Field Crops Research 292 (2023) 108824

Contents lists available at ScienceDirect

Field Crops Research

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

Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks

Gregor Perich[a,*], Mehmet Ozgur Turkoglu[b], Lukas Valentin Graf[a,d], Jan Dirk Wegner[b,c], Helge Aasen[a,d], Achim Walter[a], Frank Liebisch[e]

[a] Crop Science, ETH Zurich, Switzerland
[b] EcoVision Lab, Photogrammetry and Remote Sensing, ETH Zurich, Switzerland
[c] Institute for Computational Science, University of Zurich, Switzerland
[d] Earth Observation of Agroecosystems team, Agroecology and Environment, Agroscope, Switzerland
[e] Agroecology and Environment, Agroscope, Switzerland

ARTICLE INFO

Keywords:
Sentinel-2
Crop yield
Spectral indices
Time series analysis
Random forest
Recurrent neural network
Precision farming
Cereals

ABSTRACT

Mapping and predicting crop yield on a large scale is increasingly important for use cases such as policy-making, risk insurance and precision agriculture applications at farm and field scale. The higher spatial resolution of Sentinel-2 compared to Landsat allows for satellite-based crop yield mapping even in relatively small scaled agricultural settings such as found in Switzerland and other central European regions. In this study, five years (2017–2021) of cereal crop yield data from a combine harvester were used to model crop yield within-field, on a spatial scale corresponding to the Sentinel-2 pixel level. Three established methods from literature using (i-ii) spectral indices and (iii) raw satellite reflectance as well as (iv) a recurrent neural network (RNN) were chosen for analysis. Although the RNN approach did not outperform the other methods, it was more efficient because of the comparatively simple end-to-end training of the model, resulting in much less time spent on data cleaning and feature extraction needed for spectral index time series analysis. The RNN was also able to discriminate cloudy data by itself, reaching similar performance levels as if using pre-processed, cloud-free data. Modelling was performed on individual years, all years combined and on unseen years using leave-one-year-out cross-validation. The models performed best when using data from all years (R^2 up to 0.88, relative RMSE up to 10.49 %) and showed poor performance when predicting on unseen data years, especially for years with previously unknown weather patterns. This highlights the importance of yearly model calibration and the need for continuous data collection enabling long time series for future crop yield models.

1. Introduction

Prediction, modelling and mapping of crop yields based on remote sensing holds great potential for a multitude of applications and stakeholders. For farmers, policymakers, crop insurance and non-governmental organisations, it is of great interest to anticipate crop yield (Weiss et al., 2020). The advent of readily available satellite data has made crop yield mapping and prediction feasible on large scales, either globally (Fritz et al., 2019; Atzberger, 2013) or nation-wide as was shown for the United States (Lobell et al., 2015) and Australia (Kamir et al., 2020). In most cases, the Landsat family of satellites has been used to achieve these tasks (Deines et al., 2021; Kamir et al., 2020; Lopresti et al., 2015; Battude et al., 2016; Jain et al., 2016; Beck et al., 2006; Kang and Özdogan, 2019), however, the Sentinel-2 (S2) satellites see increasing use (Hunt et al., 2019; Skakun et al., 2019). Their higher spatial resolution should enable the mapping of relatively small-scaled agricultural systems such as found in Switzerland, southern Germany and other central European regions. With an increasing interest in a more diverse agriculture and small-scale applications of precision agriculture, such data is gaining in importance. Pixel-based remote sensing is better suited for precision agriculture (i.e., the farmer) as opposed to the field- and regional level which is more relevant for policymaking. Crop insurances are interested in both field- and regional-scale settings (Weiss et al., 2020). Most current satellite-based crop yield mapping systems involve the use of a spectral index (SI) time series such as the Normalised Difference Vegetation Index (NDVI) (Kamir et al., 2020;

[*] Corresponding author.
E-mail address: gregor.perich@usys.ethz.ch (G. Perich).

https://doi.org/10.1016/j.fcr.2023.108824
Received 18 July 2022; Received in revised form 16 November 2022; Accepted 11 January 2023
Available online 18 January 2023

Beck et al., 2006; Battude et al., 2016), the Green Chlorophyll Vegetation Index (GCVI) (Deines et al., 2021; Lobell et al., 2015) or the Enhanced Vegetation Index (EVI) (Kang and Özdogan, 2019). An alternative approach to the use of SIs, is the use of reflectance values recorded by the satellite sensors directly (Hunt et al., 2019).

Time series of optical satellites such as Landsat and S2 are inherently challenged by the presence of clouds. Various strategies to remove cloud-induced artifacts from image time series have been employed, often involving time series interpolation methods or the use of composite images. Composite satellite images integrating information from multiple satellite scenes offer cloud-free time series data (Kamir et al., 2020; Stumpf et al., 2020) but are not always feasible in the context of monitoring rapidly growing agricultural crops as they integrate sensor data over a time span. For time series analysis on field crops using satellite data, interpolation of missing values is therefore more common. Popular interpolation techniques are the fitting of smoothing functions such as 'double logistic (Battude et al., 2016; Beck et al., 2006; Kamir et al., 2020),' 'Fourier series' (Deines et al., 2021), '4253H twice smoother' (Kang and Özdogan, 2019) and splines (Cai et al., 2017; Hermance et al., 2007). Such interpolation techniques may, however, not always be applicable. In central Europe, clouds occur very frequently in particular during the winter months, leading to strongly reduced data availability from October to March. Coupled with an orbit design having higher revisit times in high-latitude regions than near the Equator (Claverie et al., 2018), optical satellite data availability is not always sufficient to extract a long enough time series for the application of a smoothing function. Roy and Yan (2020) suggest that up to 15–20 temporal observations are needed, whereas Deines et al. (2021) estimate eight observations to be sufficient. After smoothing the time series (either SI or direct reflectance), features are then extracted for a machine learning (ML) regression algorithm. Many different ML algorithms have been used for this, with Random forest regression (RFR) being one of the more popular methods (Kamir et al., 2020; Hunt et al., 2019). In their paper, Kamir et al. (2020) showed that approximately half of the tested ML algorithms had very similar performance to one another. Especially when there's much data to train on, the performance of even simple reference ML models reaches that of complex crop models (Deines et al., 2021). With increasing data availability and therefore size, data 'intensive' methods such as neural networks (NNs) are becoming more applicable. Especially Recurrent Neural Networks (RNN) have established themselves as powerful tools for modelling sequential data. They have led to significant progress for a variety of applications, notably language processing and speech recognition (Sutskever et al., 2014; Graves et al., 2013; Vinyals and Le, 2015). Recently, they also achieved state-of-the-art performance for remote sensing time series tasks such as crop classification (Rußwurm and Körner, 2017; Rußwurm and Körner, 2018; Metzger et al., 2021; Turkoglu et al., 2021a,b). They have, however, only sporadically been applied to the estimation of crop yield models, as the community has so far largely focused on SI methods. In remote sensing, ground reference data (often referred to as ground 'truth' data) availability is usually low (Weiss et al., 2020). We therefore release the data set and the code along with the paper to increase data availability for the community and help foster future method development in the field of modelling crop yield.

In this study, we aim to model and predict the crop yield of small grain cereals, including winter wheat, on the S2 pixel level using high-resolution S2 time series data. We focus on the comparison of different models for their performance and applicability to the task in a relatively small-scaled agricultural setting. Three already published models were selected along with a fourth model, which is an adaption of an also published RNN model. The selected models are: (i-ii) two models based on SIs, (iii) one model based on S2 reflectance values and iv) the adapted RNN model. We compare the models' performance across three scenarios: (i) on individual data years, (ii) on all data years combined and iii) across data years to assess the capability of the models to estimate crop yield in general (scenarios i-ii), as well as to predict it on unseen years (scenario iii).

2. Data

2.1. Yield data

Combine harvester data was obtained from a large farm in western Switzerland (46°59′15.157//N7°03′31.814//E, WGS84) with predominant soil type Gleysol. It contains yield data from different field crops for the years 2017–2021. For this study, the data was filtered for the cereal crops winter wheat, winter barley and triticale resulting in a cereals (CR) data set with 54 fields and a winter wheat (WW) sub set with 19 fields. These winter cereals are all managed very similarly in the Swiss agriculture, as the management practices adhere to the Swiss 'Proof of Ecological Performance' (Bundesamt für Landwirtschaft, 2022), which prescribes management details to be eligible to receive Swiss agricultural direct payments. All cereals were sown in autumn between the end of September and beginning of November of the preceding year; the harvest was between mid July and end of August and they were rainfed and fertilised three times. The average field size was 12.78 ha in the CR and 13.11 ha in the WW data set. The mean grain yield per field of the CR data set ranged from 4.88 t/ha in 2021 to 8.65 t/ha in 2020 (Table 1 and appendix table A.1 for more details). Mean grain yield for the WW data set ranged from 4.52 t/ha in 2021 to 8.04 t/ha in 2020. 2021 is considered an exceptional year for farming in Switzerland with precipitation largely above normal and many hail and frost events in the early cropping season leading to yield losses.

The data contains individual, georeferenced measurement points of the combine harvester, taken every two seconds at an average speed of 4 km/h, resulting in a data point every 2.3 m. The topography of the study area is negligible, with the whole area exhibiting less than 2 m height difference across all fields. The raw combine harvester data points were pre-processed as follows (see Fig. 1): In the 1st step, yield values below 0.1 t were filtered out. Then, outliers exceeding three standard deviations (e.g. 3 σ) from the global median were filtered out. The swath width of the combine harvester was 7.2 m, therefore, values below 7 m were omitted to avoid overlapping paths. To remove artifacts caused by speed, values farther than 3 σ from the median speed were filtered out. The last applied step was a 'rolling window' filter, where the three nearest neighbours of each yield point were assessed and values farther than 3 σ from the local median were omitted. The filtered yield data was buffered inwards by 20 m from the field border and then rasterised to a 10 m raster using the 'geocube' package in python with linear interpolation. This resulted in a total of 54'098 pixels with yield information for the CR data set and 20,170 pixels with yield information for the WW subset (see also Table 1 for per-year pixel numbers). From this rasterised yield data, a CSV file containing each pixels' coordinates and yield (in t/ha) was extracted for subsequent modelling (see Fig. 1 for a schematic overview of the data pre-processing). Comparison against the farmers' harvested yields for each field (e.g., the field calendar) showed a systematic over-estimation of yield in the combine harvester data (Supplementary Fig. A.8) that was corrected for by a scaling factor of 0.87. The yield data published along with this study corresponds to the outcome of the pre-processing described in this section.

Table 1
Mean yield over all fields in tonnes per hectare (t/ha) and the number of yield pixels per year for the cereals (CR) and the winter wheat (WW) data set. The WW data set is a sub set of the CR data set.

year	CR yield [t/ha]	WW yield [t/ha]	CR px	WW px
2017	7.35	5.53	10,536	1638
2018	6.87	6.29	13,108	6240
2019	7.36	7.08	13,696	4340
2020	8.65	8.04	12,727	4969
2021	4.88	4.52	4031	2983
total	7.02	6.29	54,098	20,170

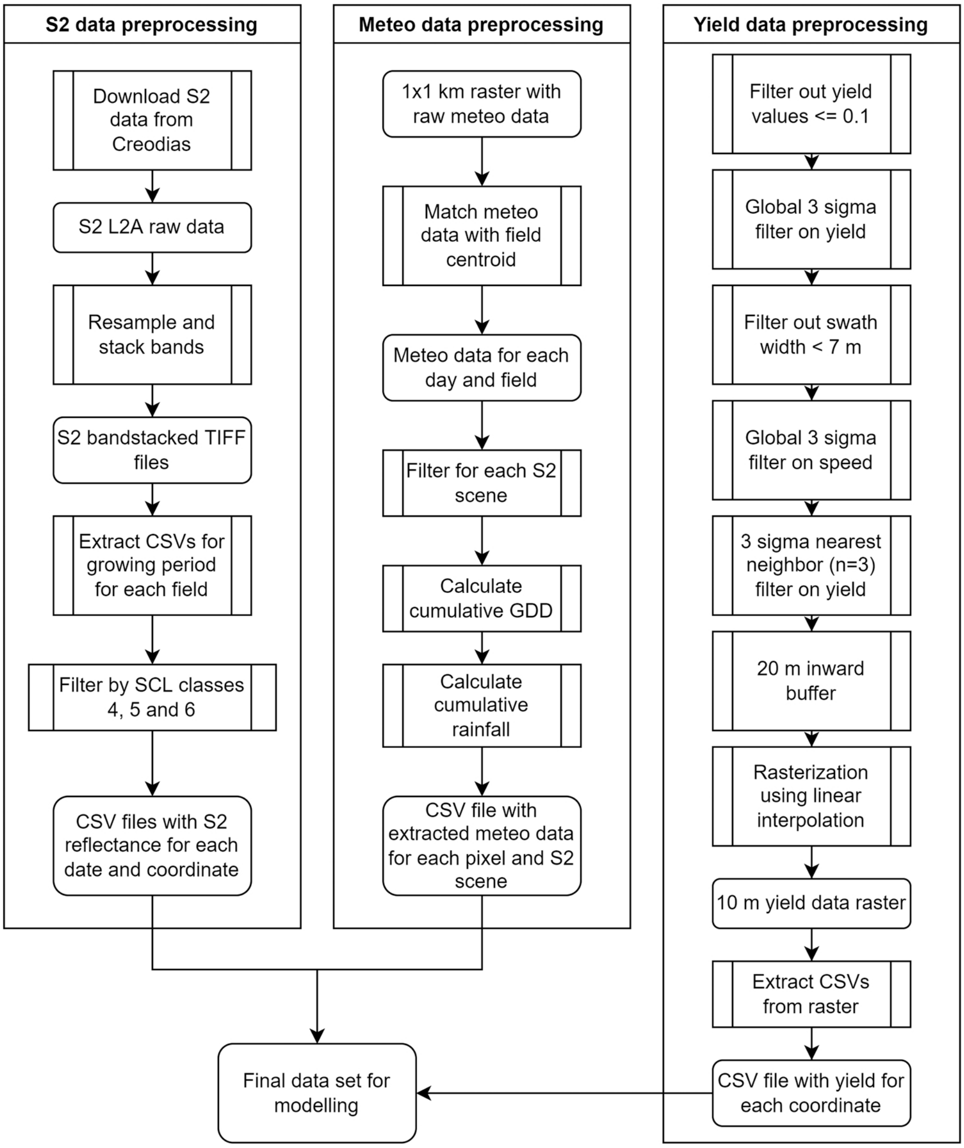
S2 data preprocessing
Download S2 data from Creodias
S2 L2A raw data
Resample and stack bands
S2 bandstacked TIFF files
Extract CSVs for growing period for each field
Filter by SCL classes 4, 5 and 6
CSV files with S2 reflectance for each date and coordinate
Meteo data preprocessing
1x1 km raster with raw meteo data
Match meteo data with field centroid
Meteo data for each day and field
Filter for each S2 scene
Calculate cumulative GDD
Calculate cumulative rainfall
CSV file with extracted meteo data for each pixel and S2 scene
Yield data preprocessing
Filter out yield values <= 0.1
Global 3 sigma filter on yield
Filter out swath width < 7 m
Global 3 sigma filter on speed
3 sigma nearest neighbor (n=3) filter on yield
20 m inward buffer
Rasterization using linear interpolation
10 m yield data raster
Extract CSVs from raster
CSV file with yield for each coordinate
Final data set for modelling

Fig. 1. Schematic of the data pre-processing performed on the different data sets.

2.2. Sentinel-2 data

All available S2 scenes containing the study region from January 2017 to December 2021 were downloaded from the Data and Information Access Service (DIAS) platform 'Creodias' and pre-processed as shown in Fig. 1. S2 scenes prior to March 2018 were only available in the top-of-atmosphere L1C format and were downloaded as such. Later scenes were downloaded in the bottom-of-atmosphere L2A format. The L1C scenes were processed to L2A product level using the 'Sen2Cor' processor version 2.9 provided by ESA. The S2 bands at 60 m resolution (bands 1, 9 and 10) were omitted from the analysis, as they lie within spectral regions affected by atmospheric disturbances caused by aerosols and water vapour (Spoto et al., 2012). For each S2 scene, the 20 m bands were resampled to 10 m using 'cubic' interpolation settings and all bands were written to a 10-band stacked TIFF file. The Scene Classification Layer (SCL) provided by ESA is a classification of each pixel in a S2 scene and is provided in 20 m resolution. It was resampled by dividing each 20 m pixel into four 10 m pixels, effectively using a 'nearest neighbour interpolation'. The resampled SCL layer was also added to the bandstack. From these bandstacks, data was extracted using the SCL layer to filter out clouds, cloud shadows, dark areas and defective pixels. Only pixels from the SCL classes 4 (vegetation), 5 (non-vegetated) and 6 (water) were kept. This resulted in a CSV file containing – for each pixel – the S2 reflectance data for each cloud-free S2 scene (e.g., date) in the growth period from sowing to harvest of the selected field. Clouds are a common occurrence in the data set, as the study region lies between three lakes. Figure 2 illustrates how the fraction of cloud-free pixels of the study region is low for the autumn and winter months October to February (0–120 Days After Sowing (DAS)) and only stabilises late in the growing season in April (around 175 DAS). After S2 data pre-processing, data from 134 S2 scenes were kept for analysis. Appendix table A.2 gives an overview of the available S2 scenes for each field used in this study. The S2 data processing routines were performed using an early version of the open source 'Earth Observation Data Analysis Library' - 'EOdal' (https://doi.org/10.5281/zenodo.7278252) available under GNU General Public version 3 license (Graf et al., 2022).

2.3. Meteorological data

Daily temperature and rainfall data was obtained from the Swiss Federal Office of Meteorology and Climatology 'MeteoSwiss'. Both weather variables are available in 1x1 km gridded tiles in the Swiss national coordinate system CH1903/LV03 (EPSG:21781). For each field

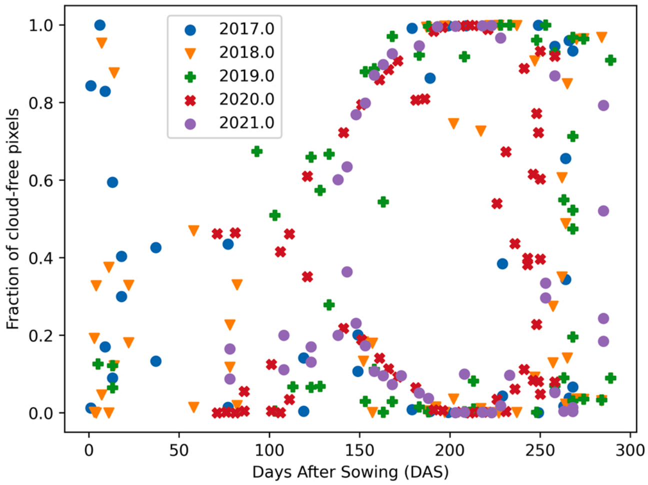

Fig. 2. Fraction of cloud-free pixels of the study region from every Sentinel-2 scene of all five data years.

polygon, the centroid was calculated and data from the nearest weather data coordinate was assigned as this field polygons' weather data. The daily rainfall data was summed up, resulting in cumulative rainfall for each S2 scene.

3. Methods

For this study, we compared four models for their advantages and potential disadvantages to learn about optimal applicability for crop yield modelling and prediction on the S2 pixel level. The focus was to assess method robustness, accuracy and data processing needs. The methods were chosen from literature based on their heterogeneity, e.g. as to include diverse time series analysis techniques, variables (e.g. using SI's vs. spectral bands directly) and ML algorithms. The two methods based on SIs, the 'partial integral at peak GCVI' (Section 3.1) and 'smoothed NDVI' (Section 3.3) were chosen due to the prevalance of SI methods in the remote sensing literature (see introduction). The 'four S2 scenes' method (Section 3.2) was chosen as it is a robust, yet high-performing method using all available spectral bands of S2 as input variables. The used RNN (Section 3.4) was chosen as there are very few examples of pixel-based crop yield modelling using RNNs. Table 2 gives a brief overview of the methods, while Fig. 3 gives a graphical illustration thereof.

3.1. Partial integral at peak GCVI

The first implemented method originates from Deines et al. (2021), which use the partial integral of the smoothed GCVI time series curve between the peak GCVI and thirty days after the peak in addition to the weather variables rainfall and temperature. This method was chosen as

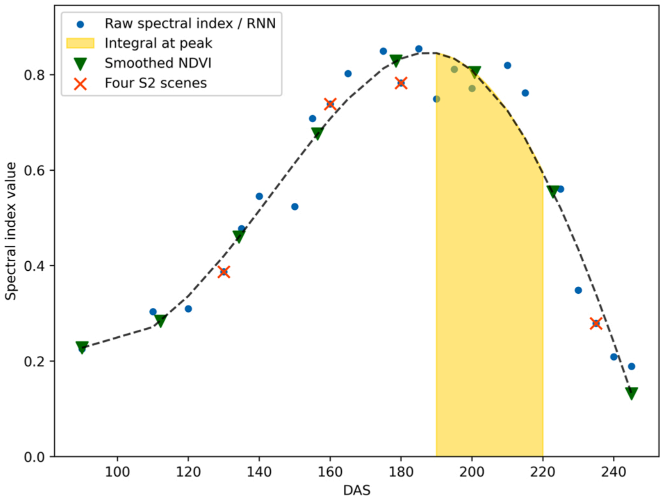

Fig. 3. Spectral index curve of an exemplary pixel, illustrating how the selected methods of this study each use more or less of the available Sentinel-2 (S2) time series. The 'integral at peak' method uses the least information from the time series (yellow area). The 'smoothed NDVI' method uses eight equidistantly sampled values along the NDVI curve (green triangles). The 'Four S2 scenes' as well as the Recurrent Neural Network (RNN) do not use the spectral index, but rather use the ten available S2 bands. The former uses information from four time points (orange crosses), whereas the latter uses the full time series (blue dots). For a detailed description see section 3.

it is a simple, easy to understand method which, in theory, allows for in-season prediction as it only takes a small window (Deines et al., 2021) call it the 'key crop growth window') of the S2 time series into account. In their paper, Deines et al. (2021) primarily describe their Scalable Crop Yield Mapper (SCYM) approach, but also benchmark against a RFR. The latter was taken as a method for this study. The GCVI time series was calculated from the pre-processed S2 data (Fig. 1) and was smoothed using a 2nd order Fourier series smoothing (FSS) as follows:

$$f(t) = c + \sum_{k=1}^{n=2} a_k cos(2\pi\omega kt) + b_k sin(2\pi\omega kt) \quad (1)$$

where $f(t)$ is the fitted SI value at time point t, ω is the frequency, a_k and b_k are the cosine and sine coefficients and c the intercept coefficient calculated as the mean SI value over the whole SI time series. The unit of each time point t was given in days after sowing (DAS). From this, the peak GCVI was calculated and the partial integral between the DAS of the peak GCVI and + 30 days was defined and the area under the curve (AUC) thereof was calculated as described in Deines et al. (2021). In addition to the AUC, the maximum temperature in June (peak growth for cereals) and the total April to June rainfall were taken as variables for the RFR model. The hyperparameters of the RFR model were chosen based on the values reported in Deines et al. (2021) as follows: A RF tree size of 200, 2 variables per split, a minimum leaf size of 2, and a bag fraction of 0.63.

Table 2
Overview table of the methods used in this study and their respective variables.

Method	Smoothing	Time series length	Selected variables	Weather variables	Total variables	Regression algorithm	References
Partial integral at peak	Fourier series	Integral of 30 days after peak GCVI	Area under the GCVI Curve (AUC)	Max June temp. + tot. Apr. to June rainfall	3	RFR	Deines et al., 2021
Four S2 scenes	–	4 selected S2 scenes	Raw S2 reflectance of 10 bands	Monthly avg. Temp. + total monthly rainfall	8	RFR	Hunt et al., 2019
Smoothed NDVI	B-Splines	8 selected NDVI observations	Smoothed NDVI values + extracted features (e.g. slope, min/max NDVI, AUC, etc.)	Avg. daily temp. + cumulative rainfall	31	RFR	Kamir et al., 2020; Battude et al., 2016
Recurrent Neural Network	–	Full available S2 time series	Raw S2 reflectance of 10 bands	Avg. daily temp. + cumulative rainfall	12	RNN	Turkoglu et al., 2021a,b

3.2. Four S2 scenes

The method proposed by Hunt et al. (2019) takes four arbitrarily selected, cloud-free S2 scenes across the growth period of winter wheat, extracts the reflectance of the ten S2 bands (Section 2.2) and uses it together with monthly aggregates of temperature and rainfall in a RFR. This method was chosen as it is a straight-forward method using all (except the 60 m) bands of S2, has shown good performance and is easy to implement. Furthermore, Hunt et al. (2019) have been using very similar point-based yield data from combine harvesters in their study, which makes their method attractive for evaluation on this study's data set. The original selection of S2 scenes (December, April, June and July) performed by Hunt et al. (2019) could not be exactly matched because no cloud-free S2 scenes were available at these dates for the study region. Instead, four S2 scenes were selected for the whole AOI according to the availability of fully cloud-free S2 scenes. The 10 available S2 bands (Section 2.2) were kept as features for the RFR model. In addition, the monthly average of the daily mean temperature and the total monthly precipitation were calculated for each of the four selected S2 scenes. RFR model parameters were set in accordance with Hunt et al. (2019) to: RF tree size of 500, 1/3 of variables used to split the data at each node and 10-fold cross-validation.

3.3. Smoothed NDVI

This method is a synthesis of methods found in literature where features from a smoothed NDVI time series were extracted for a ML algorithm (Kamir et al., 2020; Battude et al., 2016; Beck et al., 2006). It was chosen to be representative of a NDVI-based model, which is arguably one of the most widespread SIs used for crop and plant phenology models in the remote sensing community. Here, the NDVI time series of the cloud-free S2 observations was smoothed using B-Splines and a RFR was selected as the ML algorithm. In addition to the cloud filtration using the SCL (Fig. 1), S2 scenes on which an individual field had less than 90% cloud free pixels were omitted. This resulted in more dense per-pixel NDVI time series, which still included visible outliers in the form of clouds. Therefore, an additional cloud filter adapted from the MAJA cloud detection algorithm (Hagolle et al., 2017; Bolton et al., 2020) was used, which masked pixels as cloudy if:

$$(\rho_{blue}(D) - \rho_{blue}(D_{prev})) > 0.03 * (1 + (D - D_{prev})/30) \tag{2}$$

where $\rho_{blue}(D)$ is the reflectance of the blue band on date D, D_{prev} is the date previous to D and the difference between D and D_{prev} is given in days. Since a sudden variation in ρ_{blue} can also occur due to agricultural management practices or natural variations such as fires or snow (Hagolle et al., 2017), a 2nd check was implemented. Additional to equation (2), a pixel was not masked if the following condition was fulfilled:

$$(\rho_{red}(D) - \rho_{red}(D_{prev})) > 1.5 * (\rho_{blue}(D) - \rho_{blue}(D_{prev})) \tag{3}$$

where ρ_{red} is the reflectance of the red band on date D and $\rho_{red}()D_{prev}$ the reflectance of the red band on the date (D_{prev}, the date of the S2 observation previous to D. The resulting, cloud-free NDVI time series were smoothed using B-Splines with four knots and a polynomial degree of three.

For this model, Growing degree days (GDDs), a measure for temperature normalised plant growth, were used as the value of the time axis. GDDs are agronomically more relevant than days after sowing (DAS), as they include information on the phenological growth stage of the plant. GDDs were calculated using daily maximum and minimum temperatures according to McMaster and Wilhelm (1997) as follows:

$$GDD = \frac{T_{max} + T_{min}}{2} - T_{base} \tag{4}$$

where if $(T_{max} + T_{min})/2 < T_{base}$, then $GDD = T_{base}$. The base temperature T_{base} denotes the temperature below which the crop does not grow and was set to 0°C in this study (McMaster and Wilhelm, 1997). T_{max} and T_{min} are the daily maximum and minimum temperatures. GDDs were calculated for each day and summed up to obtain a cumulative GDD number for each S2 scene.

The following features were extracted from the smoothed NDVI time series: max & min NDVI, timepoint of max & min NDVI, area under the (NDVI) curve (AUC), min & max GDD. In addition to these features, eight NDVI observations as well as the cumulative rainfall and the average daily temperature at these NDVI observations were taken as features, totalling in 31 variables/features. These features were input into a Random forest regression (RFR) model. Opposed to the previous methods, where the ML models' hyperparameters were taken from their respective source papers, the RF parameters of this method were selected based on cross-validation: The RF tree size was 500, 16 variables were used for each split and the bag fraction was 0.632.

3.4. Recurrent neural network

An RNN was used as the fourth method in this study because they belong to the family of neural network approaches for time-series analysis, which is receiving increasing attention for crop analysis using remote sensing data (Rußwurm and Körner, 2018; Turkoglu et al., 2021a). The RNN architecture, originally developed for crop type classification in (Turkoglu et al., 2021a), was adapted to the regression task of yield modelling. A 2-layers Gated Recurrent Unit (GRU) architecture (Cho et al., 2014; Chung et al., 2014) with 256 hidden units was used with dropout rate set to 0.5, and learned with the 'Adam' optimiser (Adam et al., 2010) with learning rate set to 0.001. Training was done for 60 epochs with a batch size of 8, taking the equivalent of 15 min per GPU (NVIDIA RTX 2080Ti) and data year.

Contrary to the other methods, which use sub sets differing in length from the available time series (Fig. 3), the RNN was run with the full available time series for each field pixel as input data. E.g. all ten S2 bands as well as the cumulative rainfall and the average daily temperature at each available S2 scene (Section 2.2). In addition, the RNN was run with two other input data sets: i) with the raw, cloudy data (i.e., not filtered using the SCL layer of ESA) to assess the ability of the network to discriminate clouds from true S2 scene content on the ground. ii) with the same data used in the four S2 scenes method (Section 3.2), i.e., cloud-free but reduced time series length to enable better model comparison.

3.5. Model scenarios and implementation details

All models were run on three different scenarios. In the 1st scenario ('per-year'), each model was trained on an individual data year and evaluated on the test set from the same year. For each data year and individual field, 80 % of the yield pixels were taken as training, and 20 % as the test set. In the 2nd scenario ('all-year'), the training and test set data from all five years was aggregated and then used for modelling. The training and test sets split was kept the same for all methods for these two scenarios. The 3rd scenario ('cross-year') consisted of a leave-one-out cross-validation: Each of the five years was held out once, while the models were trained on the four remaining years of data. The holdout year was then predicted on to assess model performance. The performance metrics chosen for model evaluation were the coefficient of determination (R^2) and the Root Mean Square Error (RMSE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

where $\bar{y}$ is the average observed yield of all samples. Additionally, the relative RMSE (rRMSE, in %) was calculated from the yearly average cereal yields (Table 1) as follows:

$$rRMSE = \frac{RMSE}{\bar{y}} * 100 \quad (7)$$

4. Results

4.1. Per-year performance

The performance of the per-year models can be seen in Table 3. For the CR data set, the integral at peak method showed the least performance. The other methods showed similar per-year performance, with R^2 values being within 0.10 and rRMSE values being within ≈ 5 % of each other. The four S2 scenes method showed the best performance by a small margin over both the RNN and the smoothed NDVI method. The models exhibited the worst performance on the year 2017 and the best for the year 2020. For the smaller WW sub set, the integral at peak method also showed the least performance with the other methods exhibiting similar performance. The four S2 scenes method again showed the best performance on the WW sub set, albeit by a small margin. The RNN exhibited slightly better performance than the smoothed NDVI method, with the exception on the year 2017. The best model predictions were on the year 2021. Performance on the WW sub set was, like in the CR data set, worst for the year 2017. Compared to the other three methods, the integral at peak method did not exhibit a performance loss for the year 2017 when going from the CR to the WW data set. Generally, the models exhibited slightly lower coefficients of variation (R^2 values) on the smaller WW sub set than on the CR data set. The rRMSE values were generally similar between the WW and CR data sets.

4.2. Neural network with cloudy data

The results of running the RNN on the original pre-processed data (Section 2.2) as well as the two additional data sets (Section 3.4) are shown in Table 4. Overall, the RNN performs similar across all three input S2 time series. The cloudy time series exhibited the best performance among the three tested time series by a few percentage points for both R^2 and rRMSE. This held true for both the CR and the WW data set, with the performance increase of using the cloudy time series data being larger for the WW sub set.

Table 3
Performance metrics of the per-year models for both the cereals (CR) and the winter wheat (WW) data set. The per-year models were trained and evaluated on each data year individually. The best score for each metric is **bold**, the 2nd best underlined and the 3rd best in *italics*.

		R^2		RMSE [t/ha]		rRMSE [%]	
Method	Data set Year	CR	WW	CR	WW	CR	WW
Integral at peak	2017	0.34	0.33	1.60	0.69	21.77	12.48
	2018	0.39	0.39	1.19	1.15	17.32	18.28
	2019	0.42	0.20	1.36	1.17	18.48	16.53
	2020	0.36	0.36	1.34	1.19	15.49	14.80
	2021	0.71	0.77	1.18	1.04	24.18	23.01
Four S2 scenes	2017	0.80	0.55	0.89	0.56	12.11	10.13
	2018	0.83	0.80	0.64	0.66	9.32	10.49
	2019	0.79	0.73	0.81	0.67	11.01	9.46
	2020	*0.85*	0.82	*0.65*	*0.63*	7.51	**7.84**
	2021	**0.87**	**0.89**	0.79	0.73	16.19	16.15
Smoothed NDVI	2017	0.73	0.50	1.03	0.59	13.95	10.73
	2018	0.77	0.74	0.72	0.75	10.55	11.98
	2019	0.73	0.63	0.93	0.80	12.61	11.25
	2020	0.82	0.77	0.71	0.72	*8.17*	8.93
	2021	0.82	*0.86*	0.92	0.80	18.82	17.76
Recurrent Neural Network	2017	0.77	0.42	0.95	0.64	12.93	11.57
	2018	0.81	0.77	0.67	0.71	9.75	11.29
	2019	0.77	0.67	0.85	0.75	11.55	10.59
	2020	0.86	0.75	**0.63**	0.75	**7.28**	*9.33*
	2021	0.83	0.87	0.90	0.79	18.44	17.48

Table 4
Per-year performance of the Recurrent Neural Network on different sets of Sentinel-2 (S2) input data. The 'regular time series' contains the S2 data which was pre-processed as described in Section 2.2. The '4 cloud-free scenes' correspond to the same data as was used in the 'four S2 scenes' method (Section 3.2). The 'cloudy time series' contains all S2 scenes over the study region, regardless of the scene classification layer (SCL) cloud mask. The best score for each metric is bold, the 2nd best underlined and the 3rd best in *italics*.

		R^2		RMSE [t/ha]		rRMSE [%]	
Input data	Data set Year	CR	WW	CR	WW	CR	WW
regular time series	2017	0.77	0.42	0.95	0.64	12.93	11.57
	2018	0.81	0.77	0.67	0.71	9.75	11.29
	2019	0.77	0.67	0.85	0.75	11.55	10.59
	2020	**0.86**	0.75	**0.63**	0.75	**7.28**	*9.33*
	2021	*0.83*	**0.87**	0.90	0.79	18.44	17.48
4 cloud-free scenes	2017	0.75	0.41	0.98	*0.65*	13.33	11.75
	2018	0.77	0.75	0.73	0.74	10.63	11.76
	2019	0.74	0.65	0.92	0.77	12.50	10.88
	2020	0.82	0.75	0.70	0.74	*8.09*	9.20
	2021	0.81	0.86	0.96	0.83	19.67	18.36
cloudy time series	2017	0.79	0.52	0.91	**0.58**	12.38	10.49
	2018	0.82	0.80	*0.66*	0.67	9.61	10.65
	2019	0.78	0.70	0.84	0.72	11.41	10.17
	2020	0.85	0.78	0.64	0.70	7.40	**8.71**
	2021	0.83	**0.87**	0.91	0.77	18.65	17.04

4.3. All-year performance

The model performance using training and test data from all five years is shown in Table 5. For the CR data set, the integral at peak method showed the lowest performance. The other three methods showed similar R^2 values ranging between 0.82 and 0.86. The four S2 scenes and the RNN showed the lowest RMSE values at 0.76 t/ha (rRMSE = 10.82 %) and 0.79 t/ha (rRMSE = 11.25 %), respectively. All methods exhibited better model performance on the WW sub set, which was also observed in the tighter grouping of the scatterplots (Fig. 4). The largest increase in model performance was observed in the integral at peak method. The other methods performed slightly better on the WW sub set than on the CR data set with R^2 values again being very comparable (0.84–0.88). RMSE values were all below 1 t/ha with the four S2 scenes method exhibiting the lowest RMSE at 0.66 t/ha (rRMSE = 10.49%).

Figure 5 shows an exemplary yield map for a winter barley field using the predictions of the all-year models and the corresponding prediction errors as the predicted yield minus the true yield. All models showed no clear pattern in the distribution of prediction errors. This seemingly random distribution of errors (both from yield over- and underestimations) indicates, that there is no systematic, spatial error

Table 5
Model prediction results using training and test set data from all five years combined (e.g. 'all-year' scenario).

	R^2		RMSE [t/ha]		rRMSE [%]	
Data set Method	CR	WW	CR	WW	CR	WW
Integral at peak	0.55	0.65	1.35	1.12	19.23	17.80
Four S2 scenes	0.86	0.88	0.76	0.66	10.82	10.49
Smoothed NDVI	0.82	0.85	0.86	0.75	12.25	11.92
Recurrent Neural Network	0.85	0.86	0.79	0.70	11.25	11.13

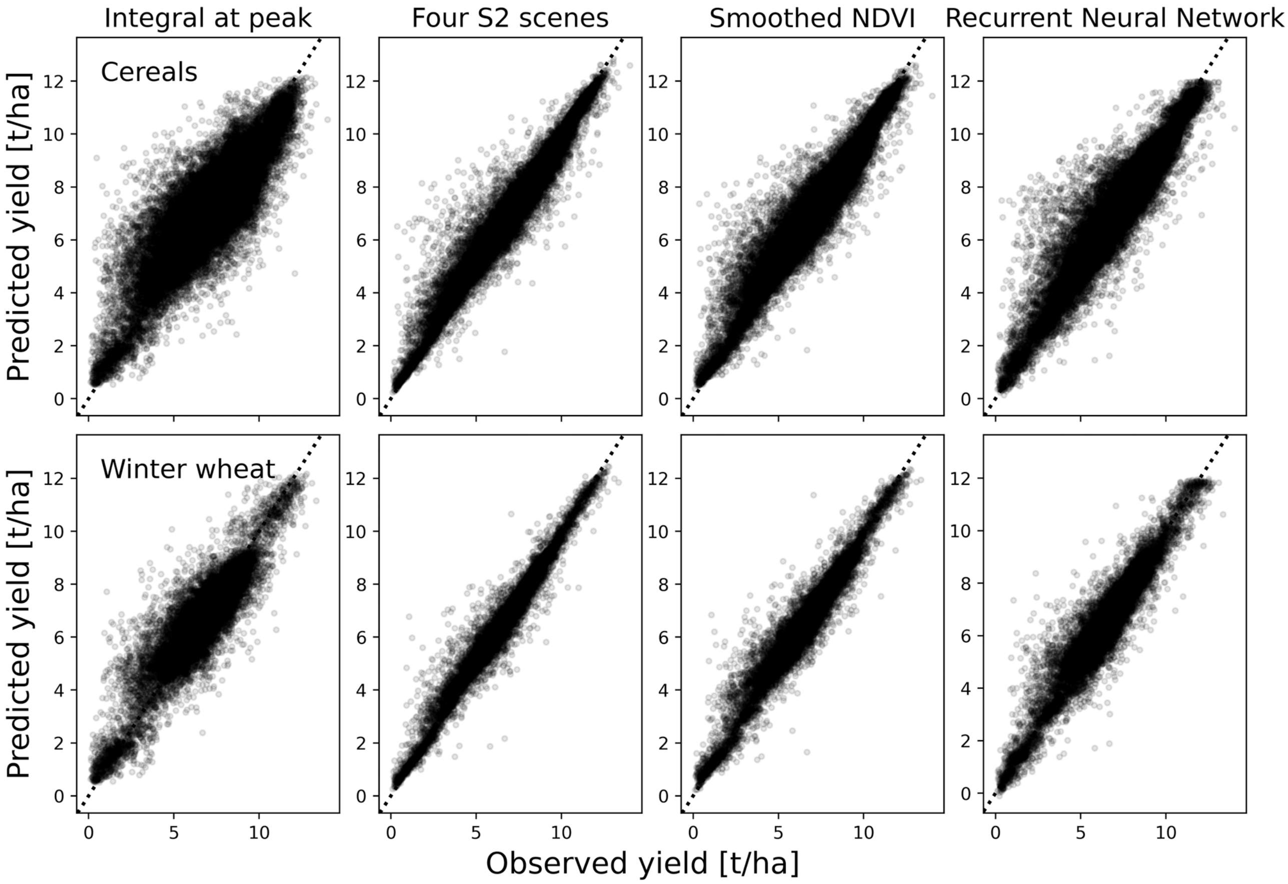

Fig. 4. Scatterplot of the relationship between the observed and predicted yields of the full cereal (top) and the winter wheat (bottom) data sets. Predictions were obtained from the all-year scenario. The dashed line has slope one. Detailed model fit metrics are given in Table 5.

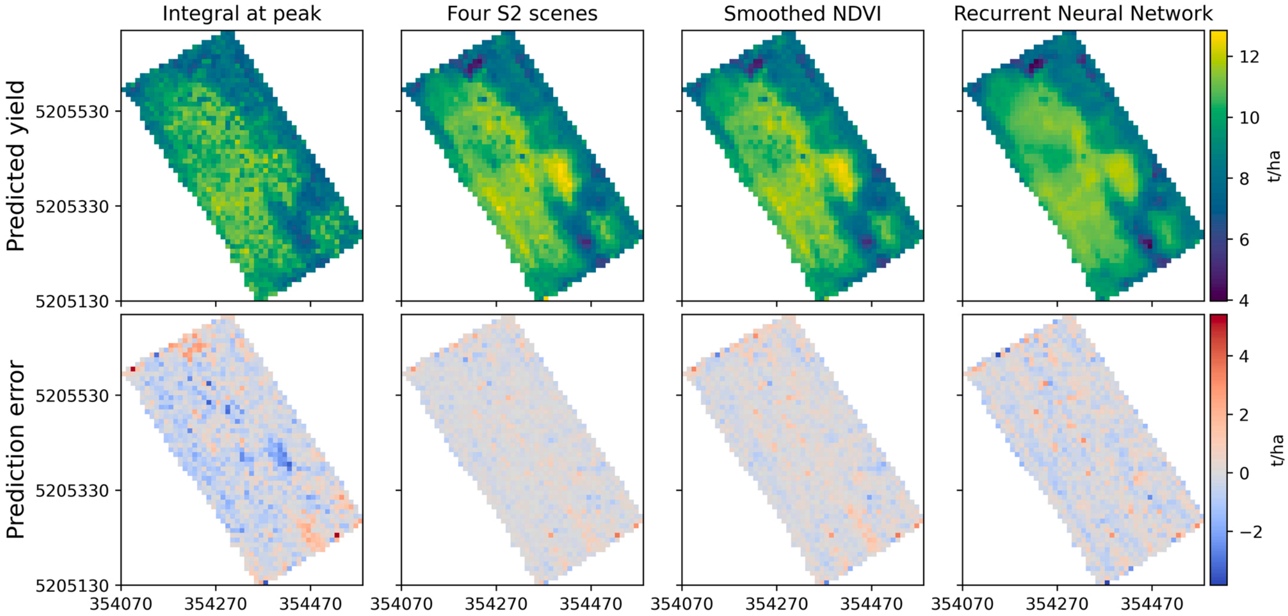

Fig. 5. Yield map of field ID 36 (winter barley) for the year 2020 showing the predicted yield and the prediction error as the predicted yield minus true yield. The predictions were obtained from the all-year scenario. Coordinates are given in metres (EPSG:32632).

present in these models. The predicted yield map of the integral at peak method exhibited the most jagged image, owing to this models' comparatively lower performance (Fig. 6). The yield map obtained from the RNN exhibited the most homogeneous map out of all methods (Fig. 5).

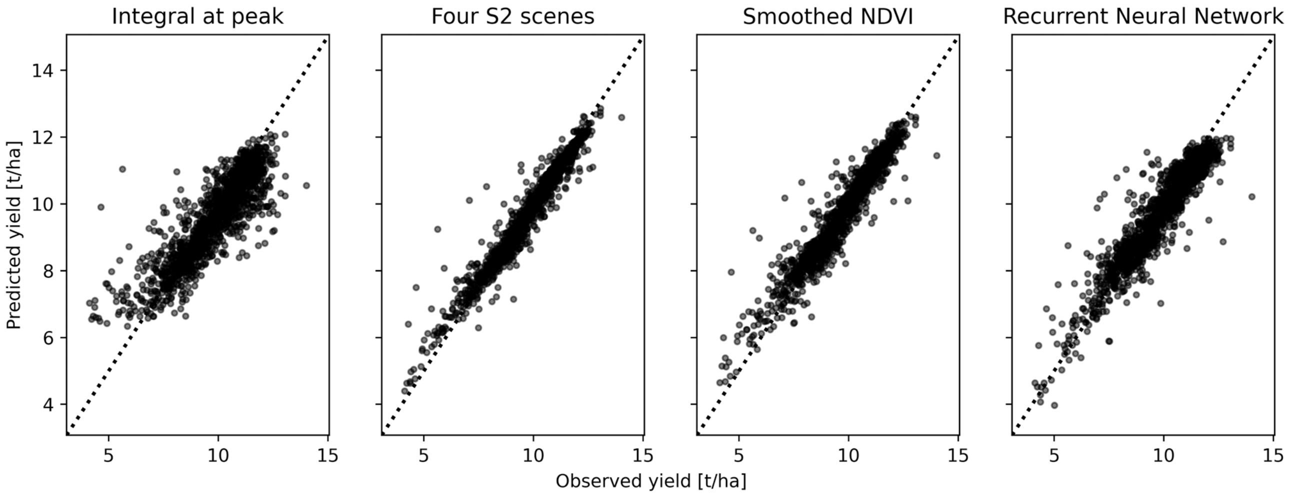

Fig. 6. Scatterplot of the relationship between the observed and predicted yields of field ID 36 for all four methods. The predictions were obtained from the all-year scenario. The dashed line has slope one.

4.4. *Cross-year performance*

The models trained on four data years were not able to correctly predict crop yield of the unseen fifth data year. Negative R^2 values were obtained for all models for the WW sub set and for most of the CR data set (Appendix table A.3). Only the four S2 scenes and the RNN methods showed positive, but low R^2 values (ranging from 0.17 to 0.42) for the holdout years 2017–2019 using the CR data set. For these years, these models also exhibit the best-in-class rRMSE values ranging from 16.74 % to 23.67 % for the four S2 scenes method and from 20.38 % to 24.35 % for the RNN. The used formula for R^2 (eq. (5)) can only be negative, if the numerator is larger than the denominator, which is the case if the residuals $\widehat{y_i}$ are farther away from the mean of the predicted data $\bar{y}$. This prevalence of negative R^2 values thus indicates, that the models' predictions were far from the mean of the test data (e.g., the unseen data year).

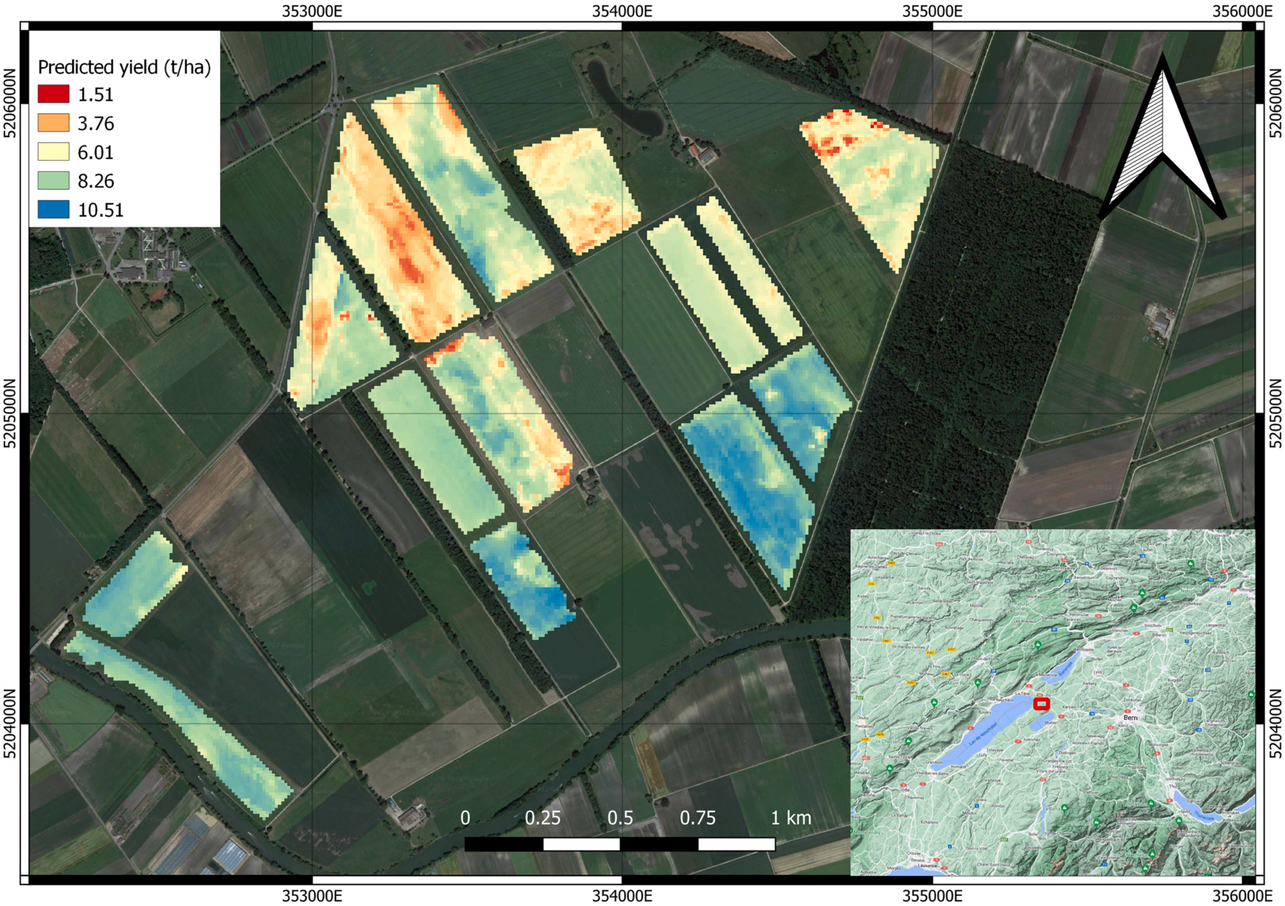

Fig. 7. Yield map of the cereal fields from study region for the year 2019 obtained from the prediction of the recurrent neural network trained on data from the all-year scenario. Background maps by Google ('Satellite' and 'Terrain' layers).

5. Discussion

The results show that precise yield modelling - and to a lesser degree also prediction - is possible in a relatively small-scaled agricultural setting using high-resolution S2 imagery (Fig. 7). This shows that satellite-based approaches become usable on smaller scales, and not just on regional scales, bringing the technology closer to producers. This enables tools to allow farmers to better assess within-field yield potentials and make informed management decisions (e.g., fertiliser input, pest control, etc.) more efficient, increasing the sustainability (Finger et al., 2019; Walter et al., 2017) and reducing the risk for their enterprises (Kantanantha et al., 2010). Policymakers can use yield maps to asses yield gaps (Lobell, 2013) and to increase reaction time related to food security (Kogan et al., 2019). Also, yield maps enable data-based crop insurance products (Kogan, 2019) and agricultural commodity trading (Filippi et al., 2019), while making the latter less prone to price volatility (Fritz et al., 2019).

5.1. Data set

The range of data used in published studies varies greatly, but usually includes a few thousand pixels: Kamir et al. (2020) used 3011 yield pixels (at 250 m resolution of MODIS), Hunt et al. (2019) used yield data from 8794 pixels at 10 m S2 resolution, whereas Deines et al. (2021) obtained a massive data set of over one million field-level yield maps at 5 m resolution (resampled to 30 m resolution of Landsat) through a collaboration with a large Ag-tech company. The data set used in this study is therefore medium-to-large, depending if only one year of data (e.g., year 2021 had 4031 pixels, year 2019 had 13,696) is used or all years combined (54,098 pixels). The data amounts were sufficient for modelling as could be seen in model performance of the per-year and the all-year scenarios (Tables 3 and 5). The model performance did, however, drop in the years with little yield pixels, especially for the WW sub set (Table 1). The available data set is unique for Switzerland, as combine harvester data is hard to obtain due to the sensible nature of this data.

The observed cereal yields in the data set are representative of the Swiss agriculture, where average cereal yields are between 6 and 7 t/ha, depending on the quality level (Bundesamt für Landwirtschaft, 2021). Combine harvester yields are eventually slightly overestimated, as combine harvesters often report slightly lower yield values at the field borders compared to the field centre (Khosla and Flynn, 2008) and these regions were omitted due to the buffer of 20 m that was deducted when rasterising the yield data (Section 2.1). This effect was also observed in the data set when comparing the combine yield data with the per-field yield data reported by the farmer (see appendix figure A.1) and had to be corrected for in this study. This highlights the need to calibrate combine harvester data (Kharel et al., 2019) but also offers a simple calibration procedure if field-scale yield data is available.

While the study region is a typical agricultural cropping area of western Switzerland, it might differ from other regions due to local factors such as the soil composition and the climate. For broad application of remote estimation of crop yield, validation studies in other regions are therefore needed for a country-wide yield prediction.

5.2. Modelling crop yield

The integral at peak method showed overall lower performance than the other methods used in this study. This is not surprising, as it only uses three variables compared to the other methods, which use up to 31 variables (smoothed NDVI, Section 3.3). It also uses a very limited ‘window’ of the available time series (30 days), which makes it unable to detect sharp changes in either the early or late growing season. Nevertheless, We obtained slightly better performance in this study as in the original work of Deines et al. (2021), who report 0.40 R^2 for their ML baseline which is run on the equivalent of the all-year scenario in this study (Section 3.1). Their RMSE value of ca. 2.45 t/ha is not directly comparable, as it is reported for corn and not cereals. Using the average corn yield of the USA (for 2018: 12.8 t/ha, USDA NASS data), we calculate a rRMSE of 19.14 % which is comparable to our model output of the CR data set for both the per-year (Table 3) and the all-year scenarios (Table 5). Our crop-specific WW model showed lower rRMSE values, but - especially for the per-year scenario – lower R^2 values. For the all-year scenario, our WW model exhibited better performance ($R^2 = 0.65$, rRMSE = 17.8 %).

Hunt et al. (2019) obtained good model performance ($R^2 = 0.91$, RMSE = 0.61 t/ha, rRMSE ≈ 6.1 %) for pixel-based yield modelling of winter wheat, which our implementation of the method was not able to reproduce in the per-year scenario (Table 3). In the all-year scenario, our four S2 scenes method showed similar R^2 (0.88) and RMSE (0.66 t/ha) values, but slightly worse rRMSE values of 10.5 %. The slightly better performance reported in Hunt et al. (2019) might be due to the fact, that they were able to use a broader range of cloud-free S2 time series incorporating December, April, June and July. Our implementation of the method only got cloud-free S2 imagery from April to harvest in July across all data years used in this study. Using a scene from December can be disputed, as the WW has usually only just emerged and should be barely discernible from the soil signal in a S2 image. However, it is also a good way to ‘force’ the inclusion of a soil or a barely vegetated pixel signal, providing the regression algorithm with the data from the start of the vegetation curve (also see Fig. 3. A S2 scene from December would mean a DAS of around 30–60).

The smoothed NDVI method cannot be compared as directly as the previous two methods, as it is a synthesis of methods from literature (Section 3.3). The method described by Kamir et al. (2020) is the most directly comparable, as they worked on pixel-based (250 m resolution of MODIS) modelling of winter wheat yield maps obtained from farmers. They report R^2 values of 0.77 and RMSE of 0.55 t/ha (rRMSE of 30.72 % with the avg. wheat yield of 1.79 t/ha in their data used). Their model is also run on the equivalent of the all-year scenario of this paper. When comparing our smoothed NDVI method, we obtain a better model fit ($R^2 = 0.84$, rRMSE = 14.14 %) for WW. The low resolution of MODIS possibly explains part of the worse performance. The method of Battude et al. (2016) is less directly comparable, as they performed a point-based field sampling campaign on maize (instead of using yield maps) and used the Simple Algorithm for Yield Estimates (SAFY) (Duchemin et al., 2008) in conjunction with the satellite (Landsat, SPOT, and more) data. They did, however, perform similar smoothing of the satellite time-series data. For their yield estimation at the field-level, they report a correlation (not R^2!) of 0.81 and a rRMSE of 8.82 %. Skakun et al. (2019) also used a similar method as in this paper by smoothing a SI curve obtained from the Harmonised Landsat S2 (HLS) product. They did, however, not predict on the pixel-level, but rather on the regional level. They report R^2 values of 0.73 and a rRMSE of 5.4 %.

There are considerably less papers applying RNNs (or other types of NNs) on satellite data time series to estimate crop yield. The few that we found, usually model crop yield at very coarse, regional levels such as the US county-level using the national agricultural statistics service (NASS) data set (Terliksiz and Altýlar, 2019; Ghazaryan et al., 2020). Others performed similar, regional-scale analysis in Brazil (Schwalbert et al., 2020) or Australia (Cai et al., 2019). Only Khan et al. (Khan et al., 2020) estimated crop biomass of mint on field-level in India for one data year and report a R^2 of 0.76 and a RMSE of 2.74 t/ha (rRMSE = 15.58 %). This small amount of NN approaches to model and predict crop yield on the field- or pixel-scale exemplifies the lack of adequate data sets where the NNs can be trained and validated on.

A brief analysis of the all-year models’ variable importances (Appendix Figs. A.3 for the CR data set and A.4 for the WW data set) showed that the weather variables were generally ranked with high importance scores across all models. Especially the smoothed NDVI method ranked the temperature and the cumulative rainfall from both the early and late growth stages as the most influential variables, with two instances of

NDVI in the early growth stages being also highly ranked. This is in contrast to Kamir et al. (2020), who reported NDVI variables ranking highest. The four S2 scenes' monthly aggregated weather variables, especially the cumulative rainfall, were highly ranked. The RNN was an exception, as it ranked the S2 bands B05 (Red edge 1) and B03 (Green) before the two weather variables. Interestingly, the Near-infra-red bands (B08 and B8A), who are often associated with vegetation growth (Thenkabail et al., 2013; Gnyp et al., 2014), were not highly ranked at all in the RNN and only of medium importance in the four S2 scenes method.

Overall, the per-year and all-year models performed well. This shows that the selected methods can be used to model crop yield on the pixel-level. This enables their use for farm and field management information for farmers (that have large enough fields to make use of the 10 m pixel resolution of S2), agricultural statistics for either crop insurance schemes or policymakers to better make decisions regarding food safety.

5.2.1. *Using spectral indices and multiple spectral bands*

The yield modelling methods used in this paper can be classified into two types: i) using SIs (integral at peak and smoothed NDVI) and ii) using S2 reflectance data from multiple bands (four S2 scenes and RNN). The latter methods tend to perform better (Tables 3 and 5) as they use more of the available spectral information. However, the increase in performance is not so large to consider them superior to the SI-based methods in every regard. The main advantage of the methods using raw reflectance is their reduced need for data pre-processing. This decreases the time needed for implementation and operation. The main drawback of these methods is that they are more complex to analyse and therefore, it is often more difficult to communicate the results to stakeholders. The main drawback of the two methods based on SIs is the requirement to pre-process the satellite time series data in order to be applicable. Such a pre-processing requires a minimum amount of data points in the time series (Roy and Yan, 2020; Deines et al., 2021). This minimum amount was not always given in our time series due to the heavily cloud-contaminated winter and early spring months (Fig. 2). Thus, it was not possible to implement a reliable double logistic smoothing, which is very often used for NDVI time series (Kamir et al., 2020; Battude et al., 2016; Beck et al., 2006), for the smoothed NDVI method. As some clouds within a S2 scene would only partially cover an individual field, the time series length of the different pixels within that field would vary greatly. Especially in such cases, the Fourier series smoothing performed in the integral at peak method introduced artifacts at both ends of the time series due to the sinusoidal form of the Fourier series. This could also be observed in the poor average fitting parameters of the Fourier series smoothing (for cereals: $R^2 = 0.44$, RMSE = 2.34, for WW: $R^2 = 0.56$, RMSE = 1.5, GCVI values ranged from 0.75 to 12.2 with a median of 3.5). This did not matter for this study, as the integral at peak method, which used Fourier series smoothing, focused on the peak region, where the fit was much improved. For the smoothed NDVI method, however, such artifacts did seriously impact model performance as was shown in prior testing using Fourier series smoothing. Additionally, errors may be introduced when smoothing a satellite time series using a technique which usually expects an equidistant distribution of data points in time. So far, there's little discussion in literature on such errors. In this work, B-Splines were used for the smoothed NDVI method, as they were able to smooth the NDVI using less data points than the double logistic smoothing and handled the NDVI time series ends without introducing obvious artifacts. Fitting parameters of the B-Splines smoothing were much improved over the Fourier series smoothing (for cereals: $R^2 = 0.99$, RMSE = 0.02, and for WW: $R^2 = 0.98$, RMSE = 0.03, NDVI values ranged from 0.22 to 0.94 with a median of 0.81). As just discussed, interpolation is very time consuming and requires prior knowledge of the behaviour of the SI curve to select an adequate smoothing function. Another promising approach for gap-filling the cloud-contaminated S2 time series would be the incorporation of Sentinel-1 Synthetic Aperture Radar (SAR) data, which has the ability to penetrate clouds. This would however, require the complex pre-processing of SAR data and the sensor fusion with the optical S2 data. Despite the drawbacks of interpolation, using SI-based methods have a large benefit: They are easy to interpret and thus to communicate to stakeholders.

5.2.2. *Neural networks*

The RNN models' performance was similar to that of the four S2 scenes method, which was the best performing model found in this study. In addition to the advantages and disadvantages discussed in the previous section, the RNN was able to perform well using satellite input data of varying data quantity and quality (Table 4). It worked well with only very short time series such as the four cloud-free scenes as used in the four S2 scenes method as well as with the full, unprocessed S2 time series including clouds. This meant that the data cleaning and pre-processing could be reduced to a minimum as there was no need for time-consuming feature extraction (Sagan et al., 2021), which greatly decreased the time required for the method implementation and operation. The WW subset for 2017, which only included 1638 yield pixels (Table 1), even saw an increase in RNN model performance when the longer, cloudy S2 time series data was used compared to the pre-processed series. RNNs have been shown to be able to deal with cloudy time series for crop classification tasks (Metzger et al., 2021), but to the best of our knowledge, this is the first time a RNN has been used to model crop yield using cloudy satellite time series. Additionally, a homogenisation effect on the output crop yield map (Fig. 5) could be observed. This is very realistic, as the within-field yield is usually distributed with a smooth gradient (with the exception of externally-induced damages such as hail, floods and damping-off of crop seedlings). We conclude that RNNs cannot only be used to model crop yield on regional scale, but are also a very promising method for high-resolution, pixel-based crop yield modelling.

5.3. *Predicting crop yield*

In this study, we investigated two ways of yield prediction. The first is the integral at peak method that can be regarded as a within-season forecasting of crop yield as it takes information from the peak GCVI curve usually observed at the end of April to mid-May for cereals in the study region. The second is the prediction of the held-out, unseen data years, which is a more stringently separated cross-year prediction method.

The integral at peak method showed good performance, especially in the all-year scenario, where - for WW - 65 % of the yield variation could be explained with a rRMSE of 17.8 % (Table 5). Since the integral at peak method was developed for corn (Deines et al., 2021), it could likely be further optimised for small grain cereals, improving the crop yield prediction. Possible approaches to improve the method would be to better parameterise the integral width, which is now arbitrarily selected to be 30 days after the peak or even to expand the integral to include the earlier growth period of the crop and the inclusion of a more refined set of weather variables.

Predicting held-out, unseen data years showed poor model performance (Section 4.4). This highlighted the limitation of empirical, data-driven ML models to their training data. Unseen patterns in the data, originating either from large yearly differences in crop yield (Table 1) or from weather events, lead to poor model performance. Especially for the years 2020 and 2021, the average crop yields were much higher (2020) or lower, respectively (2021, see also Appendix Table A.1 for more details). This explains the prevalence of negative R^2 values especially for these two years. The cold and hail events in the early and heavy rainfall in the peak growing season of 2021 (Appendix Fig. A.2) also negatively impacted the model performance. If the years with more 'regular' weather patterns were predicted, the models' performance was better (Appendix Table A.3). Five years of data and corresponding weather data is not enough to predict crop yield of unseen data years or to extrapolate. It would be interesting to see, if an empirical ML would be able to reliably predict held-out years if trained on a massive data set as was used in the work of Deines et al. (2021). Therefore, to improve empirical ML models, more data needs to be collected to better cover a wider variety of both target (yield) and input variables (weather). An improved composition of weather variables could also increase prediction performance. With the exception of storms, floods or other extreme events, weather usually impacts crops in a delayed manner. Therefore, the use of a 'rolling window' for weather variables might be better suited to capture the influence of the past *n* days of weather. This may be an explanation for the better performance of the four S2 scenes method in the cross-year scenario compared to the RNN (appendix table A.3), since it uses monthly aggregated weather variables. The inclusion of extreme weather events themselves should be considered as well, as they can have a massive impact on crop growth. However, completely unknown patterns are still impossible to learn for an empirical model. Despite the failure to predict the mean crop yield of an unseen data year (equation (5)), a mean rRMSE of 37.5 % for predicting unseen cereal yields was achieved for the CR data set (Appendix table A.3). This equates to a model RMSE of approximately 2.7 t/ha using the mean yield observed (Table 1) which nonetheless allows for a rough crop yield estimation.

To overcome these limitations of empirical ML models, data assimilation techniques (Jin et al., 2018) or mechanistic crop models (Weiss et al., 2020) could be used. Such models can be updated within-season as new satellite and weather observations become available. This would also reduce the dependency on timely satellite (S2 or other) observations, as the development of the crops could be forecast without access to S2 data. Such approaches have already been performed (Battude et al., 2016; Azzari et al., 2017; Kang and Özdogan, 2019) and show good performance. Therefore, we think that the combination of empirical and mechanistic approaches holds great promise to predict crop yield.

6. Conclusion

This study has shown that modelling crop yield is possible within-field on the 10 m resolution of S2 in a relatively small-scaled agricultural setting by using either methods based on SIs or methods using the S2 reflectance directly. The latter methods perform slightly better, but are less straightforward to interpret. A downside of SI methods is the need to interpolate the cloud-contaminated S2 time series, which is a time consuming process that requires high user knowledge. The implemented RNN approach was able to discriminate between cloudy and non-cloudy pixels by itself, eliminating the need for any time series preprocessing and indicating the high potential of RNNs for crop yield modelling and prediction. The main drawback of using a RNN is the large amount of training data needed, which explains the low prevalence of RNNs in literature for such applications, as pixel-based yield data is hard to obtain. The in-season prediction of crop yield is possible, albeit at reduced performance (R^2 up to 0.65, rRMSE up to 17.8 %) compared to analyzing the full time series from sowing until harvest (R^2 up to 0.88, rRMSE up to 10.49 %). The empirical methods used in this study exhibited low performance when predicting the crop yield of unseen data years due to the lack of information on unseen data patterns caused mostly by the weather. An alternative would be the use of mechanistic models to model the crop growth and the accumulation of biomass for crop yield along the crop development to make the prediction more robust to unseen data points. In general, more data, e.g. more data years and different locations to enable better model calibration and validation is paramount for further model development to enable robust yield mapping and timely yield forecasting for a multitude of uses and stakeholders in the agricultural system.

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 link to the data and code used in the study is in the manuscript.

Acknowledgements

GP received funding through the project 'DeepField' of the Swiss Federal Office for Agriculture. LVG and HA acknowledge funding of the Swiss Science Foundation for the project 'PhenomEn' (grant number IZCOZ0_198091). We thank Alfred Burri and Bernhard Streit for giving us access to the combine harvester data and for general support in questions related to the farm. We acknowledge data access from the Swiss Federal Office of Meteorology and Climatology 'MeteoSwiss' and are thankful to Prof. Peter Molnar for his support with the weather data.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fcr.2023.108824.

References

Adam, E., Mutanga, O., Rugege, D., 2010. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. Wetl. Ecol. Manag. 18, 281–296. https://doi.org/10.1007/s11273-009-9169-z.

Atzberger, C., 2013. Advances in remote sensing of agriculture: context description, existing operational monitoring systems and major information needs. Remote Sens. 5, 949–981. https://doi.org/10.3390/rs5020949. ⟨https://www.mdpi.com/2072-4292/5/2/949⟩. number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Azzari, G., Jain, M., Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: testing multiple methods and satellites in three countries. Remote Sens. Environ. 202, 129–141. https://doi.org/10.1016/j.rse.2017.04.014. ⟨https://www.sciencedirect.com/science/article/pii/S0034425717301645⟩.

Battude, M., AlBitar, A., Morin, D., Cros, J., Huc, M., MaraisSicre, C., LeDantec, V., Demarez, V., 2016. Estimating maize biomass and yield over large areas using high spatial and temporal resolution Sentinel-2 like remote sensing data. Remote Sens. Environ. 184, 668–681. https://doi.org/10.1016/j.rse.2016.07.030. ⟨https://www.sciencedirect.com/science/article/pii/S0034425716302875⟩.

Beck, P.S.A., Atzberger, C., Høgda, K.A., Johansen, B., Skidmore, A.K., 2006. Improved monitoring of vegetation dynamics at very high latitudes: a new method using MODIS NDVI. Remote Sens. Environ. 100, 321–334. https://doi.org/10.1016/j.rse.2005.10.021. ⟨https://www.sciencedirect.com/science/article/pii/S0034425705003640⟩.

Bolton, D.K., Gray, J.M., Melaas, E.K., Moon, M., Eklundh, L., Friedl, M.A., 2020. Continental-scale land surface phenology from harmonized Landsat 8 and Sentinel-2 imagery. Remote Sens. Environ. 240, 111685 https://doi.org/10.1016/j.rse.2020.111685. ⟨https://www.sciencedirect.com/science/article/pii/S0034425720300547⟩.

Bundesamt für Landwirtschaft, 2021.Agrarbericht 2021. Technical Report. Bundesamt für Landwirtschaft (BLW).⟨https://www.agrarbericht.ch/de/markt/pflanzliche-produkte/getreide?_k=5I544Kzs⟩.

Bundesamt für Landwirtschaft, 2022.Ökologischer Leistungsnachweis.⟨https://www.blw.admin.ch/blw/de/home/instrumente/direktzahlungen/oekologischer-leistungsnachweis.html⟩.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. For. Meteorol. 274, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010. ⟨https://www.sciencedirect.com/science/article/pii/S0168192319301224⟩.

Cai, Z., Jönsson, P., Jin, H., Eklundh, L., 2017. Performance of smoothing methods for reconstructing NDVI time-series and estimating vegetation phenology from MODIS data. Remote Sens. 9, 1271. https://doi.org/10.3390/rs9121271. ⟨https://www.mdpi.com/2072-4292/9/12/1271⟩. number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Cho, K., van Merrienboer, B., Bahdanau, D., Bengio, Y., 2014.On the properties of neural machine translation: encoder-decoder approaches.⟨http://arxiv.org/abs/1409.1259⟩, 10.48550/arXiv.1409.1259.number: arXiv:1409.1259 arXiv:1409.1259 [cs, stat].

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.⟨http://arxiv.org/abs/1412.3555⟩, 10.48550/arXiv.1412.3555.number: arXiv:1412.3555 arXiv:1412.3555 [cs].

Claverie, M., Ju, J., Masek, J.G., Dungan, J.L., Vermote, E.F., Roger, J.C., Skakun, S.V., Justice, C., 2018. The harmonized landsat and sentinel-2 surface reflectance data set.

Remote Sens. Environ. 219, 145–161. https://doi.org/10.1016/j.rse.2018.09.002. ⟨https://linkinghub.elsevier.com/retrieve/pii/S0034425718304139⟩.

Deines, J.M., Patel, R., Liang, S.Z., Dado, W., Lobell, D.B., 2021. A million kernels of truth: insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. Remote Sens. Environ. 253, 112174 https://doi.org/10.1016/j.rse.2020.112174. ⟨https://www.sciencedirect.com/science/article/pii/S0034425720305472⟩.

Duchemin, B., Maisongrande, P., Boulet, G., Benhadj, I., 2008. A simple algorithm for yield estimates: evaluation for semi-arid irrigated winter wheat monitored with green leaf area index. Environ. Model. Softw. 23, 876–892. https://doi.org/10.1016/j.envsoft.2007.10.003. ⟨https://www.sciencedirect.com/science/article/pii/S1364815207002010⟩.

Filippi, P., Jones, E.J., Wimalathunge, N.S., Somarathna, P.D.S.N., Pozza, L.E., Ugbaje, S. U., Jephcott, T.G., Paterson, S.E., Whelan, B.M., Bishop, T.F.A., 2019. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precis. Agric. 20, 1015–1029. https://doi.org/10.1007/s11119-018-09628-4.

Finger, R., Swinton, S.M., El Benni, N., Walter, A., 2019. Precision farming at the nexus of agricultural production and the environment. Annu. Rev. Resour. Econ. 11, 313–335. https://doi.org/10.1146/annurev-resource-100518-093929. ⟨https://www.annualreviews.org/doi/10.1146/annurev-resource-100518-093929⟩.

Fritz, S., See, L., Bayas, J.C.L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., Rembold, F., Rojas, O., Schucknecht, A., Van der Velde, M., Verdin, J., Wu, B., Yan, N., You, L., Gilliams, S., Mücher, S., Tetrault, R., Moorthy, I., McCallum, I., 2019. A comparison of global agricultural monitoring systems and current gaps. Agric. Syst. 168, 258–272. https://doi.org/10.1016/j.agsy.2018.05.010. ⟨https://www.sciencedirect.com/science/article/pii/S0308521X17312027⟩.

Ghazaryan, G., Skakun, S., König, S., Rezaei, E.E., Siebert, S., Dubovyk, O., 2020 Crop yield estimation using multi-source satellite image series and deep learning, In: Proceedings of the IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 5163–5166.10.1109/IGARSS39084.2020.9324027.iSSN: 2153–7003.

Gnyp, M.L., Miao, Y., Yuan, F., Ustin, S.L., Yu, K., Yao, Y., Huang, S., Bareth, G., 2014. Hyperspectral canopy sensing of paddy rice aboveground biomass at different growth stages. Field Crops Res. 155, 42–55. https://doi.org/10.1016/j.fcr.2013.09.023. ⟨http://www.sciencedirect.com/science/article/pii/S0378429013003298⟩.

Graf, L.V., Perich, G., Aasen, H., 2022. EOdal: an open-source python package for ecosystem scale agroecological research using earth observation and gridded environmental data. Computers and335Electronics in Agriculture, 203:107487, Dec. 2022. ISSN 0168-1699. doi: 10.1016/j.compag.2022.336107487.

Graves, A., Mohamed, A.r., Hinton, G., 2013.Speech recognition with deep recurrent neural networks, in: ICASSP.

Hagolle, O., Huc, M., Desjardins, C., Auer, S., Richter, R., 2017.MAJA algorithm theoretical basis document ⟨https://zenodo.org/record/1209633⟩, 10.5281/zenodo.1209633.publisher: Zenodo.

Hermance, J.F., Jacob, R.W., Bradley, B.A., Mustard, J.F., 2007. Extracting phenological signals from multiyear AVHRR NDVI time series: framework for applying high-order annual splines with roughness damping. IEEE Trans. Geosci. Remote Sens. 45, 3264–3276. https://doi.org/10.1109/TGRS.2007.903044 (conference Name: IEEE Transactions on Geoscience and Remote Sensing).

Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W., Rowland, C.S., 2019. High resolution wheat yield mapping using Sentinel-2. Remote Sens. Environ. 233, 111410 https://doi.org/10.1016/j.rse.2019.111410. ⟨https://www.sciencedirect.com/science/article/pii/S0034425719304298⟩.

Jain, M., Srivastava, A., Balwinder-Singh, Joon, R., McDonald, A., Royal, K., Lisaius, M., Lobell, D., 2016. Mapping smallholder wheat yields and sowing dates using micro-satellite data. Remote Sens. 8, 860. https://doi.org/10.3390/rs8100860. ⟨http://www.mdpi.com/2072-4292/8/10/860⟩.

Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., Wang, J., 2018. A review of data assimilation of remote sensing and crop models. Eur. J. Agron. 92, 141–152. https://doi.org/10.1016/j.eja.2017.11.002. ⟨https://www.sciencedirect.com/science/article/pii/S1161030117301685⟩.

Kamir, E., Waldner, F., Hochman, Z., 2020. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. ISPRS J. Photogramm. Remote Sens. 160, 124–135. https://doi.org/10.1016/j.isprsjprs.2019.11.008. ⟨https://www.sciencedirect.com/science/article/pii/S092427161930262X⟩.

Kang, Y., Özdogan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163. https://doi.org/10.1016/j.rse.2019.04.005. ⟨https://www.sciencedirect.com/science/article/pii/S0034425719301427⟩.

Kantanantha, N., Serban, N., Griffin, P., 2010. Yield and price forecasting for stochastic crop decision planning. J. Agric., Biol., Environ. Stat. 15, 362–380. https://doi.org/10.1007/s13253-010-0025-7.

Khan, M.S., Semwal, M., Sharma, A., Verma, R.K., 2020. An artificial neural network model for estimating Mentha crop biomass yield using Landsat 8 OLI. Precis. Agric. 21, 18–33. https://doi.org/10.1007/s11119-019-09655-9.

Kharel, T.P., Swink, S.N., Maresma, A., Youngerman, C., Kharel, D., Czymmek, K.J., Ketterings, Q.M., 2019. Yield monitor data cleaning is essential for accurate corn grain and silage yield determination. Agron. J. 111, 509–516. https://doi.org/10.2134/agronj2018.05.0317.

Khosla, R., Flynn, B., 2008. Understanding and cleaning yield monitor data. In: Soil Science Step-by-Step Field Analysis. John Wiley & Sons Ltd, pp. 113–130. https://doi.org/10.2136/2008.soilsciencestepbystep.c9 (pp).

Kogan, F., 2019. Vegetation health for insuring drought-related yield losses and food security enhancement. In: Kogan, F. (Ed.), Remote Sensing for Food Security. Springer International Publishing, Cham, pp. 163–173. https://doi.org/10.1007/978-3-319-96256-6_7 (Sustainable Development Goals Series, pp).

Kogan, F., Guo, W., Yang, W., 2019. Drought and food security prediction from NOAA new generation of operational satellites. Geomat., Nat. Hazards Risk 10, 651–666. https://doi.org/10.1080/19475705.2018.1541257 publisher: Taylor & Francis.

Lobell, D.B., 2013. The use of satellite data for crop yield gap analysis. Field Crops Res. 143, 56–64. https://doi.org/10.1016/j.fcr.2012.08.008. ⟨https://www.sciencedirect.com/science/article/pii/S0378429012002754⟩.

Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sens. Environ. 164, 324–333. https://doi.org/10.1016/j.rse.2015.04.021. ⟨https://www.sciencedirect.com/science/article/pii/S0034425715001637⟩.

Lopresti, M.F., DiBella, C.M., Degioanni, A.J., 2015. Relationship between MODIS-NDVI data and wheat yield: a case study in Northern Buenos Aires province, Argentina. Inf. Process. Agric. 2, 73–84. https://doi.org/10.1016/j.inpa.2015.06.001. ⟨https://linkinghub.elsevier.com/retrieve/pii/S221431731500027X⟩.

McMaster, G.S., Wilhelm, W.W., 1997. Growing degree-days: one equation, two interpretations. Agric. For. Meteorol. 87, 291–300. https://doi.org/10.1016/S0168-1923(97)00027-0. ⟨https://www.sciencedirect.com/science/article/pii/S0168192397000270⟩.

Metzger, N., Turkoglu, M.O., D'Aronco, S., Wegner, J.D., Schindler, K., 2021. Crop classification under varying cloud cover with neural ordinary differential equations. IEEE Trans. Geosci. Remote Sens. 60, 1–12.

Roy, D.P., Yan, L., 2020. Robust Landsat-based crop time series modelling. Remote Sens. Environ. 238, 110810 https://doi.org/10.1016/j.rse.2018.06.038. ⟨https://www.sciencedirect.com/science/article/pii/S003442571830316X⟩.

Rußwurm, M., Körner, M., 2017.Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images, In: Proceedings of the In'l Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops.

Rußwurm, M., Körner, M., 2018. Multi-temporal land cover classification with sequential recurrent encoders. ISPRS Int. J. Geo-Inf. 7, 129.

Sagan, V., Maimaitijiang, M., Bhadra, S., Maimaitiyiming, M., Brown, D.R., Sidike, P., Fritschi, F.B., 2021. Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. ISPRS J. Photogramm. Remote Sens. 174, 265–281. https://doi.org/10.1016/j.isprsjprs.2021.02.008. ⟨https://www.sciencedirect.com/science/article/pii/S0924271621000411⟩.

Schwalbert, R.A., Amado, T., Corassa, G., Pott, L.P., Prasad, P.V.V., Ciampitti, I.A., 2020. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agric. For. Meteorol. 284, 107886 https://doi.org/10.1016/j.agrformet.2019.107886. ⟨https://www.sciencedirect.com/science/article/pii/S0168192319305027⟩.

Skakun, S., Vermote, E., Franch, B., Roger, J.C., Kussul, N., Ju, J., Masek, J., 2019. Winter wheat yield assessment from landsat 8 and sentinel-2 data: incorporating surface reflectance, through phenological fitting, into regression yield models. Remote Sens. 11, 1768. https://doi.org/10.3390/rs11151768 number: 15 Publisher: Multidisciplinary Digital Publishing Institute.

Spoto, F., Sy, O., Laberinti, P., Martimort, P., Fernandez, V., Colin, O., Hoersch, B., Meygret, A., 2012.Overview Of Sentinel-2, In: Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, 1707–1710.10.1109/IGARSS.2012.6351195.iSSN: 2153–7003.

Stumpf, F., Schneider, M.K., Keller, A., Mayr, A., Rentschler, T., Meuli, R.G., Schaepman, M., Liebisch, F., 2020. Spatial monitoring of grassland management using multi-temporal satellite imagery. Ecol. Indic. 113, 106201 https://doi.org/10.1016/j.ecolind.2020.106201. ⟨https://linkinghub.elsevier.com/retrieve/pii/S1470160X20301382⟩.

Sutskever, I., Vinyals, O., Le, Q.V., 2014.Sequence to sequence learning with neural networks, in: Advances in neural information processing systems.

Terliksiz, A.S., Altýlar, D.T., 2019.Use of deep neural networks for crop yield prediction: a case study of soybean yield in Lauderdale County, Alabama, USA, in: 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), pp.1–4.10.1109/Agro-Geoinformatics.2019.8820257.

Thenkabail, P.S., Mariotto, I., Gumma, M.K., Middleton, E.M., Landis, D.R., Huemmrich, K.F., 2013. Selection of hyperspectral narrowbands (HNBs) and composition of hyperspectral twoband vegetation indices (HVIs) for biophysical characterization and discrimination of crop types using field reflectance and hyperion/EO-1 data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 427–439. https://doi.org/10.1109/JSTARS.2013.2252601. ⟨http://ieeexplore.ieee.org/document/6507245/⟩.

Turkoglu, M.O., D'Aronco, S., Perich, G., Liebisch, F., Streit, C., Schindler, K., Wegner, J. D., 2021a. Crop mapping from image time series: deep learning with multi-scale label hierarchies. Remote Sens. Environ. 264, 112603 https://doi.org/10.1016/j.rse.2021.112603. ⟨https://www.sciencedirect.com/science/article/pii/S0034425721003230⟩.

Turkoglu, M.O., D'Aronco, S., Wegner, J., Schindler, K., 2021b. Gating revisited: deep multi-layer rnns that can be trained. IEEE Trans. Pattern Anal. Mach. Intell.

Vinyals, O., Le, Q., 2015.A neural conversational model. arXiv preprint arXiv: 1506.05869.

Walter, A., Finger, R., Huber, R., Buchmann, N., 2017. Opinion: smart farming is key to developing sustainable agriculture. Proc. Natl. Acad. Sci. USA 114, 6148–6150. https://doi.org/10.1073/pnas.1707462114. ⟨https://www.pnas.org/content/114/24/6148⟩ (publisher: National Academy of Sciences Section: Opinion).

Weiss, M., Jacob, F., Duveiller, G., 2020. Remote sensing for agricultural applications: a meta-review. Remote Sens. Environ. 236, 111402 https://doi.org/10.1016/j.rse.2019.111402. ⟨http://www.sciencedirect.com/science/article/pii/S0034425719304213⟩.