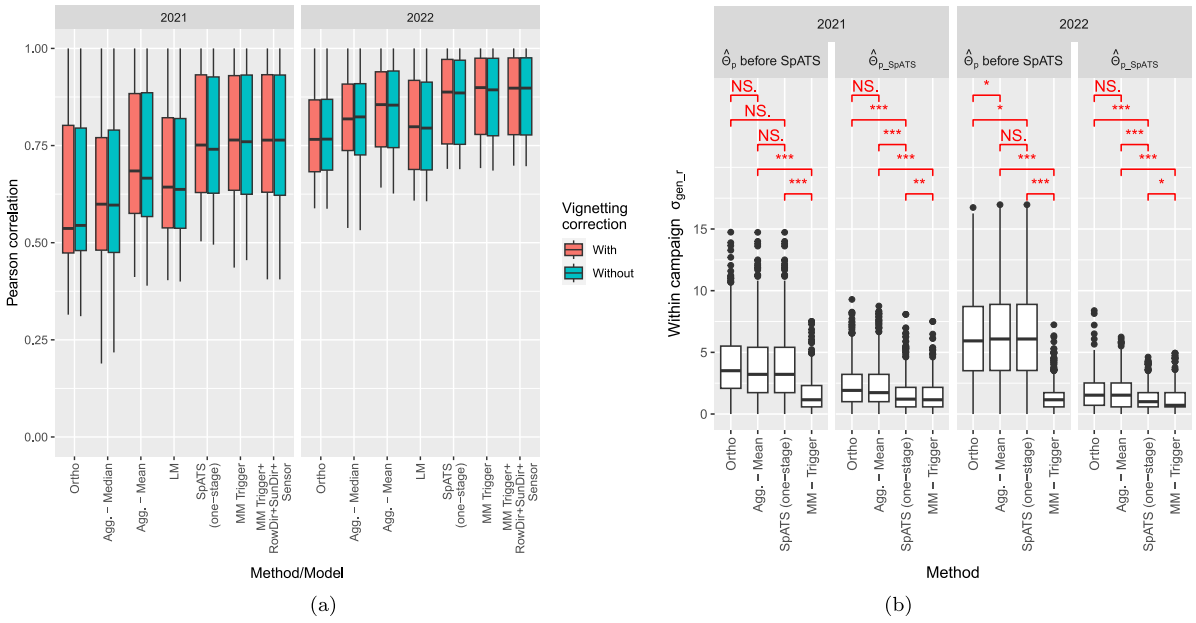


Fig. 8. Consistency of plot-wise CT estimates and genotype CT ranking. (a) Pearson's correlations of plot-wise CT measurements $\hat{\theta}_{P\_SpATS}$ within EuVar. Correlations were calculated for each flight within both years, but not across years. CT was estimated with the orthomosaic method, two 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. (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: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

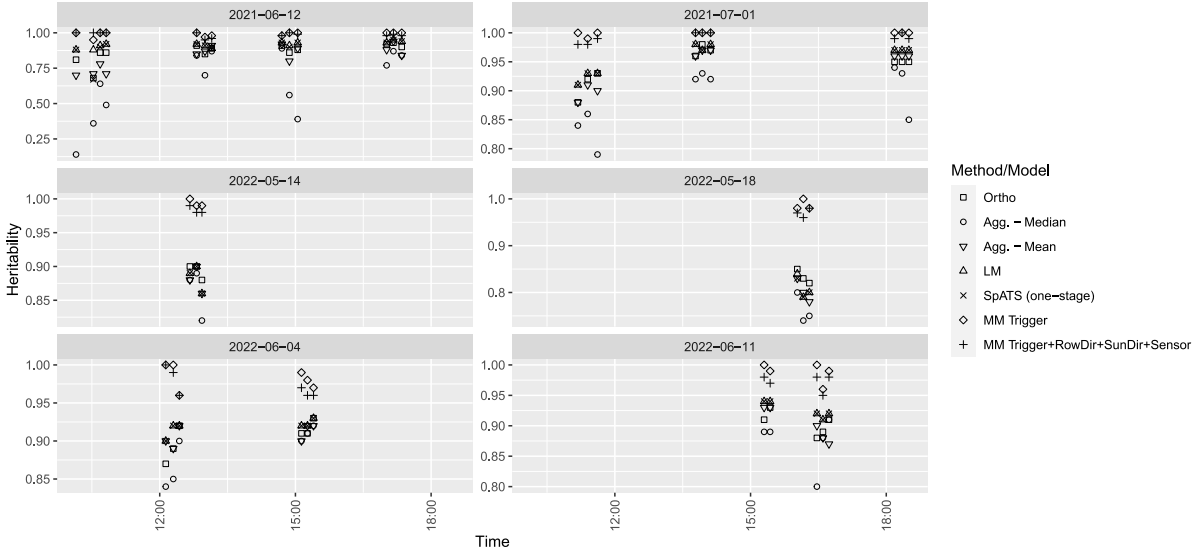


Fig. 9. Heritability for all flights on EuVar grouped by date and time. The shapes indicate the different methods and models used in plot-wise CT estimation. Each group of two to four flights within the different time slots represents a campaign. Note that the scale of heritability is varying between the plots to allow to represent very different value ranges between different dates.

radiation, wind speed, wind direction, relative humidity or VPD could be found (Fig. A9 and A10).

4. Discussion

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. 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. (10)). From the CT values arranged by genotypes, it therefore became evident that heritability

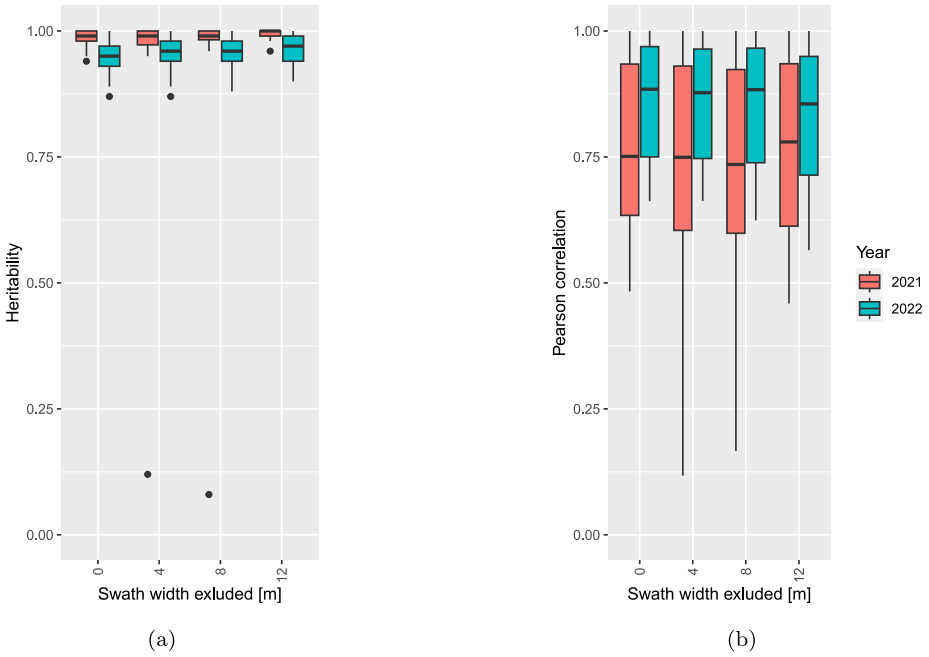


Fig. 10. Heritabilities (a) and the correlations (b) for all flights, for which data was excluded in nadir oriented swaths of different widths.

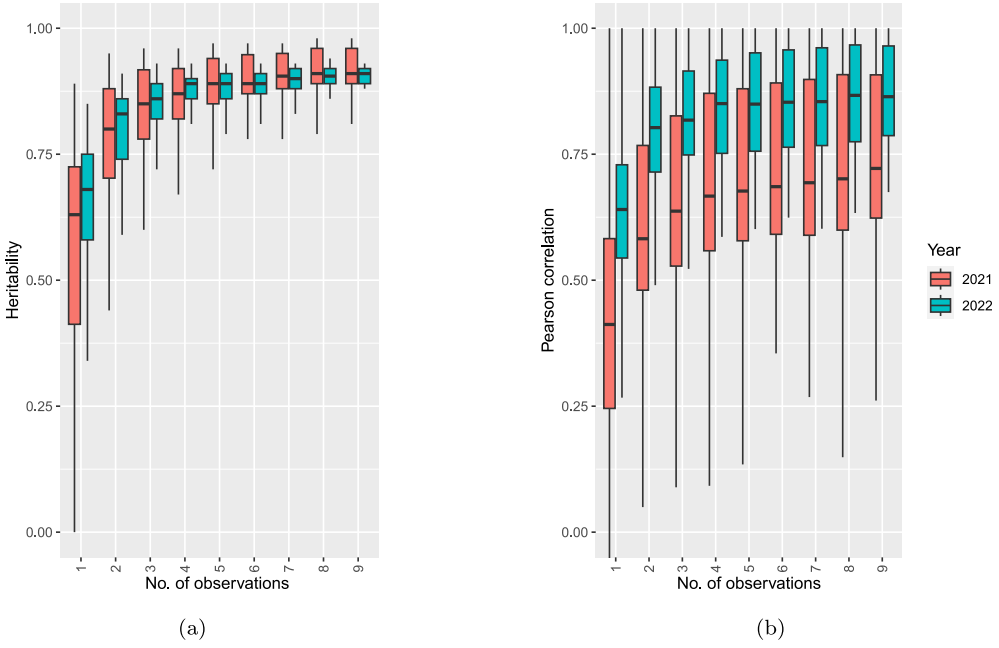


Fig. 11. Heritabilities (a) and the correlations (b) for all flights and specific numbers of observations per flight.

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. 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 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.

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 $2\,m\,s^{-1}$ 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 Berni 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.

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.

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.

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 (Jones 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 (Jones 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.

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 (Jones 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.

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. 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 analyzing experiments separately.

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 50th 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 A17 and A18, respectively.

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 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 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 et al. (2016, 2019) and Perich et al. (2020) used correlations between measurements at different times and heritabilities as quality criteria of the experiment, Jones 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

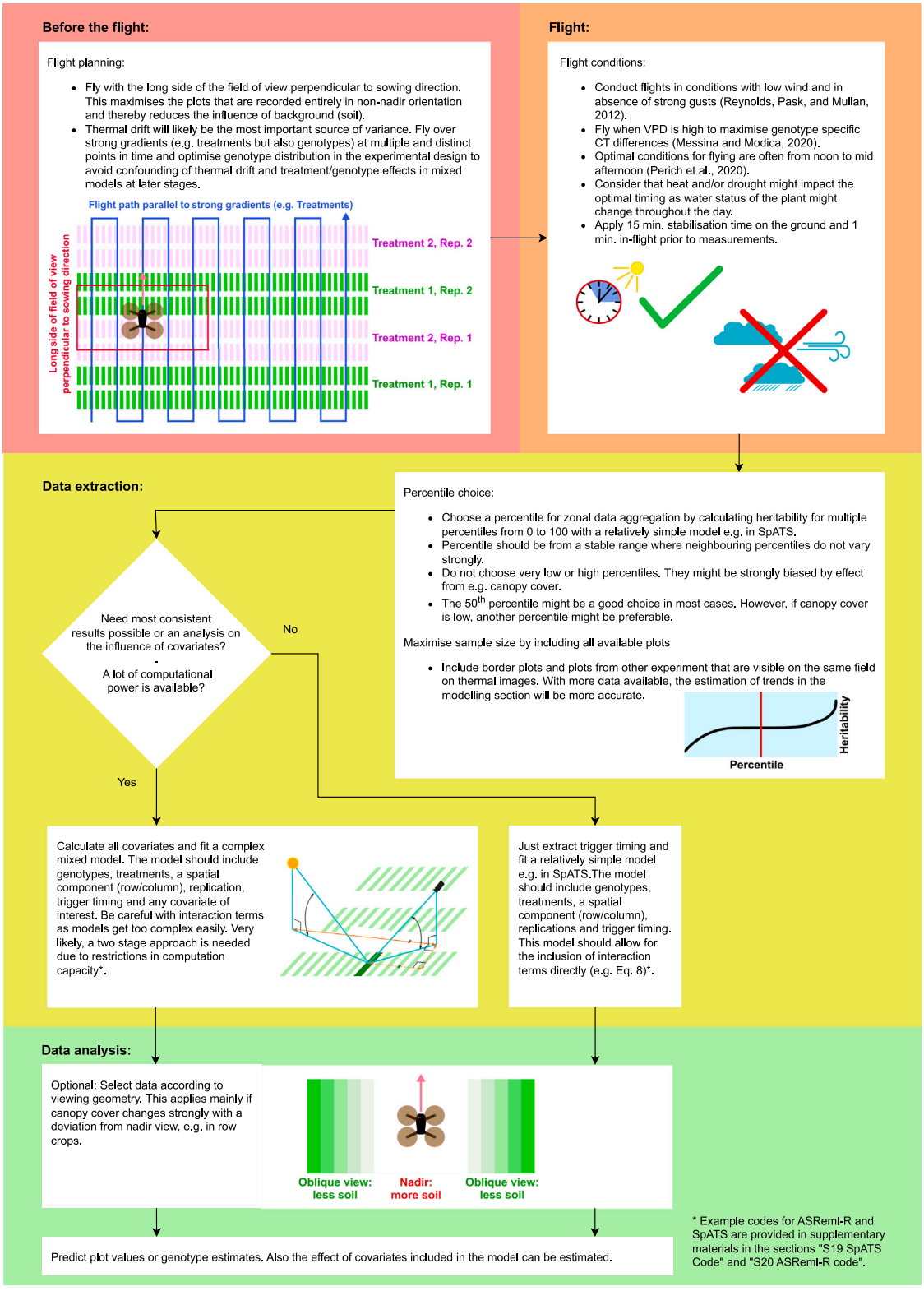


Fig. 12. Cheat sheet, giving an overview on most relevant considerations when measuring CT with a drone based multi-view approach.

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 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.

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. 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.

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).

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 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.

Funding

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

CRediT authorship contribution statement

Simon Treier: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Juan M. Herrera:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Andreas Hund:** Writing – review & editing. **Norbert Kirchgessner:** Writing – review & editing, Conceptualization. **Helge Aasen:** Writing – review & editing. **Achim Walter:** Writing – review & editing. **Lukas Roth:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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).

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; Marogot Visse-Mansiaux for support in setting up the experiments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2024.09.015.

References

Aasen, H., Bolten, A., 2018. Multi-temporal high-resolution imaging spectroscopy with hyperspectral 2D imagers – From theory to application. Remote Sens. Environ. 205, 374–389. http://dx.doi.org/10.1016/j.rse.2017.10.043.

Anderegg, J., Yu, K., Aasen, H., Walter, A., Liebisch, F., Hund, A., 2020. Spectral vegetation indices to track senescence dynamics in diverse wheat germplasm. Front. Plant Sci. 10, 1–20. http://dx.doi.org/10.3389/fpls.2019.01749.

Aragon, B., Johansen, K., Parkes, S., Malbeteau, Y., Al-Mashharawi, S., Al-Amoudi, T., Andrade, C.F., Turner, D., Lucieer, A., McCabe, M.F., 2020. A calibration procedure for field and UAV-based uncooled thermal infrared instruments. Sensors 20, 3316. http://dx.doi.org/10.3390/s20113316.

Baxter, S., 2007. World reference base for soil resources. World soil resources report 103. Rome: Food and Agriculture Organization of the United Nations (2006), p. 132. Exp. Agric. 43, 264. http://dx.doi.org/10.1017/S0014479706394902.

Benassi, F., Dall'Asta, E., Diotri, F., Forlani, G., Morra di Cella, U., Roncella, R., Santise, M., 2017. Testing accuracy and repeatability of UAV blocks oriented with GNSS-supported aerial triangulation. Remote Sens. 9, 172. http://dx.doi.org/10.3390/rs9020172.

Berni, J.A.J., Zarco-Tejada, P.J., Suarez, L., Fereres, E., 2009. Thermal and narrowband multispectral remote sensing for vegetation monitoring from an unmanned aerial vehicle. IEEE Trans. Geosci. Remote Sens. 47, 722–738. http://dx.doi.org/10.1109/TGRS.2008.2010457.

Blum, A., Mayer, J., Gozlan, G., 1982. Infrared thermal sensing of plant canopies as a screening technique for dehydration avoidance in wheat. Field Crops Res. 5, 137–146. http://dx.doi.org/10.1016/0378-4290(82)90014-4.

Butler, D., 2019. asreml: Fits the Linear Mixed Model. R package version 4.1. 0.110. VSN International Ltd., Hemel Hempstead, UK.

Chandel, N.S., Rajwade, Y.A., Dubey, K., Chandel, A.K., Subeesh, A., Tiwari, M.K., 2022. Water stress identification of winter wheat crop with state-of-the-art AI techniques and high-resolution thermal-RGB imagery. Plants 11, 3344. http://dx.doi.org/10.3390/plants11233344.

Cullis, B.R., Smith, A.B., Coombes, N.E., 2006. On the design of early generation variety trials with correlated data. J. Agric. Biol. Environ. Stat. 11, 381–393. http://dx.doi.org/10.1198/108571106X154443.

Das, S., Chapman, S., Christopher, J., Choudhury, M.R., Menzies, N.W., Apan, A., Dang, Y.P., 2021a. 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. Remote Sens. Appl.: Soc. Environ. 23, 100583. http://dx.doi.org/10.1016/j.rsase.2021.100583.

Das, S., Christopher, J., Apan, A., Choudhury, M.R., Chapman, S., Menzies, N.W., Dang, Y.P., 2021b. Evaluation of water status of wheat genotypes to aid prediction of yield on sodic soils using UAV-thermal imaging and machine learning. Agricult. Forest. Meterol. 307, 108477. http://dx.doi.org/10.1016/j.agrformet.2021.108477.

Das, S., Christopher, J., Apan, A., Ro. Choudhury, M., Chapman, S., Menzies, N.W., Dang, Y.P., 2021c. UAV-thermal imaging and agglomerative hierarchical clustering techniques to evaluate and rank physiological performance of wheat genotypes on sodic soil. ISPRS J. Photogramm. Remote Sens. 173, 221–237. http://dx.doi.org/10.1016/j.isprsjprs.2021.01.014.

de Cárcer, P.S., Sinaj, S., Santonja, M., Fossati, D., Jeangros, B., 2019. Long-term effects of crop succession, soil tillage and climate on wheat yield and soil properties. Soil Tillage Res. 190, 209–219. http://dx.doi.org/10.1016/j.still.2019.01.012.

Deery, D.M., Rebetzke, G.J., Jimenez-Berni, J.A., Bovill, W.D., James, R.A., Condon, A.G., Furbank, R.T., Chapman, S.C., Fischer, R.A., 2019. Evaluation of the phenotypic repeatability of canopy temperature in wheat using continuous-terrestrial and airborne measurements. Front. Plant Sci. 10, 875. http://dx.doi.org/10.3389/fpls.2019.00875.

Deery, D.M., Rebetzke, G.J., Jimenez-Berni, J.A., James, R.A., Condon, A.G., Bovill, W.D., Hutchinson, P., Scarrow, J., Davy, R., Furbank, R.T., 2016. Methodology for high-throughput field phenotyping of canopy temperature using airborne thermography. Front. Plant Sci. 7, http://dx.doi.org/10.3389/fpls.2016.01808.

Francesconi, S., Harfouche, A., Maesano, M., Balestra, G.M., 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. Front. Plant Sci. 12, 1–19. http://dx.doi.org/10.3389/fpls.2021.628575.

Gilmour, A.R., Cullis, B.R., Verbyla, A.P., 1997. Accounting for natural and extraneous variation in the analysis of field experiments. J. Agric. Biol. Environ. Stat. 26, 9–293. http://dx.doi.org/10.2307/1400446.

Hartung, K., Piepho, H.P., 2007. Are ordinal rating scales better than percent ratings? A statistical and "psychological" view. Euphytica 155, 15–26. http://dx.doi.org/10.1007/s10681-006-9296-z.

Idso, S.B., Jackson, R.D., Pinter, P.J., Reginato, R.J., Hatfield, J.L., 1981. Normalizing the stress-degree-day parameter for environmental variability. Agric. Meteorol. 24, 45–55. http://dx.doi.org/10.1016/0002-1571(81)90032-7.

Jones, H.G., Serraj, R., Loveys, B.R., Xiong, L., Wheaton, A., Price, A.H., 2009. Thermal infrared imaging of crop canopies for the remote diagnosis and quantification of plant responses to water stress in the field. Funct. Plant Biol. 36, 978. http://dx.doi.org/10.1071/FP09123.

Kelly, J., Kljun, N., Olsson, P.O., Mihai, L., Liljeblad, B., Weslien, P., Klemedtsson, L., Eklundh, L., 2019. Challenges and best practices for deriving temperature data from an uncalibrated uav thermal infrared camera. Remote Sens. 11, 567. http://dx.doi.org/10.3390/rs11050567.

Lepekhov, S.B., 2022. Canopy temperature depression for droughtand heat stress tolerance in wheat breeding. Vavilov J. Genet. Breed. 26, 196–201. http://dx.doi.org/10.18699/VJGB-22-24.

Malbéteau, Y., Johansen, K., Aragon, B., Al-Mashhawari, S.K., McCabe, M.F., 2021. Overcoming the challenges of thermal infrared orthomosaics using a swath-based approach to correct for dynamic temperature and wind effects. Remote Sens. 13, 3255. http://dx.doi.org/10.3390/rs13163255.

Mesas-Carrascosa, F.J., Pérez-Porras, F., Meroño De Larriva, J., Men. Frau, C., Agüera-Vega, F., Carvajal-Ramírez, F., Martínez-Carricondo, P., García-Ferrer, A., 2018. Drift correction of lightweight microbolometer thermal sensors on-board unmanned aerial vehicles. Remote Sens. 10, 615. http://dx.doi.org/10.3390/rs10040615.

Messina, G., Modica, G., 2020. Applications of UAV thermal imagery in precision agriculture: State of the art and future research outlook. Remote Sens. 12, http://dx.doi.org/10.3390/RS12091491.

Nicodemus, F.E., 1977. Geometrical Considerations and Nomenclature for Reflectance, vol. 160, US Department of Commerce, National Bureau of Standards Washington, DC, USA, http://dx.doi.org/10.6028/NBS.MONO.160.

Nugent, P.W., Shaw, J.A., Pust, N.J., 2013. Correcting for focal-plane-array temperature dependence in microbolometer infrared cameras lacking thermal stabilization. Opt. Eng., Bellingham 52, 061304. http://dx.doi.org/10.1117/1.OE.52.6.061304.

Oakey, H., Verbyla, A., Pitchford, W., Cullis, B., Kuchel, H., 2006. Joint modeling of additive and non-additive genetic line effects in single field trials. Theoret. Appl. Genet. 113, 809–819. http://dx.doi.org/10.1007/s00122-006-0333-z.

Pask, A.J.D., Pietragalla, J., Mullan, D.M., Reynolds, M.P., 2012. Physiological Breeding II: A Field Guide to Wheat Phenotyping. Cimmyt.

Perich, G., Hund, A., Anderegg, J., Roth, L., Boer, M.P., Walter, A., Liebisch, F., Aasen, H., 2020. Assessment of multi-image unmanned aerial vehicle based high-throughput field phenotyping of canopy temperature. Front. Plant Sci. 11, 1–17. http://dx.doi.org/10.3389/fpls.2020.00150.

Piepho, H.P., Möhring, J., Schulz-Streeck, T., Ogutu, J.O., 2012. A stage-wise approach for the analysis of multi-environment trials: Stage-wise analysis of trials. Biom. J. 54, 844–860. http://dx.doi.org/10.1002/bimj.201100219.

Piepho, H.P., Williams, E.R., 2010. Linear variance models for plant breeding trials. Plant Breed. 129, 1–8. http://dx.doi.org/10.1111/j.1439-0523.2009.01654.x.

Prashar, A., Jones, H., 2014. Infra-red thermography as a high-throughput tool for field phenotyping. Agronomy 4, 397–417. http://dx.doi.org/10.3390/agronomy4030397.

QGIS Development Team, 2022. QGIS geographic information system. URL: https://www.qgis.org.

R Development Core Team, 2022. R: A language and environment for statistical computing. URL: http://www.r-project.org. place: Vienna, Austria.

Reynolds, M.P., Pask, A.J.D., Mullan, D.M., 2012. Physiological Breeding I: Interdisciplinary Approaches to Improve Crop Adaptation. CIMMYT.

Ribeiro-Gomes, K., Hernández-López, D., Ortega, J., Ballesteros, R., Poblete, T., Moreno, M., 2017. Uncooled thermal camera calibration and optimization of the photogrammetry process for UAV applications in agriculture. Sensors 17, 2173. http://dx.doi.org/10.3390/s17102173.

Rodríguez-Álvarez, M.X., Boer, M.P., van Eeuwijk, F.A., Eilers, P.H.C., 2018. Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spatial Stat. 23, 52–71. http://dx.doi.org/10.1016/j.spasta.2017.10.003.

Romano, G., Zia, S., Spreer, W., Sanchez, C., Cairns, J., Araus, J.L., Müller, J., 2011. Use of thermography for high throughput phenotyping of tropical maize adaptation in water stress. Comput. Electron. Agric. 79, 67–74. http://dx.doi.org/10.1016/j.compag.2011.08.011.

Roth, L., Aasen, H., Walter, A., Liebisch, F., 2018. Extracting leaf area index using viewing geometry effects–A new perspective on high-resolution unmanned aerial system photography. ISPRS J. Photogramm. Remote Sens. 141, 161–175. http://dx.doi.org/10.1016/j.isprsjprs.2018.04.012.

Roth, L., Camenzind, M., Aasen, H., Kronenberg, L., Barendregt, C., Camp, K.H., Walter, A., Kirchgessner, N., Hund, A., 2020. Repeated multiview imaging for estimating seedling tiller counts of wheat genotypes using drones. Plant Phenomics 2020, http://dx.doi.org/10.34133/2020/3729715, 2020/3729715.

Roth, L., Rodríguez-Álvarez, M.X., Eeuwijk, F.van., Piepho, H.P., Hund, A., 2021. Phenomics data processing: A plot-level model for repeated measurements to extract the timing of key stages and quantities at defined time points. Field Crops Res. 274, 108314. http://dx.doi.org/10.1016/j.fcr.2021.108314.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Stoica, P., Selen, Y., 2004. Model-order selection. IEEE Signal Process. Mag. 21, 36–47. http://dx.doi.org/10.1109/MSP.2004.1311138.

Swiss Federal Council, 2013. Verordnung über Die Direktzahlungen an Die Landwirtschaft (Direktzahlungsverordnung, Dzv). Technical Report, Federal Council of Switzerland.

van Rossum, Guido, Drake, Fred L., 2009. Python 3 Reference Manual. Scotts Valley, CA, Place.

Velazco, J.G., Rodríguez-Álvarez, M.X., Boer, M.P., Jordan, D.R., Eilers, P.H.C., Malosetti, M., Va. Eeuwijk, F.A., 2017. Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. Theoret. Appl. Genet. 130, 1375–1392. http://dx.doi.org/10.1007/s00122-017-2894-4.

Wang, Z., Zhou, J., Ma, J., Wang, Y., Liu, S., Ding, L., Tang, W., Pakezhamu, N., Meng, L., 2023. Removing temperature drift and temporal variation in thermal infrared images of a UAV uncooled thermal infrared imager. ISPRS J. Photogramm. Remote Sens. 203, 392–411. http://dx.doi.org/10.1016/j.isprsjprs.2023.08.011.

Wu, L., 2010. Mixed Effects Models for Complex Data. In: Monographs on Statistics and Applied Probability, vol. 113, Chapman & Hall/CRC Press, Boca Raton.

Yuan, W., Hua, W., 2022. A case study of vignetting nonuniformity in UAV-based uncooled thermal cameras. Drones 6, 394. http://dx.doi.org/10.3390/drones6120394.