ENVIRONMENTAL RESEARCH
LETTERS

PURPOSE-LED PUBLISHING™ Part of the ENVIRONMENTAL RESEARCH SERIES

LETTER • OPEN ACCESS

A weakly supervised framework for high-resolution crop yield forecasts

To cite this article: Dilli Paudel *et al* 2023 *Environ. Res. Lett.* **18** 094062

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

A weakly supervised framework for high-resolution crop yield forecasts

Dilli Paudel[1,*], **Diego Marcos**[2], **Allard de Wit**[3], **Hendrik Boogaard**[3] **and Ioannis N Athanasiadis**[4]

[1] Information Technology Group, Wageningen University & Research, Hollandseweg 1, 6706 KN Wageningen, The Netherlands
[2] Inria, University of Montpellier, Montpellier, France
[3] Wageningen Environmental Research, Wageningen University & Research, P O Box 47, 6700 AA Wageningen, The Netherlands
[4] Laboratory for Geo-information Science and Remote Sensing, and Wageningen Data Competence Center, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB Wageningen, The Netherlands
* Author to whom any correspondence should be addressed.

E-mail: dilli.paudel@wur.nl

Keywords: crop yield, deep learning, weak supervision, disaggregation, spatial variability

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
20 March 2023

REVISED
25 August 2023

ACCEPTED FOR PUBLICATION
30 August 2023

PUBLISHED
18 September 2023

Abstract

Predictor inputs and labels (e.g. yield data) for crop yield forecasting are not always available at the same spatial resolution. Common statistical and machine learning methods require inputs and labels at the same resolution. Therefore, they cannot produce high resolution (HR) yield forecasts in the absence of HR yield data. We propose a weakly supervised (WS) deep learning framework that uses HR inputs and low resolution (LR) labels (crop areas and yields) to produce HR forecasts. The forecasting model was calibrated by aggregating HR forecasts and comparing with LR crop area and yield statistics. The framework was evaluated by disaggregating yields from parent statistical regions to sub-regions for five countries and two crops in Europe. Similarly, corn yields were disaggregated from counties to 10 km grids in the US. The performance of WS models was compared with naive disaggregation (ND) models, which assigned LR forecasts for a region or county to all HR sub-units, and strongly supervised models trained with HR yield labels. In Europe, all models (ND, WS and strongly supervised) were statistically similar, mainly due to the effect of yield trend. In the US, the WS models performed even better than the strongly supervised models. Based on Kendall's rank correlation coefficient, the WS model forecasts captured significant amounts of HR yield variability. Combining information from WS with Trend model (using LR yield trend) and WS No Trend model (not using yield trend) provided good estimates of yields as well as spatial variability among sub-regions or grids. High resolution crop yield forecasts are useful to policymakers and other stakeholders for local analysis and monitoring. Our weakly supervised framework produces such forecasts even in the absence of high resolution yield data.

1. Introduction

Predictor inputs and label data for crop yield forecasting are often not available at the same spatial resolution. Weather inputs are available at grid-level (Thornton *et al* 2020, EC-JRC 2022) and soil and remote sensing data at sub-kilometer resolutions (ESDAC 2021, Poggio *et al* 2021, Copernicus ESA 2022). Label data (e.g. yield statistics) are published for administrative regions, such as counties or provinces. Common statistical and machine learning methods require strong supervision, i.e. each data point has to have inputs and a corresponding label at the same spatial level. This means strongly supervised models can be built only at the administrative levels where yield statistics are published. Therefore, predictor inputs are aggregated to the level of yield data. High resolution (HR) labels may be unavailable for various reasons. For example, yield statistics are rarely published at grid level, and farm level yield data are typically held by private companies (Deines *et al* 2021). In the absence of HR labels, weakly supervised

learning (Zhou 2018) is still possible using HR inputs and low resolution (LR) labels. Deep learning models can be weakly supervised by using HR inputs to produce HR yield forecasts, which can be aggregated to LR and compared with the labels there. Weakly supervised models limit the spatial aggregation required for input data and produce HR yield forecasts in the absence of HR yield data.

Many studies have used deep learning for crop yield forecasting (Khaki *et al* 2020, Wolanin *et al* 2020, Fan *et al* 2021, Shahhosseini *et al* 2021), but they do not disaggregate yields to high resolutions. Folberth *et al* (2019) attempted disaggregation using gradient boosting (Friedman 2001) and random forests (Breiman 2001). The models were trained on LR inputs and labels and later applied to HR inputs without any further learning or fine tuning. This approach assumes that data for both resolutions come from the same distribution, which is generally unlikely. Other methods of disaggregating crop yields exist, for example, area-to-point kriging (Brus *et al* 2018, Steinbuch *et al* 2020) and spatial allocation based on cross-entropy method (You *et al* 2014) or remote sensing indicators (Kang and Özdoğan 2019, Shirsath *et al* 2020). We draw inspiration from Jacobs *et al* (2018), who trained a convolutional neural network (CNN) and an aggregation layer to predict pixel-level population density from HR satellite images and LR density statistics. To our knowledge, weakly supervised methods have not been used to disaggregate crop yields to HR.

We propose a weakly supervised (WS) deep learning framework that uses HR inputs and LR labels to produce crop yield forecasts for both HR and LR. Since crop area statistics may be unavailable at HR, the framework also estimates the crop area weights for aggregation. Our objective was to evaluate WS models that can produce HR yield forecasts even when HR yields and crop areas are unavailable. This objective was divided into three sub-objectives. First, we assessed the ability of WS models to disaggregate crop yields from low to high resolution. Second, we evaluated the quality of LR yield forecasts produced using HR inputs. Our analysis included two crops (soft wheat and potatoes) and five countries (Germany, Spain, France, Hungary, Italy) in Europe and corn in the US. Third, we analyzed how well weak supervision captures yield variability at HR for an extreme harvest and the following season's harvest.

The contributions of this paper are as follows: (1) we tackled the task of producing HR crop yield forecasts assuming that predictor inputs are available at both HR and LR, but labels are available only at LR. (2) We designed an approach to learn aggregation weights for HR forecasts and to propagate weak supervision signals from LR labels. (3) We demonstrated the performance and benefits of weak supervision in two different settings, Europe and the US, both in terms of agro-environmental factors and spatial resolutions. Our approach is useful to researchers working on similar problems where HR inputs are available, but labels are missing for various reasons.

2. Methods

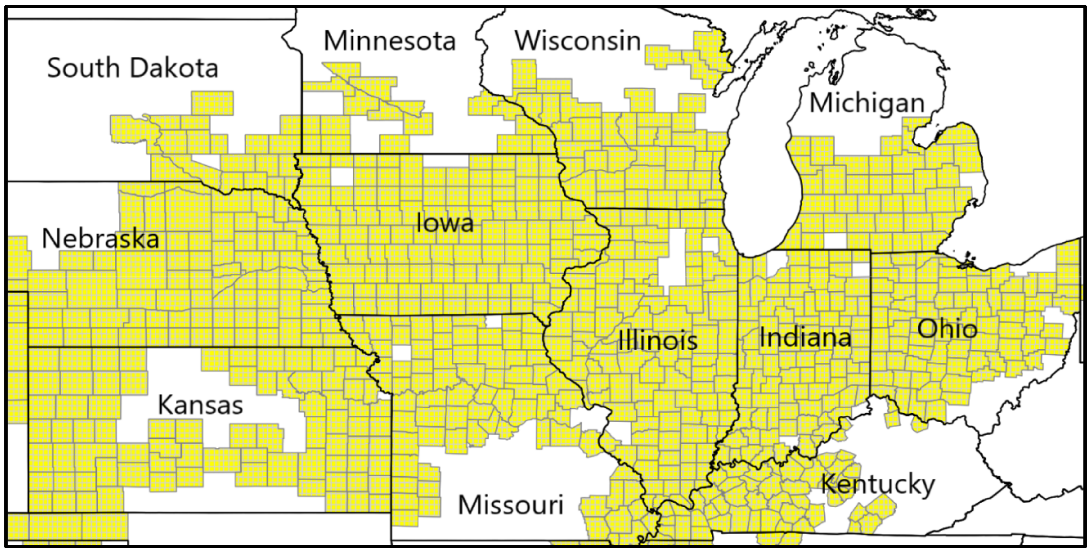
We evaluated WS models in Europe and the US (figure 1). In Europe, WS models used labels from Nomenclature of Territorial Units for Statistics Level 2 (NUTS2) regions (LR) and inputs from NUTS3 regions (HR). NUTS is a hierarchical system of dividing the territory of the European Union for statistics and policy (Eurostat 2016). In the US, WS models used labels from counties (LR) and inputs from 10 km grids (HR).

2.1. Data

European data came from the MARS Crop Yield Forecasting System of the European Commission's Joint Research Centre (MARSWiki 2021) and the Eurostat (Eurostat 2021). The data covered two crops (soft wheat and potatoes) and five countries: Germany, Spain, France, Hungary and Italy. Data from all countries was combined to build one model per crop, mainly due to the small number of NUTS2 labels. Seasonal data included outputs of the WOrld FOod STudies (WOFOST) crop model (van Diepen *et al* 1989, Supit *et al* 1994, de Wit *et al* 2019), weather variables and remote sensing indicators aggregated to NUTS3 and NUTS2 (table 1). The yield trend was captured using yield values of five previous years. Static differences among regions were captured by soil water holding capacity and agro-environmental features, such as elevation, slope and field sizes (Paudel *et al* 2022). In addition, agro-environmental zones and countries were added as categorical variables to account for other agro-climatic and administrative differences. Yield and crop area statistics served as labels. In most cases, we had data from 1999 to 2018. The most recent 30% of the years were allocated to the test set. From the remaining 70% training years, five most recent years were used in a custom five-fold sliding validation (figure A.2) to optimize hyperparameters (i.e. parameters not learned during model training).

For the US, county crop yields and crop areas were exported from the National Agricultural Statistics Service of the US Department of Agriculture (USDA-NASS 2022). 10 km grid inputs came from the Climate Data Store of the Copernicus Climate Change Service (Copernicus CDS 2022) and Copernicus Global Land Service (Copernicus GLS 2020) (table A.1). Grid-level yields published by Deines *et al* (2021), produced using the scalable crop yield mapper approach of Lobell *et al* (2015), were considered ground-truths for grid-level validation since yield statistics are not available for 10 km grids. Overall, we

(a)



(b)

Figure 1. Study areas with low and high resolution units. (a) Europe, (b) the United States. In (a), NUTS2 regions (gray) are shown with their constituent NUTS3 regions (light gray). In (b), US counties (gray) are shown with constituent 10 km grids (light gray).

had data from 2000 to 2018. Training and test splits followed a 70%–30% scheme, similar to European data. Since the data size was larger (about 10× compared to Europe), hyperparameter optimization used a single validation set (five most recent years from the training set), instead of a five-fold sliding validation.

2.2. The weakly supervised framework

The WS framework modified the deep learning framework from Paudel *et al* (2023) to include LR trend features and an aggregation layer. Long short-term memory (LSTM) recurrent neural networks (RNNs) were used to extract features from seasonal data at HR, including crop productivity indicators, weather and remote sensing indicators. Features from LSTM were concatenated with static data and LR yield trend features and passed to the output layer (figure 2), which produced HR yield forecasts and crop area fractions. We believe remote sensing indicators can help predict crop area fractions (crop production area/total land area), but not the absolute crop areas. The aggregation layer

Table 1. Data sources for Europe. In Europe, data sources covered two crops and five countries: soft wheat (DE, ES, FR, IT) and potatoes (DE, FR, HU, IT). The US data covered corn. Data sources for the US are shown in table A.1.

Data	Type of data	Indicators, source
Crop productivity and water balance indicators	Seasonal time series	Water-limited dry weight biomass (kg ha^{-1}), water-limited dry weight storage organs (kg ha^{-1}), water-limited leaf area index (m^2 m^{-2}), development stage (0–200), root-zone soil moisture as % of soil water holding capacity, sum of water limited transpiration (cm). ***Source***: MCYFS. See Lecerf *et al* (2019).
Weather variables	Seasonal time series	Maximum, minimum, average daily air temperature (°C), sum of daily precipitation (PREC, mm), sum of daily evapotranspiration of short vegetation (ET0, mm) (Penman-Monteith, Allen *et al* (1998)), climate water balance = PREC − ET0 (mm), sum of daily global incoming shortwave radiation (kg m^{-2} d^{-1}). ***Source***: MCYFS. See Lecerf *et al* (2019).
Remote sensing indicators	Seasonal time series	Fraction of absorbed photosynthetically active radiation (Smoothed). ***Source***: MCYFS. See Copernicus GLS (2020).
GAES	Static	Agro-environmental zone identifiers. ***Source***: Global agro-environmental stratification (Mücher *et al* 2016).
Crop areas	Yearly	Crop production areas (ha). ***Source***: Eurostat (Eurostat 2021) and MCYFS (EC-JRC 2022).
Irrigated area	Static	Irrigated total area and irrigated crop-specific area (ha). ***Source***: EC-JRC (2022).
Elevation, slope	Static	Average and standard deviation of elevation (m) and slope (degrees). ***Source***: USGS-EROS (2021).
Soil	Static	Soil water holding capacity. ***Source***: MCYFS. See Lecerf *et al* (2019).
Field size	Static	Average and standard deviation (ha). ***Source***: Lesiv *et al* (2019).
Yield	Yearly	Yield at NUTS3 level (t ha^{-1}). NUTS2 level yields were produced by aggregating NUTS3 yields. ***Source***: FR-Agreste (2020), DE-RegionalStatistiks (2020), Eurostat (2021), EC-JRC (2022).

multiplied predicted crop area fractions with land areas to produce HR crop areas, and used them to calculate crop area weights for aggregation. HR yield forecasts were then aggregated to LR. The framework was supervised with NUTS2 or county-level yields and crop areas. Data from all NUTS3 regions within an NUTS2 region, or 10 km grids within a county, formed a batch to enable aggregation of HR forecasts. Model weights were optimized using the Adam optimizer (Kingma and Ba 2014). Hyperparameters optimized included the learning rate and weight decay (aka L2-penalty). Models were retrained with optimal hyperparameters and evaluated on the validation set with early stopping: training stopped after the validation error increased for two successive epochs. Before the final evaluation on the test set, models were retrained on the entire training set (including validation set) with optimal hyperparameters and early stopping epoch.

2.3. Evaluation

Performance of the WS models was compared with two types of models: strongly supervised models and naive disaggregation (ND) models (figure 3). Strongly supervised models were built at HR and LR with inputs and labels from the corresponding spatial level. ND models assigned the forecasts from LR models for parent region or county to all constituent sub-regions or grids. Thus, ND models served as the 'null' models with no prediction skill, while the strongly supervised models provided the bar to beat. LR LSTM and HR LSTM adapted the LSTM framework from Paudel *et al* (2023). The HR LSTM and the WS models used the same HR inputs, except yield trend features. The HR LSTM had access to the HR yield trend; the WS models only had the LR trend. All forecasts were made 60 days before harvest.

We first evaluated WS with Trend forecasts to validate that they are accurate when aggregated to

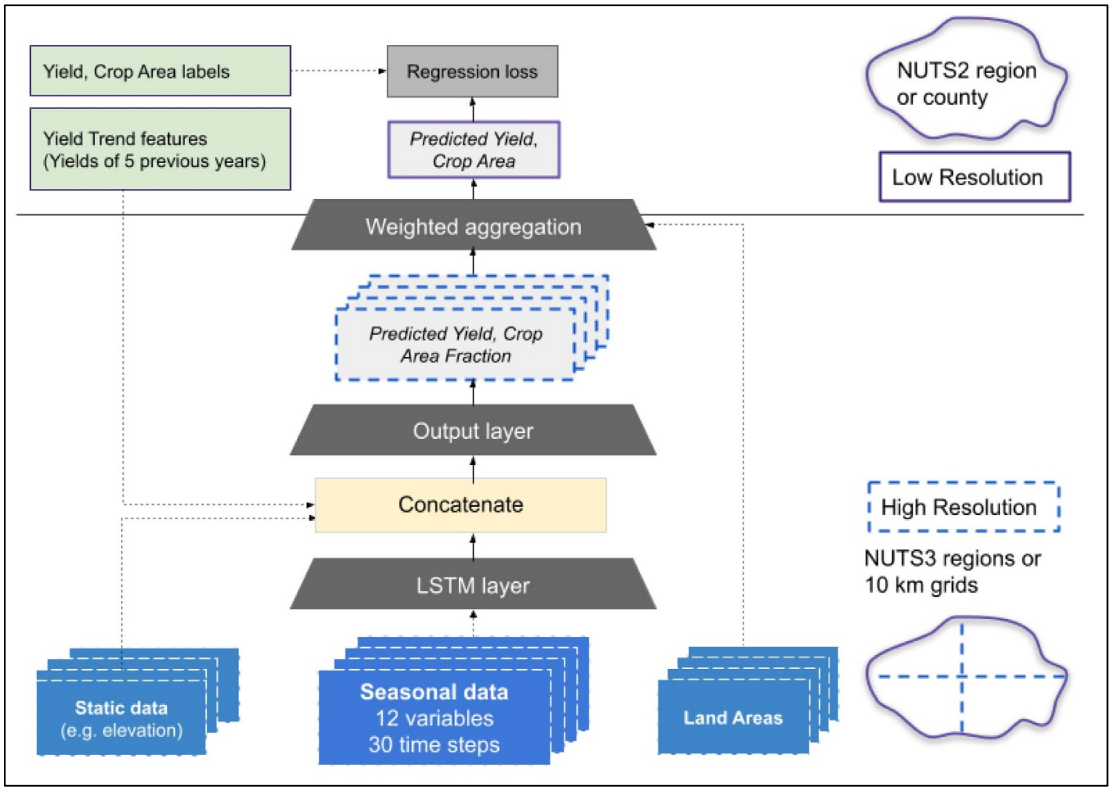


Figure 2. Weakly supervised framework for high resolution crop yield forecasts. The framework used seasonal and static data from high resolution and yield trend from low resolution to produce high resolution forecasts, which were then aggregated to low resolution. The framework was weakly supervised by comparing the aggregated forecasts with low resolution yields and crop areas. The stacked boxes represent data from different high resolution units.

Figure 3. Evaluation framework. The low resolution (LR) models and high resolution (HR) models were trained with inputs and labels from the corresponding (low or high) resolution. The weakly supervised (WS) models were trained with high resolution inputs and low resolution labels. High resolution forecasts from WS models were compared with naive disaggregation (ND) LSTM and HR LSTM. Low resolution forecasts from the WS models were compared with LR LSTM. LR LSTM and HR LSTM are based on the LSTM framework of Paudel *et al* (2023).

low resolution. For this, we compared performance with LR LSTM. More importantly, we were interested in the quality of HR forecasts; we compared them with ND LSTM and HR LSTM forecasts. Model comparison followed the scheme used by Paudel *et al* (2023). Model forecasts were collected from ten models to account for the effect of random weight initializations. We used the average normalized root mean squared errors (NRMSEs), normalized by average yield of the test set, of ten models to compare performance. Variance and outliers were analyzed using boxplots of prediction residuals (predicted yield − reported yield). Significance of model performance was evaluated using the Mann–Whitney

U test (Mann and Whitney 1947), which is a non-parametric version of Student's *t*-test for independent samples. Prediction residuals used for boxplots and statistical tests were averaged across the ten models.

We also analyzed the spatial variability of HR forecasts for soft wheat in Europe and corn in the US. A significant part of yield variability is explained by the yield trend attributed to factors such as technological improvements (see Lecerf *et al* 2019). In figure 2, we expected the LR trend to make the WS model more accurate, but suppress spatial variability at HR. Therefore, we ran another version of the WS model without NUTS2 or county trend. The two versions are called WS with Trend and WS No Trend (figure 3). Kendall's rank correlation coefficient, or Kendall's tau (Kendall 1938), was used to quantify the skill to capture spatial yield variability at HR. For example, the ranking of NUTS3 yield forecasts within the same NUTS2 region was compared with the ranking of NUTS3 yields to compute Kendall's tau. Kendall's tau of the WS models were compared with those of HR LSTM. A high correlation (and significance based on *p*-value) would show that forecasts captured the relative differences in yields among NUTS3 regions. To illustrate the spatial yield variability captured by different models, maps of yield forecasts vs. yield statistics were plotted for an extreme harvest and the following season's harvest. In Europe, NUTS3 regions were selected based on maximum acreage for soft wheat (France) and years (2016 and 2017) based on significant yield losses reported in the north of France in 2016 (see Ben-Ari *et al* 2018). In the US, spatial variability of 10 km grid yields was analyzed for 2012, when there was a severe drought (Rippey 2015), and 2013.

3. Results

In this section, performance comparison results are reported for the WS with Trend model and spatial variability analysis includes the WS No Trend model as well.

3.1. Evaluation of low resolution yield forecasts

For both soft wheat and potatoes in Europe, LR (NUTS2) forecasts of WS with Trend and LR LSTM were statistically similar (*p*-value 0.8534 and 0.5274 respectively) (table A.4). Box plots of prediction residuals and per-country average NRMSEs were also generally similar, although WS with Trend had more stable NRMSEs (i.e. lower variance), especially for potatoes (figures 4(a) and (b); table A.2). Overall, WS with Trend model was not significantly better than the LR LSTM models despite using HR inputs.

In the US, HR inputs did make WS with Trend model significantly better; county-level corn forecasts were better than those from LR LSTM model (*p*-values near 0) (table A.6). The LR LSTM had a similar average NRMSE, but with a higher variance; it also underestimated the yields more compared to WS with Trend (figure 4(c)).

3.2. Evaluation of high resolution yield forecasts

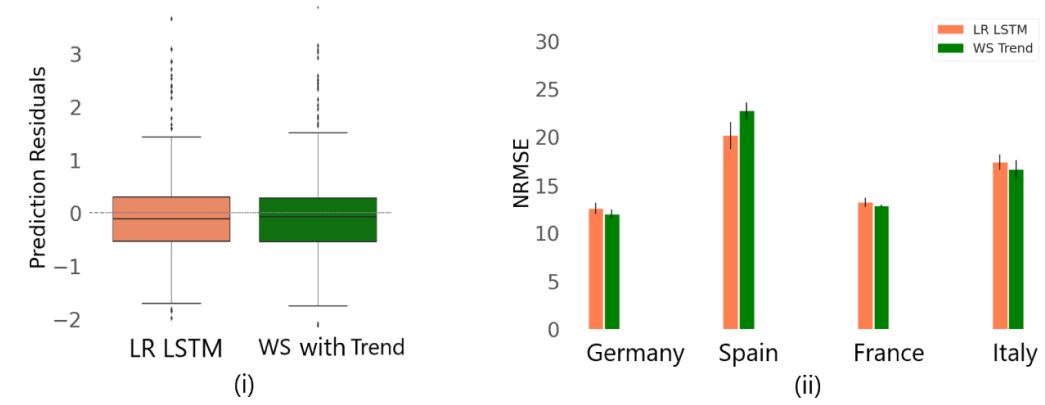
For soft wheat and potatoes in Europe, HR (NUTS3) forecasts of all three models—WS with Trend, ND LSTM and HR LSTM—were statistically similar (tables A.7 and A.8). Box plots of ND LSTM and WS with Trend were similar, but the latter had slightly higher average NRMSEs, especially for soft wheat in Germany and Spain (figure 5(a); table A.3). For potatoes, the similarity between WS with Trend and HR LSTM was less conclusive because median residuals were quite different (0.3588 vs. 0.2123) and HR LSTM had much lower average NRMSEs (figure 5(b)). Overall, performance results in Europe did not show weak supervision to be better than naive disaggregation to NUTS3.

For corn in the US, weak supervision did produce better HR forecasts than ND LSTM as well as the HR LSTM (*p*-values near zero) (table A.7). HR LSTM had the lowest average NRMSE, but box plots showed that WS with Trend had prediction residuals closer to zero than ND LSTM and HR LSTM (figure 5(c); table A.3). The median prediction residuals were −0.546 for ND LSTM, −0.132 for WS with Trend and −0.443 for HR LSTM.

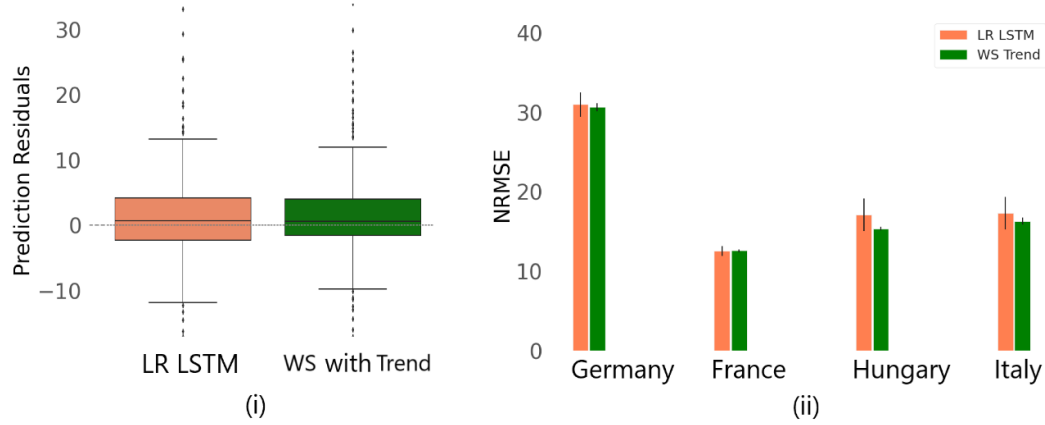
3.3. Spatial variability of high resolution forecasts

At HR in Europe, WS with Trend forecasts for soft wheat were not better than naively disaggregated values and the same was true for HR LSTM forecasts. Even then, ND models provide no information about HR yield variability because they assign the same value to all NUTS3 regions within a NUTS2 region. Kendall's tau for WS models showed that weak supervision does provide information about spatial yield variability. Kendall's tau values were 0.265 for WS with Trend, 0.357 for WS No Trend and 0.578 for HR LSTM (*p*-values near zero indicating significance). As expected, WS No Trend model had a higher correlation coefficient than WS with Trend model. For corn in the US, Kendall's tau values were 0.278 for WS with Trend and 0.327 for WS No Trend and 0.532 for HR LSTM (*p*-values near zero).

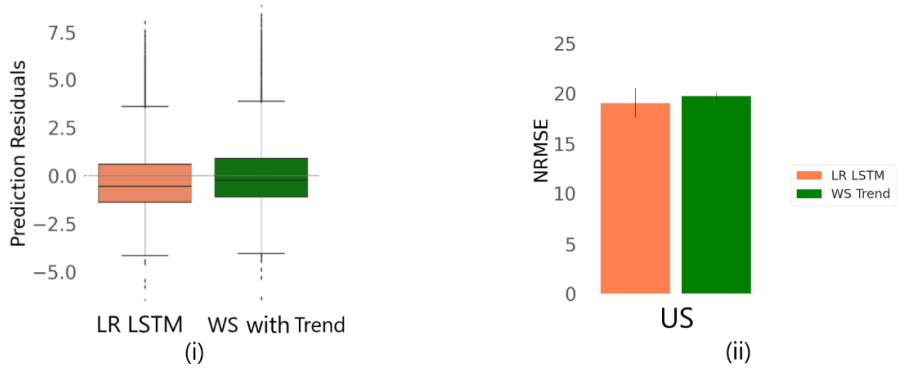
For soft wheat in France, ND LSTM predicted higher yields in 2016, with an average prediction residual of 0.362. The maps showed that WS with Trend and HR LSTM were also influenced by the yield trend and overestimated yields (average prediction residuals: 0.280 and 0.395 respectively). Their forecasts looked quite similar (figure 6(a)). WS No Trend model captured the yield losses better with an average residual of 0.064. In 2017, ND LSTM produced more accurate forecasts (average prediction residual −0.144), but provided no information about NUTS3 level yield variability (figure 6(b)). WS with Trend model forecasts looked similar to ND LSTM forecasts and did not show visible differences among NUTS3



(a) Soft wheat, NUTS2 (Europe). (i): Box plots of prediction residuals. (ii): Per-country average NRMSEs.



(b) Potatoes, NUTS2 (Europe). (i): Box plots of prediction residuals. (ii): Per-country average NRMSEs.

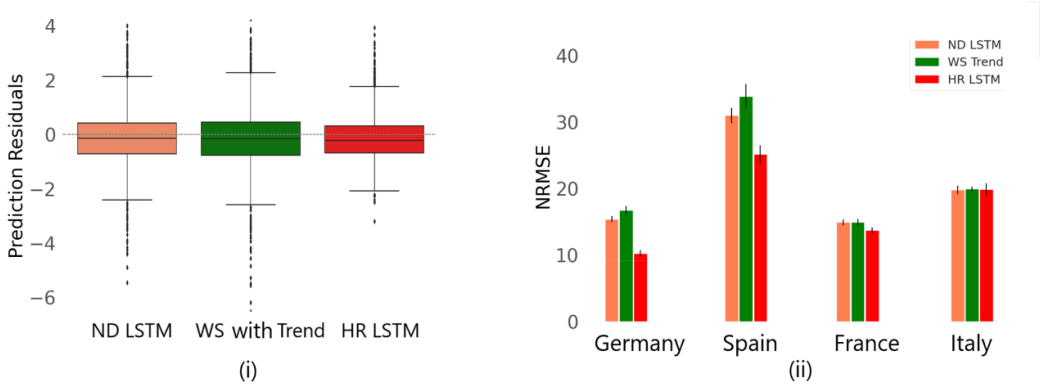


(c) Corn, county (US). (i): Box plots of prediction residuals. (ii): Average NRMSEs.

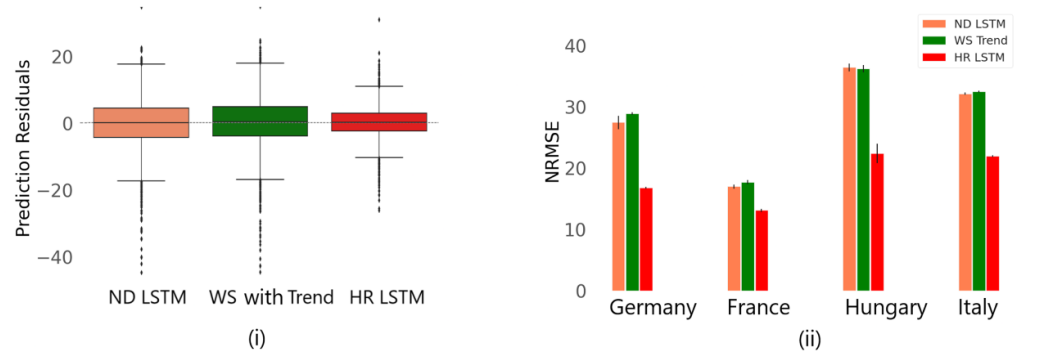
Figure 4. Evaluation of low resolution forecasts 60 days before harvest. Weak supervision with trend (WS with Trend) is compared with low resolution LSTM (LR LSTM). Prediction residuals used for the boxplots and per-country NRMSEs were averaged across ten models. The whiskers in the bar plots indicate the standard deviation of NRMSEs for the ten models.

sub-regions within NUTS2 regions. WS No Trend model captured such differences better, and the forecasts looked similar to HR LSTM forecasts. Because it did not use yield trend, WS No Trend model underestimated the yields: the average prediction residual was −0.516 compared to −0.135 for WS with Trend and −0.061 for HR LSTM.

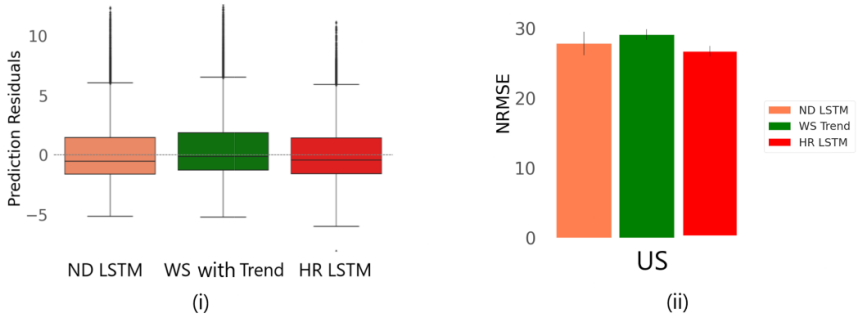
For corn in the US, we evaluated spatial variability for 2012 and 2013 because of the well-known drought of 2012 (Rippey 2015). For 2012, WS No Trend model captured the yield losses better, while WS with Trend model overestimated the yields (figure A.3). The average prediction residuals were 4.76 for ND LSTM, 5.27 for WS with Trend and 3.60 for WS No Trend. The WS No Trend model also captured differences among grids better than the WS with Trend model (Kendall's tau: 0.315 vs. 0.265), especially within some counties in Illinois. In 2013, all models produced more accurate forecasts compared to 2012 (figure A.4). Grid-level variability was difficult to compare visually, but



(a) Soft wheat, NUTS3 (Europe). (i): Boxplots of prediction residuals. (ii): Per-country NRMSEs.



(b) Potatoes, NUTS3 (Europe). (i): Boxplots of prediction residuals. (ii): Per-country NRMSEs.



(c) Corn, 10-km grids (US). (i): Boxplots of prediction residuals. (ii): NRMSEs.

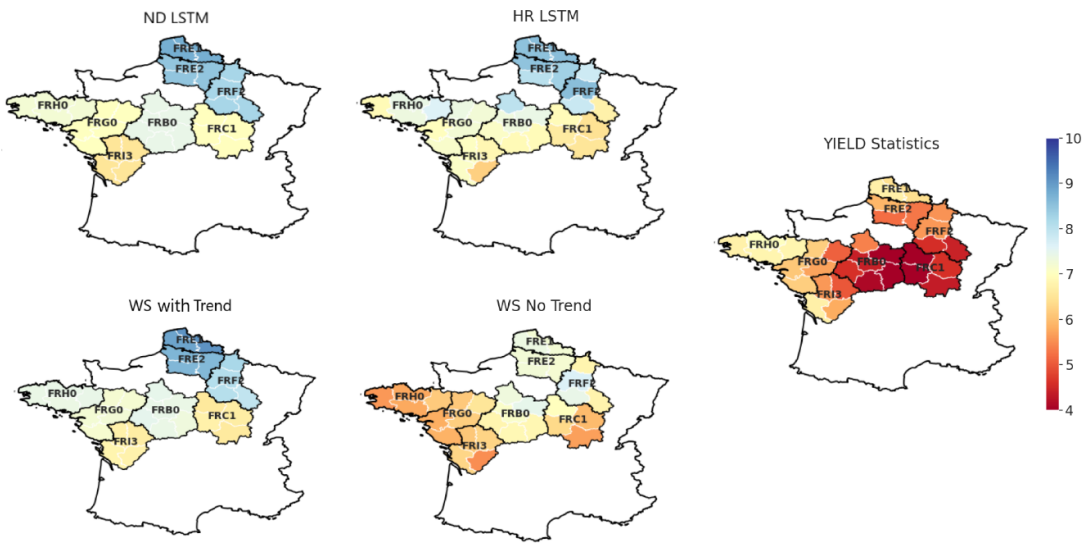
Figure 5. Evaluation of high resolution forecasts 60 days before harvest. Weak supervision with trend (WS with Trend) is compared with naive disaggregation LSTM (ND LSTM) and high resolution LSTM (HR LSTM). Prediction residuals used for the boxplots and per-country NRMSEs were averaged across ten models. The whiskers in the bar plots indicate the standard deviation of NRMSEs for the ten models.

WS No Trend had a higher Kendall's tau than WS with Trend (0.303 vs. 0.242).

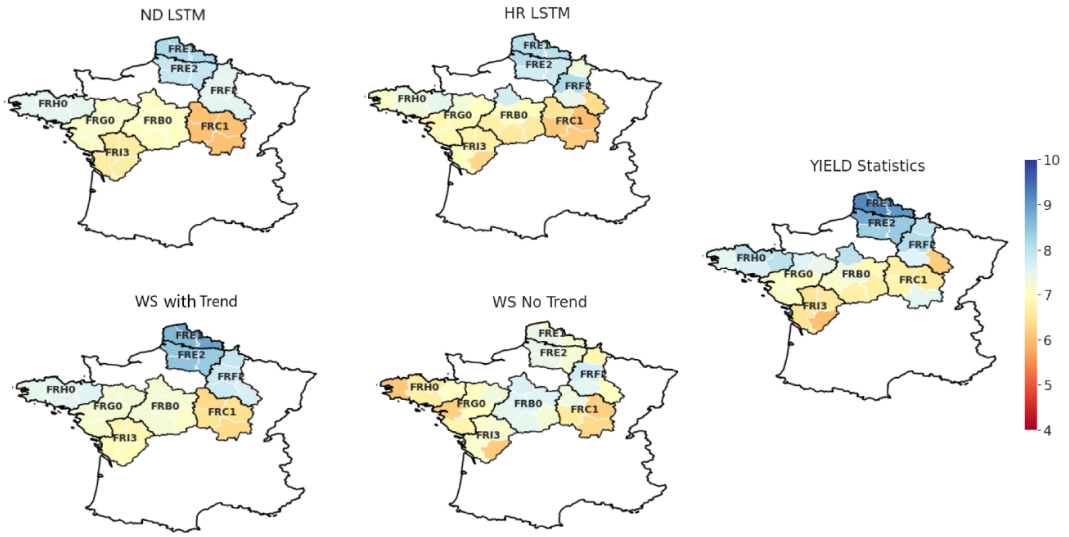
Overall, the LR trend was more useful to WS with Trend in Europe than in the US. The Pearson's r for NUTS2 trend and NUTS3 yields was 0.81 for soft wheat and 0.61 for potatoes. The corresponding value for county trend and 10 km grid yields was 0.33. This makes sense because NUTS3 regions are much larger than 10 km grids and yield trend is more pronounced at larger spatial levels, where variability due to other factors tends to average out.

4. Discussion

The reliance of standard machine learning methods on strong supervision can be a limitation when yield labels are unavailable. Weakly supervised deep learning methods address this limitation by learning from HR inputs and LR labels. We have shown that weak supervision can disaggregate crop yields from low to high resolution. In Europe, WS with Trend models were not better than ND LSTM models, but they were statistically similar to HR LSTM models, which were themselves similar to ND LSTM models. Therefore, WS with Trend forecasts were as good as those from HR LSTM, especially for soft wheat. At the same time, the WS models, especially WS No Trend model, captured some NUTS3-level yield variability. Information from WS with Trend and WS No Trend could be combined, for example using a weighted average, to produce more accurate forecasts. Another approach would be to train the WS No Trend model



(a) Soft wheat, France (2016). An extreme harvest in the north of France.



(b) Soft wheat, France (2017).

Figure 6. Spatial variability of soft wheat yields and forecasts. Forecasts from weak supervised with trend (WS with Trend) and weak supervised no trend (WS No Trend) models are compared with naive disaggregation LSTM (ND LSTM) and high resolution LSTM (HR LSTM). The WS No Trend model produced more accurate forecasts for 2012 and also captured NUTS3 level yield variability better than ND LSTM and WS with Trend.

to predict yield residuals from the trend. In general, WS with Trend forecasts indicate where the yield level should be. WS No Trend forecasts provide information about deviations from that yield level. Future work could develop a consistent method of selecting one WS model or combining information from the two models.

Our WS framework was adapted to forecast corn yields for 10 km grids in the US with minimal changes related to data preprocessing and larger (approximately 10×) data size. WS with Trend forecasts were significantly better than LR LSTM and HR LSTM forecasts for counties and grids respectively. Both WS models captured some grid-level variability within counties, and WS No Trend model also captured some yield losses due to drought in 2012 (figure A.3). County-level NRMSEs were quite similar to those reported by other studies. For example, Khaki *et al* (2020) used a CNN–RNN framework and reported an NRMSE of 9% for 2016–2018. WS with Trend had a corresponding NRMSE of 10.54%. Future work could experiment with other architectures, another crop model (e.g. WOFOST or Agricultural Production Systems sIMulator (APSIM) (Holzworth *et al* 2014)) and additional farm management information to improve the performance of WS models.

Performance of WS models was quite different between Europe and the US. In the US, WS with Trend models were better than even the strongly supervised HR LSTM. In Europe, they were similar to ND LSTM as well as HR LSTM. To understand this similarity, we analyzed the effect of yield trend using Trend Only

models. LR and HR Trend Only models fitted a line through yield values of five previous years. The ND Trend Only model naively disaggregated LR Trend Only forecasts to high resolution. In Europe, ND Trend Only models were statistically similar to HR Trend Only models for both soft wheat and potatoes. This similarity between LR and HR trends, coupled with high correlation between LR trend and HR yields (Pearson's r: 0.81 for soft wheat and 0.61 for potatoes), had a significant influence in making the models similar to each other. In the US, ND Trend Only and HR Trend Only models were statistically different, and LR trend was correlated less with HR yields (Pearson's r: 0.33). The weaker influence of LR trend may have helped the WS models to learn better from HR inputs. Consequently, WS with Trend was statistically better than strongly supervised models at both resolutions. Apart from the influence of yield trend, weak supervision could be affected by differences in spatial resolution, the number of sub-regions or grids within a region or county, the quality of estimated crop area weights, quality of LR labels and yield variability at HR. Research into the effect of these factors on the performance of WS models will bring clarity to when and where weak supervision will work.

In this paper, we have scratched the surface of high resolution crop yield forecasting with weak supervision. We see three areas that need further research to gauge the benefits and limitations of weakly supervised methods. First, more work is needed to understand the scale differences that can be handled by weak supervision. For example, weak supervision worked well between counties to 10 km grids in the US and less so between NUTS2 and NUTS3 regions in Europe. For very large differences in resolution (e.g. NUTS3 regions or counties to 1 km grids), supervision signals from LR labels may be insufficient to capture HR differences. Second, predictor inputs must be suitable to capture yield variability at selected resolutions. Crop simulation outputs and weather variables may correlate well with yields at NUTS3 or 10 km grids, but become less relevant at farm or parcel level. HR remote sensing data, for example from Sentinel satellites (Copernicus ESA 2022), and ground measurements may provide better predictors for farm level yields. Third, we experimented with standard neural network architectures. Future work could investigate other architectures that are more suitable for weak supervision. As mentioned above, architectures that combine strengths of CNNs and RNNs to learn both spatial and temporal features are also worth exploring. Data size and quality will always play a role due to the data-driven nature of neural networks.

Crop yield predictors will become available at increasingly high resolution. Yield data may be missing due to many reasons, including privacy concerns. When there is an imbalance between spatial resolutions of inputs and yields, weakly supervised methods provide a solution. Our approach will continue to work when HR yield data becomes available for some regions but not others. Deep learning may also provide a way to better optimize the crop area weights. HR crop areas, when available, will remove the need to estimate them and further improve the quality of yield forecasts. HR crop yield forecasts improve the effectiveness of policy interventions targeted to food security, agricultural production and resource sustainability (You *et al* 2014). We have shown that weakly supervised methods can produce such forecasts in the absence of HR labels.

5. Conclusions

We designed a weakly supervised deep learning framework that uses HR inputs and LR labels to produce crop yield forecasts for both resolutions. Evidence from NUTS2 to NUTS3 in Europe and county to 10 km grids in the US showed that weak supervision performs similarly well or better than strong supervision in different settings, both in terms of agro-climatic factors and spatial resolutions. Forecasts from WS No Trend models captured a significant amount of HR yield variability and produced more accurate forecasts for extreme harvests. The framework can be improved with additional data sources, including HR crop areas, a better understanding of factors affecting weak supervision and neural network architectures that can capture both spatial and temporal differences. Overall, high resolution crop yield forecasts are useful to farmers, policymakers and other stakeholders as they provide more detailed information about local yield variability. Weakly supervised methods provide a way to produce such forecasts when HR yield data is unavailable.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.7751190.

Acknowledgments

This work was partially supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 825355 (CYBELE).

We would like to thank S Niemeyer from the European Commission's Joint Research Centre (JRC) for the permission to use the MARS Crop Yield Forecasting System (MCYFS) data. Similarly, we would like to thank M van der Velde, L Nisini and I Cerrani from JRC for preparing and sharing the Eurostat regional yield statistics and crop areas.

ORCID iDs

Dilli Paudel https://orcid.org/0000-0003-4080-4276
Allard de Wit https://orcid.org/0000-0002-5517-6404

References

Allen R G, Pereira L S, Raes D and Smith M 1998 Crop evapotranspiration—guidelines for computing crop water requirements irrigation and drainage, paper 56 (FAO)
Ben-Ari T, Boé J, Ciais P, Lecerf R, Van der Velde M and Makowski D 2018 Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France *Nat. Commun.* **9** 1–10
Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
Brus D, Boogaard H, Ceccarelli T, Orton T, Traore S and Zhang M 2018 Geostatistical disaggregation of polygon maps of average crop yields by area-to-point kriging *Eur. J. Agron.* **97** 48–59
Copernicus CDS 2022 *Copernicus Climate Data Store* (Copernicus Climate Change Service) (available at: https://cds.climate.copernicus.eu/)
Copernicus ESA 2022 Sentinel earth observation data. Copernicus open access hub (available at: https://scihub.copernicus.eu/)
Copernicus GLS 2020 Fraction of absorbed photosynthetically active radiation (Copernicus Global Land Service) (https://doi.org/10.24381/cds.7e59b01a)
de Wit A, Boogaard H, Fumagalli D, Janssen S, Knapen R, van Kraalingen D, Supit I, van der Wijngaart R and van Diepen K 2019 25 years of the WOFOST cropping systems model *Agric. Syst.* **168** 154–67
DE-RegionalStatistiks 2020 Regionaldatenbank Deutschland (available at: www.regionalstatistik.de/genesis/online/data) (Accessed 11 May 2020)
Deines J M, Patel R, Liang S Z, Dado W and 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
EC-JRC 2022 JRC Agri4Cast data portal (available at: https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx) (Accessed 11 February 2022)
ESDAC 2021 European soil database (available at: https://esdac.jrc.ec.europa.eu/resource-type/datasets) (Accessed 28 April 2021)
Eurostat 2016 Nomenclature of territorial units for statistics (available at: https://ec.europa.eu/eurostat/web/nuts/background) (Accessed 11 May 2020)
Eurostat 2021 Eurostat—agricultural production—crops (available at: https://ec.europa.eu/eurostat/statistics-explained/index.php/Agricultural_production_-_crops) (Accessed 11 May 2021)
Fan J, Bai J, Li Z, Ortiz-Bobea A and Gomes C P 2021 A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction (arXiv:2111.08900v2)
Folberth C, Baklanov A, Balkovič J, Skalský R, Khabarov N and Obersteiner M 2019 Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning *Agric. For. Meteorol.* **264** 1–15
FR-Agreste 2020 Agreste web data portal (available at: https://agreste.agriculture.gouv.fr/agreste-web/) (Accessed 11 May 2020)
Friedman J H 2001 Greedy function approximation: a gradient boosting machine *Ann. Stat.* **29** 1189–232
Holzworth D P *et al* 2014 Apsim–evolution towards a new generation of agricultural systems simulation *Environ. Model. Softw.* **62** 327–50
Jacobs N, Kraft A, Rafique M U and Sharma R D 2018 A weakly supervised approach for estimating spatial density functions from high-resolution satellite imagery *Proc. 26th ACM SIGSPATIAL Int. Conf. on Advances in Geographic Information Systems* pp 33–42
Kang Y and Özdoğan M 2019 Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach *Remote Sens. Environ.* **228** 144–63
Kendall M G 1938 A new measure of rank correlation *Biometrika* **30** 81–93
Khaki S, Wang L and Archontoulis S V 2020 A CNN-RNN framework for crop yield prediction *Front. Plant Sci.* **10** 1750
Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.69803)
Lecerf R, Ceglar A, López-Lozano R, Van Der Velde M and Baruth B 2019 Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe *Agric. Syst.* **168** 191–202
Lesiv M *et al* 2019 Estimating the global distribution of field size using crowdsourcing *Globe Change Biol.* **25** 174–86
Lobell D B, Thau D, Seifert C, Engle E and Little B 2015 A scalable satellite-based crop yield mapper *Remote Sens. Environ.* **164** 324–33
Mann H B and Whitney D R 1947 On a test of whether one of two random variables is stochastically larger than the other *Ann. Math. Stat.* **18** 50–60
MARSWiki 2021 MARS Crop Yield Forecasting System (available at: https://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Welcome_to_WikiMCYFS) (Accessed 11 May 2021)
Mücher S, De Simone L, Kramer H, de Wit A, Roupioz L, Hazeu G, Boogaard H, Schuiling R, Fritz S, Latham J and Cormont A 2016 A new global agro-environmental stratification (GAES) *Technical Report* (Wageningen Environmental Research) (available at: https://edepot.wur.nl/400815) (Accessed 14 June 2021)
Paudel D, Boogaard H, de Wit A, van der Velde M, Claverie M, Nisini L, Janssen S, Osinga S and Athanasiadis I N 2022 Machine learning for regional crop yield forecasting in Europe *Field Crops Res.* **276** 108377
Paudel D, de Wit A, Boogaard H, Marcos D, Osinga S and Athanasiadis I N 2023 Interpretability of deep learning models for crop yield forecasting *Comput. Electron. Agric.* **206** 107663
Poggio L, De Sousa L M, Batjes N H, Heuvelink G, Kempen B, Ribeiro E and Rossiter D 2021 Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty *Soil* **7** 217–40
Rippey B R 2015 The US drought of 2012 *Weather Clim. Extrem.* **10** 57–64
Shahhosseini M, Hu G, Huber I and Archontoulis S V 2021 Coupling machine learning and crop modeling improves crop yield prediction in the US corn belt *Sci. Rep.* **11** 1–15
Shirsath P B, Sehgal V K and Aggarwal P K 2020 Downscaling regional crop yields to local scale using remote sensing *Agriculture* **10** 58
Steinbuch L, Orton T G and Brus D J 2020 Model-based geostatistics from a Bayesian perspective: investigating area-to-point kriging with small data sets *Math. Geosci.* **52** 397–423
Supit I, Hooijer A and Van Diepen C 1994 System description of the WOFOST 6.0 crop simulation model implemented in CGMS. Vol. 1. Theory and algorithms *EUR Publication No. 15959 EN* (Office for Official Publications of the European Communities) p 146
Thornton M, Wei Y, Thornton P, Shrestha R, Kao S and Wilson B 2020 Daymet: station-level inputs and cross-validation result for North America, version 4 (available at: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1850)
USDA-NASS 2022 Statistics by subject—crops (available at: www.nass.usda.gov/Statistics_by_Subject/index.php?sector=CROPS) (Accessed 10 August 2022)

USGS-EROS 2021 USGS EROS archive—digital elevation—global 30 arc-second elevation (GTOPO30) (available at: www.usgs.gov/centers/eros/data) (Accessed 11 May 2021)

van Diepen C, Wolf J, Van Keulen H and Rappoldt C 1989 WOFOST: a simulation model of crop production *Soil Use Manag.* **5** 16–24

Wolanin A, Mateo-García G, Camps-Valls G, Gómez-Chova L, Meroni M, Duveiller G, Liangzhi Y and Guanter L 2020 Estimating and understanding crop yields with explainable deep learning in the Indian wheat belt *Environ. Res. Lett.* **15** 024019

You L, Wood S, Wood-Sichra U and Wu W 2014 Generating global crop distribution maps: from census to grid *Agric. Syst.* **127** 53–60

Zhou Z H 2018 A brief introduction to weakly supervised learning *Natl Sci. Rev.* **5** 44–53