Field Crops Research 276 (2022) 108377

Contents lists available at ScienceDirect

Field Crops Research

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

Machine learning for regional crop yield forecasting in Europe

Dilli Paudel [a,*,1], Hendrik Boogaard [b], Allard de Wit [b], Marijn van der Velde [d], Martin Claverie [d], Luigi Nisini [d], Sander Janssen [b], Sjoukje Osinga [a], Ioannis N. Athanasiadis [c]

[a] Information Technology Group, Wageningen University and Research, Hollandseweg 1, 6706 KN Wageningen, The Netherlands
[b] Wageningen Environmental Research, PO Box 47, 6700 AA Wageningen, The Netherlands
[c] Geo-information and Remote Sensing Group and Wageningen Data Competence Center, Wageningen University and Research, PO Box 47, 6700 AA Wageningen, The Netherlands
[d] European Commission, Joint Research Centre (JRC), Ispra 21027, Italy

ARTICLE INFO

Keywords:
Crop yield
Machine learning
Large-scale crop yield forecasting
Scalability
Regional differences

ABSTRACT

Crop yield forecasting at national level relies on predictors aggregated from smaller spatial units to larger ones according to harvested crop areas. Such crop areas come from land cover maps or reported statistics, both of which can have errors and uncertainties. Sub-national or regional crop yield forecasting minimizes the propagation of these errors to some extent. In addition, regional forecasts provide added value and insights to stakeholders on regional differences within a country, which would otherwise compensate each other at national level. We propose a crop yield forecasting approach for multiple spatial levels based on regional crop yield forecasts from machine learning. Machine learning, with its data-driven approach, can leverage larger data sizes and capture nonlinear relationships between predictors and yield at regional level. We designed a generic machine learning workflow to demonstrate the benefits of regional crop yield forecasting in Europe. To evaluate the quality and usefulness of regional forecasts, we predicted crop yields for 35 case studies, including nine countries that are major producers of six crops (soft wheat, spring barley, sunflower, grain maize, sugar beets and potatoes). Machine learning models at regional level had lower normalized root mean squared errors (NRMSE) and uncertainty than a linear trend model, with Wilcoxon p-values of 3e-7 and 2e-7 for 60 days before harvest and end of season respectively. Similarly, regional machine learning forecasts aggregated to national level had lower NRMSEs than forecasts from an operational system in 18 out of 35 cases 60 days before harvest, with a Wilcoxon p-value of 0.95 indicating similar performance. Our models have room for improvement, especially during extreme years. Nevertheless, regional crop yield forecasts from machine learning and aggregated national forecasts provide a consistent forecasting method across spatial levels and insights from regional differences to support important policy decisions.

1. Introduction

Crop yields vary across space because of differences in soil, climatic conditions and agro-management practices. Crop yield forecasts at different spatial levels benefit various stakeholders, including farmers and policymakers. Such forecasts provide added value when they are available at smaller units or higher spatial resolutions. Reliable forecasts at higher spatial resolution help explain yield variability at coarser levels and also provide information to adapt agricultural policies to more specific areas (García-León et al., 2020).

Most large-scale crop yield forecasting systems worldwide, such as the MARS Crop Yield Forecasting System (MCYFS) of the European Commission's Joint Research Centre (MARSWiki, 2021), the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA-NASS, 2012), and Statistics Canada (Statistics Canada, 2021), use different methods to forecast crop yields at various spatial levels. While NASS estimates crop yield at Agricultural Statistics Districts (ASD) and aggregates them to state level, Statistics Canada and

* Corresponding author.
E-mail addresses: dilli.paudel@wur.nl (D. Paudel), hendrik.boogaard@wur.n (H. Boogaard), allard.dewit@wur.nl (A. de Wit), marijn.van-der-velde@ec.europa.eu (M. van der Velde), martin.claverie@ec.europa.eu (M. Claverie), luigi.nisini-scacchiafichi@ext.ec.europa.eu (L. Nisini), sander.janssen@wur.nl (S. Janssen), sjoukje.osinga@wur.nl (S. Osinga), ioannis.athanasiadis@wur.nl (I.N. Athanasiadis).
[1] https://orcid.org/0000-0003-4080-4276.

https://doi.org/10.1016/j.fcr.2021.108377
Received 14 October 2021; Received in revised form 18 November 2021; Accepted 19 November 2021
Available online 2 December 2021

MCYFS aggregate input data from small spatial units to build forecasting models at ecological or provincial level and national level respectively. Within MCYFS, predictors such as crop model outputs, weather variables and remote sensing indicators are aggregated from one spatial level to the next based on crop areas derived from land cover maps and crop area statistics. Land cover maps for most crops (except rice) are not crop-specific (Bartholome and Belward, 2005; Buttner et al., 2004) and crop area statistics are collected using a diverse set of country-specific methods. Therefore, aggregation of inputs to national level accumulates uncertainties and errors associated with crop masks as well as data collection and interpolation methods (Cerrani and López Lozano, 2017). Forecasting crop yields at regional level can minimize some of the aggregation errors. Using data from Canada, Chipanshi et al. (2015) showed that predicting yields at smaller spatial units and aggregating them to larger ones produced better results than aggregating predictors and building models at larger units.

Crop yield forecasting at smaller spatial units has its share of challenges. As we go to smaller units, we find that data quality deteriorates. Regional yield statistics are not curated as well as the national statistics, and data collection protocols and data quality vary from one country to another (López-Lozano et al., 2015). For systems that predict at national level, such as MCYFS, regional yield forecasting introduces challenges of scaling the analyst-driven methodology to hundreds of regions. Despite the challenges, the benefits of operationalizing regional yield forecasting could outweigh the costs. As spatial differences and uncertainties cancel out at national level (Porwollik et al., 2017), national forecasts may capture temporal yield variability well. However, they do not provide information about spatial variability. In particular, unfavorable crop conditions in some regions may be compensated by favorable conditions in other regions (Seguini et al., 2019). Regional forecasts provide useful information at the same level as well as at larger spatial units. Yield variability at the provincial or national level can be explained in terms of patterns in constituent regions. Similarly, predictors at regional levels suffer less from aggregation errors and may correlate better with yield values, producing more reliable prediction models (see Bussay et al., 2015). Another side effect of regional crop yield forecasting would be an increased understanding of data quality and the motivation to improve data collection and curation protocols. Furthermore, forecasting crop yields at regional level and subsequently aggregating regional forecasts to larger spatial units provides consistency in the forecasting method at all spatial levels involved.

Machine learning, with its data-driven approach, could benefit from the increased data size at regional level. Similarly, machine learning algorithms can model nonlinear relationships between multiple data sources and yields at regional level. Machine learning methods have been used to predict crop yield at sub-national levels outside of Europe (Han et al., 2020; Cai et al., 2019; Crane-Droesch, 2018; You et al., 2017). In Europe, most of the studies on regional yield forecasting (Pagani et al., 2019; Ceglar et al., 2016; Gouache et al., 2015; López-Lozano et al., 2015; Bussay et al., 2015) do not use machine learning. Paudel et al. (2021) have previously shown the promise of regional crop yield forecasting using machine learning for five crops in Germany, France and the Netherlands. Machine learning can also address scaling issues associated with regional crop yield forecasting. In systems such as MCYFS, analysts build a large number of statistical models at national level and select models based on expertise and contextual information (Van der Velde and Nisini, 2019). At regional level, analyst-driven crop yield forecasting would require a lot more time and effort. Machine learning methods can use regional data to build one model per country and automate many steps, such as feature selection and hyperparameter optimization. A generic and scalable machine learning workflow could be complementary to the analyst-driven crop yield forecasting: enable analysts to leverage the data-driven approach in most cases and apply the expertise-based approach to cases where machine learning does not provide reliable predictions.

In this paper, we propose a crop yield forecasting approach for multiple spatial levels based on regional forecasts from machine learning. Our objective is to build models at regional level and evaluate their quality and usefulness in capturing spatial and temporal yield variability across regions as well as larger spatial divisions. We extended the machine learning workflow introduced by Paudel et al. (2021) and predicted crop yields at the NUTS level (Eurostat, 2016b) where yield and crop area statistics are available. The data for evaluation came from MCYFS and Eurostat, and included nine European countries that are major producers of six crops (soft wheat, spring barley, sunflower, grain maize, sugar beets, potatoes). Prediction skill of machine learning models was compared with a linear trend model at regional level and past MCYFS forecasts at national level. The uncertainty of regional forecasts was estimated for cases where regional differences would cancel out at national level. Similarly, regional forecasts for an average harvest and two extreme harvests were analyzed to demonstrate how well they capture the spatial yield variability. Our approach introduces a consistent and reproducible method to forecast crop yield at multiple spatial levels.

The rest of the paper is structured as follows. *Section 2* describes the data and methods, *Section 3* presents the results, *Section 4* discusses our findings and outlines areas for future work and *Section 5* summarizes our conclusions. Appendix A and Appendix B provide details and supporting evidence not included in Section 2 (*Materials and Methods*) and Section 3 (*Results*) and Section 4 (*Discussion*).

2. Material and methods

As stated in *Section 1* above, our objective is to evaluate the prediction skill and usefulness of crop yield forecasts from machine learning at regional level as well as larger spatial levels. Using 35 *case studies* (i.e. crop and country combinations) from Europe, machine learning models were built to produce crop yield forecasts at regional and national levels. To assess prediction skill, regional forecasts were compared with trend forecasts and national forecasts with MCYFS forecasts. To gauge usefulness, we looked at the uncertainty of machine learning forecasts and how well they captured spatial and temporal yield variability early in the season. Selected case studies included combinations of nine countries (Bulgaria (BG), Germany (DE), Spain (ES), France (FR), Hungary (HU), Italy (IT), the Netherlands (NL), Poland (PL), Romania (RO)) and six crops (soft wheat, spring barley, sunflower, grain maize, sugar beets, potatoes) (Fig. 1; Table A.1).

2.1. Theoretical framework

Machine learning models to forecast crop yields were built using regional data and compared with a linear trend model to gauge basic prediction skill (Fig. 2). We use a trend model to evaluate prediction skill because there are no official regional forecasts in Europe. The machine learning workflow was run using a *configuration* that controlled options, such as crop, country, forecast dekad (10-day period relative to harvest), crop calendar and prediction algorithms. Models trained using the workflow configuration of this paper were also compared with those from our previous work (Paudel et al., 2021; Section 2.2) to assess the impact of workflow updates. In addition, we evaluated the uncertainty of machine learning forecasts for cases that showed cancellation effects of regional differences. Such cases illustrate how national averages may look good without providing information about regional differences. Furthermore, we looked at the ability of machine learning models to capture spatial variability of crop yields for an average harvest and two extreme harvests. These cases highlighted the strengths and limitations of our machine learning models. The details of each evaluation step are provided in Section 2.5.

Regional machine learning forecasts were aggregated to the national level using crop area weights and compared with MCYFS forecasts to assess their added value. Although machine learning forecasts could be produced for intermediate spatial levels, that step was skipped because

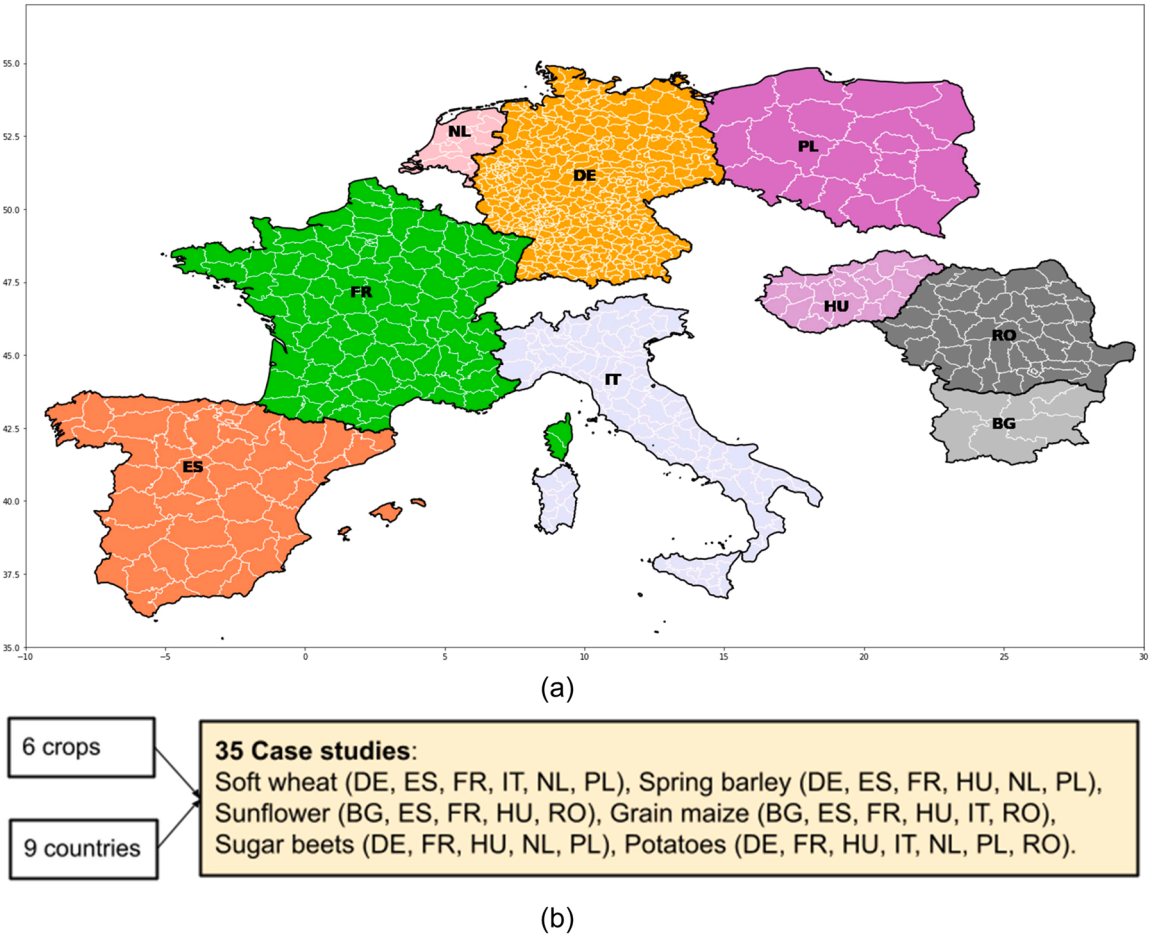

Fig. 1. (a) Selected countries and their NUTS regions. (b) Selected case studies. Case studies included nine major crop-growing countries of Europe for soft wheat, spring barley, sunflower, grain maize, sugar beets and potatoes.

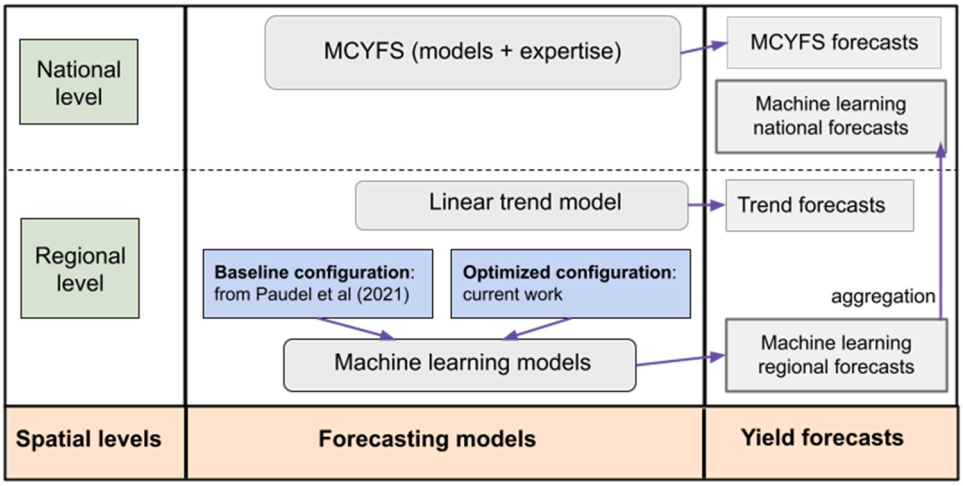

Fig. 2. Framework to evaluate the quality of machine learning forecasts. Regional data was used to build a linear trend model and two sets of machine learning models. Regional machine learning forecasts were first compared with linear trend model forecasts and later aggregated to national level to compare with MCYFS national forecasts. MCYFS forecasts were provided by the European Commission's Joint Research Centre (JRC).

these levels do not have official forecasts. The European Commission's Joint Research Centre (JRC) uses MCYFS to provide regular yield forecasts at national level. We wanted to find out whether machine learning and MCYFS performed similarly in the selected case studies or complemented each other. At national level, another focus was how well machine learning and MCYFS forecasts captured the temporal (or year-to-year) yield variability.

Crop yield forecasts were made early in the season and at harvest to

understand whether our forecasts provided useful information to support policy decisions. The workflow supports forecasts at dekadal (10-day) intervals. An *experiment* executed the workflow with the chosen configuration (i.e. crop, country and forecast dekad) and produced regional and national forecasts. For each crop and country (see Fig. 1*b*), experiments were run to make early season forecasts at 120, 90, 60 and 30 days before harvest and end of season forecasts at harvest. In this paper, we primarily report forecasts 60 days before harvest because results from other experiments do not significantly alter our observations or conclusions. All experiments were run with the baseline configuration (Section 2.2) and the optimized configuration (Section 2.3). Machine learning models were built per crop and country, i.e. data for all regions within a country (for the selected crop) were pooled to build a model for that country. The model predicted crop yields for all regions and years included.

2.2. Machine learning baseline

The machine learning baseline (Paudel et al., 2021) is a generic, modular and reusable workflow that combines agronomic principles of crop modeling with machine learning. The input data consist of crop model simulation outputs, weather observations, remote sensing indicators and soil water holding capacity. Regional yield statistics from the national statistics portals (e.g. NL-CBS, 2020) serve as the ground truth or labels for training and evaluating machine learning models. The crop calendar is inferred from crop model-simulated development stages (Table A.2) and used to design features that capture the impact of various indicators during different stages of crop development. The indicators selected for feature design are shown in Table A.3. On the machine learning side, the baseline uses grid search to find the optimal hyperparameters from a small set of values for four algorithms (see Section 2.5).

2.3. Improvements to the machine learning workflow

We updated the machine learning baseline from Paudel et al. (2021) by adding improvements to data preprocessing, feature design and machine learning steps. Here we briefly describe the improvements. Additional details about each improvement are included in Appendix A.

In preprocessing and feature design, we made four changes. *First*, we added data cleaning to preprocessing by identifying sequences of duplicate or missing yield values. An entire region was removed if it had long sequences (length >= 5) or multiple short sequences (length 2–4). In the case of one short sequence, only the data points were removed. *Second*, we used a dynamic crop calendar that varied by region and year. In contrast, the machine learning baseline used the same crop calendar for the whole country. *Third*, we designed features for extreme conditions to be less sparse. In the machine learning baseline, those features counted the number of days or dekads with values crossing certain thresholds and had many data points with zero values. We replaced them with the standard scores or z-scores based on the long-term average and standard deviation for the selected crop calendar period (see Table A.4). Z-score features were less sparse and also captured the magnitude of the extremes. *Fourth*, we added data to capture spatial differences in elevation, slope, field size, crop area and irrigated crop area (Table 1; Table A.5). In the baseline and the improved workflow, the yield trend is captured from yield values of five previous years to account for factors such as technological improvements.

On the machine learning side, we added three improvements. *First*, highly correlated features (correlation > 0.9) were dropped. We also removed feature selection methods based on mutual information (a univariate method) and unioning of features selected by other methods. *Second*, we added a robust hyperparameter search based on Bayesian optimization (Brochu et al., 2010; Shahriari et al., 2015). Bayesian optimization selects a new set of hyperparameter values by fitting an acquisition function to the results of hyperparameter settings tried before. *Third*, we made some changes to training, validation and test splits. We still started with a 70%−30% training and test split, but made a small change to the time-based 5-fold sliding validation used for feature selection and hyperparameter optimization. In the baseline, the training folds moved forward by one year when the validation fold moved. In the updated workflow, the training folds always started with the minimum training year to utilize all available training data (Fig. 3). For example, NL data was available from 1999 to 2018. The training data was from 1999 to 2011 and the test data from 2012 to 2018. During 5-fold sliding validation, the first iteration used 1999–2006 for training and 2007 for validation, the second iteration used 1999–2007 for training and 2008 for validation, the third iteration used 1999–2008 for training and 2009 for validation, and so on. During evaluation on the test set (the 30% in Fig. 3), we refitted a model for every test year using data up to the previous year, thus utilizing additional data available for training. For example, a model for soft wheat (NL) was trained with data up to 2011 and evaluated on the test year 2012 as is. For 2013, the model was refitted with data up to 2012. Similarly for 2014, the model was refitted using data up to 2013 and so on (Fig. 4). This approach is comparable to how operational systems such as MCYFS work.

We designed improvements to the machine learning baseline with emphasis on reusability and scalability of the approach to a large-scale system, such as MCYFS. In this paper, experiments for all case studies were run by combining the improvements described above. We call this the optimized configuration as opposed to the baseline configuration from Paudel et al. (2021). Our design includes configuration options to select a different combination based on expertise or validation set performance.

Table 1
Data sources summary.

Data	Indicators, Source
WOFOST crop model outputs	Water-limited dry weight biomass (kg ha^{-1}), water-limited dry weight storage organs (kg ha^{-1}), water-limited leaf area divided by surface area (m^2 m^{-2}), development stage (0–200), root-zone soil moisture as % of water holding capacity, sum of water limited transpiration (cm). **Source**: MCYFS. See Lecerf et al. (2019).
Meteo	Maximum, minimum, average daily air temperature (°C), sum of daily precipitation (PREC) (mm), sum of daily evapotranspiration of short vegetation (ET0) (Penman-Monteith, Allen et al., 1998) (mm)), sum of daily global incoming shortwave radiation (KJ m^{-2} d^{-1}), climate water balance = (PREC - ET0). ***Source***: MCYFS. See Lecerf et al. (2019).
Remote Sensing	Fraction of Absorbed Photosynthetically Active Radiation (Smoothed) (FAPAR). ***Source***: MCYFS. See Copernicus GLS (2020).
Crop Areas	Absolute crop areas (ha). Fraction of parent region's crop area. ***Source***: Eurostat (Eurostat, 2021a) and MCYFS (EC-JRC, 2021a).
Irrigated area	Irrigated total area and irrigated crop-specific area. ***Source***: EC-JRC (2021a).
Elevation, slope	Average and standard deviation. ***Source***: USGS-EROS (2021).
Soil	Soil water holding capacity. ***Source***: MCYFS. See Lecerf et al., (2019).
Field Size	Average and standard deviation. ***Source***: Lesiv et al (2019).
Yield	Yield at regional (NUTS2 or NUTS3) and national (NUTS0) level. ***Regional Source***: NL-CBS (2020), FR-Agreste (2020), DE-RegionalStatistiks (2020),Eurostat (2021a), EC-JRC (2021a). ***National Source:***Eurostat (2021a), EC-JRC (2021a).
MCYFS crop yield forecasts	Date and forecast value. ***Source:*** MCYFS. See Van der Velde and Nisini (2019).

Appendix A, Section A.4 provides additional details about the data sources.

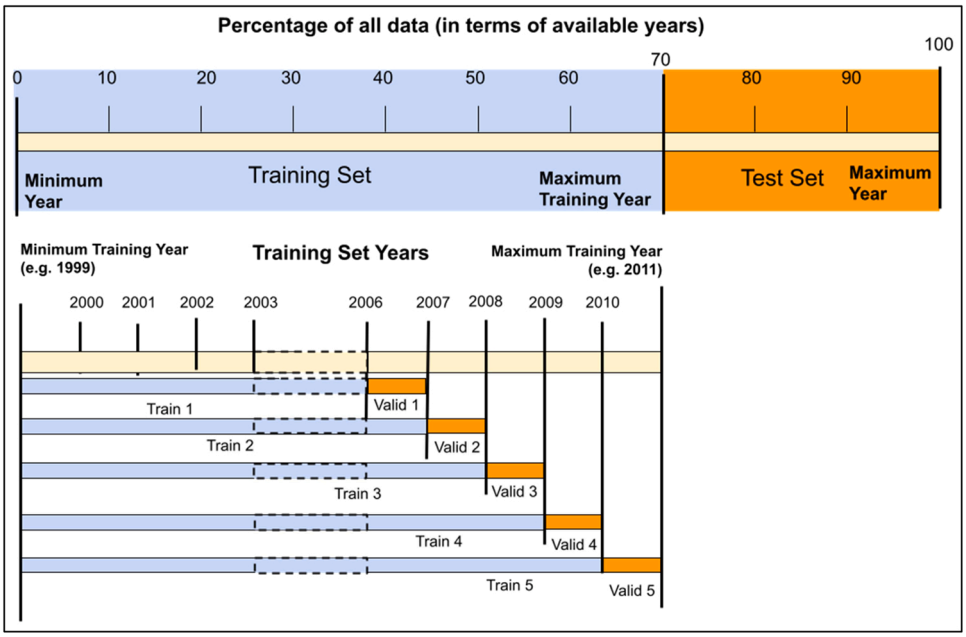

Fig. 3. Training, validation and test splits. Data for each region (for the selected crop and country) was split into training and test sets using time-based ordering of available years. The training data was further split using 5-fold sliding validation for feature selection and hyperparameter optimization (the lower panel). The dashed area is not drawn to scale.

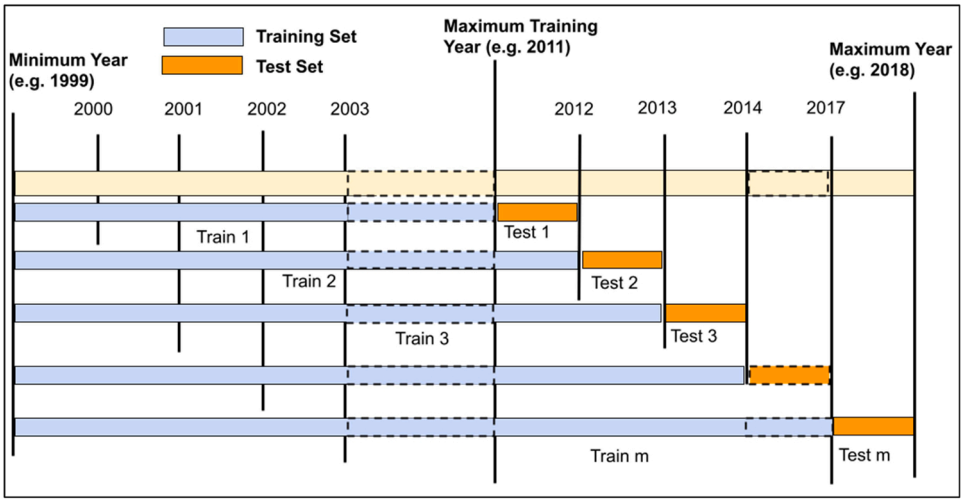

Fig. 4. Per test year model refitting. After 5-fold sliding validation to find optimal features and hyperparameters, we fitted a model on the entire training set. For every test year, we refitted this model on the training years and the previous test years, thus utilizing the additional data available. The dashed areas are not drawn to scale.

2.4. Data

Our data came from MCYFS (see EC-JRC (2021a); MARSWiki (2021); Appendix A.1) and Eurostat (Eurostat 2021a,b). We designed features from outputs of the WOFOST crop model (van Diepen et al., 1989; Supit et al., 1994; De Wit et al., 2019), weather observations, remote sensing indicators, soil, elevation, slope, crop area, irrigated crop area and average field sizes (see Table 1; Tables A.4, A.5). For labels or ground-truth data, we used yield statistics reported by the EU member states to Eurostat. Yield data was available at NUTS2 level for BG, NL and PL; and at NUTS3 level for DE, ES, FR, HU, IT and RO. Other data sources were aligned to the NUTS level of yield statistics to predict crop yields at that level. The length of the time series was determined by the availability of remote sensing and yield data. For most cases, we had data from 1999 to 2018.

2.5. Evaluation

We evaluated the quality and usefulness of machine learning forecasts using three steps (Table 2). *First*, we assessed the prediction skill and uncertainty of regional forecasts. Regional forecasts were compared with those of a per-region linear trend model that used a five-year window. *Second*, we analyzed the regional differences between reported and predicted yields for an average harvest and two extreme harvests. *Finally*, we aggregated regional predictions to the national level and compared them with the past MCYFS forecasts.

We assessed the performance of four machine learning algorithms: (i) Ridge Regression (Hoerl and Kennard, 1970), (ii) K-nearest Neighbors (KNN) Regression (Cover and Hart, 1967; Aha et al., 1991), (iii) Support Vector Machines Regression (SVR) (Boser et al., 1992; Cortes and Vapnik, 1995), and (iv) Gradient Boosted Decision Trees (GBDT)

Table 2
Summary of methods to evaluate the regional and national predictions.

Motivation	Method	Expected outcomes
1.1 Evaluate the impact of workflow improvements. 1.2 Assess prediction skill of machine learning at regional level. 1.3 Evaluate the overall uncertainty of regional forecasts. 1.4 Evaluate uncertainty of regional forecasts in cases where regional differences cancel out.	1.1 Compare the test set NRMSE, MAPE of the optimized models with the baseline. 1.2 Compare the test set NRMSE, MAPE of machine learning models with the trend model. 1.3 Box plots of prediction residuals for the optimized machine learning models and the trend model. 1.4 Compare the coefficient of variation for cases with low average and high standard deviation of trend residuals. Section: 2.5.1	1.1 Lower errors for optimized models show added value of workflow improvements. Higher errors indicate improvements did not help and likely led to overfitting. 1.2 Lower errors for machine learning models compared to trend models show prediction skill. 1.3 Lower variance and smaller number of outliers for machine learning prediction residuals indicate low uncertainty. 1.4 Lower coefficient of variation would indicate lower uncertainty. Section: 3.1
2 Evaluate how well predictions capture spatial yield variability for average and extreme harvests.	2 Divide the reported and predicted yields into 5 classes. Compare the yield classes in a confusion matrix. Assess the spatial distribution of regions with yield class mismatch. Section: 2.5.2	2 A large percentage of matching yield classes would show better prediction results. Section: 3.2
3.1 Assess the prediction skill of machine learning at national level. 3.2 Assess how well national forecasts capture temporal variability.	3.1 Compare NRMSE and MAPE for machine learning predictions with MCYFS forecasts. 3.2 Compare temporal variation of reported vs predicted yields for machine learning and MCYFS. Section: 2.5.3	3.1 Lower errors for machine learning models compared to MCYFS show the improved prediction skill. 3.2 Similarity between reported and predicted time series shows reliability of predictions. Section: 3.3

Rows 1 and 2 evaluate regional forecasts; Row 3 evaluates the aggregated national forecasts.

Regression (see Friedman, 2001; Hastie et al., 2009). These algorithms represent four ways to learn the relationships between predictors and yield. Ridge Regression can capture linear relationships only. The other three algorithms can learn nonlinear relationships in different ways. KNN makes predictions based on similarities between instances in feature space. SVR can model both linear and non-linear relationships. It maps nonlinear data to a higher dimensional space using kernel functions to capture complex relationships. GBDT is an ensemble method (similar to Random Forests (Breiman, 2001)) that relies on gradient boosting (Friedman, 2001) to grow decision trees, and is often more accurate than Random Forests (see Hastie et al., 2009). Overall, the four algorithms we selected represent four important families of machine learning algorithms.

Model performance was compared using the mean absolute percentage error (MAPE), normalized root mean squared error (NRMSE) and the coefficient of determination or R^2. The normalized RMSE was defined to be RMSE divided by the mean yield of the test set. Significance of model performance was evaluated using the Wilcoxon signed-rank test, which is a standard non-parametric method to compare models across different datasets or case studies (Demsar, 2006; Kadra et al., 2021).

2.5.1. Prediction skill and uncertainty of regional forecasts

To understand the impact of workflow improvements, the test set NRMSE and MAPE of the optimized models were compared with the baseline. Similarly, to evaluate the prediction skill, NRMSE and MAPE of machine learning models were compared with a per-region linear trend model. Wilcoxon p-value was used to evaluate the statistical significance of NRMSE differences between machine learning models and the trend model. Because the test set contained all regions of a country for many test years, metrics like NRMSE and MAPE provide a high level estimate of uncertainty. To get more information about variance and outliers, we created boxplots of the prediction residuals (100% x (predicted yield - reported yield)/reported yield) 60 days before harvest. To emphasize spatial variability of yields and the interaction of regional differences, we identified test years in which trend prediction residuals had a low average (<=10%), but high standard deviation (>= 25%). Such instances showed the compensating effect of yield overestimations in some regions canceling out yield underestimations in others. For these instances, we counted the number of cases in which the machine learning model had a lower coefficient of variation (i.e., standard deviation / mean) than the trend model. A lower coefficient of variation would imply lower uncertainty and higher reliability.

2.5.2. Regional differences in average and extreme years

To assess how well machine learning forecasts capture spatial variability, we compared them with reported yields for one average harvest and two extreme harvests from the test set. Potatoes (2013), an average harvest, was selected based on the previous five-year average (see MARS bulletin for 2013 Vol 21 No. 10, EC-JRC (2021b)). Grain maize (2015) was selected because of high yield losses in Central Europe (see MARS bulletin for 2015 Vol 23 No. 9, EC-JRC (2021b)). Similarly, soft wheat (2016) was selected because of well-known yield losses in north-central France (see Ben-Ari et al., 2018). For these cases, reported and predicted yields were divided into 5 classes (very low (0–20%), low (20–40%), medium (40–60%), high (60–80%), very high (80–100%)) covering 20% intervals between minimum and maximum yields for each country. We decided to compare yield classes instead of reported and forecasted values because the ranges of yield values varied across countries. Per-country yield classes provided similar meaning (very low, low, etc.) while still highlighting country-specific differences in yield values. Agreement between reported and predicted yield classes was quantified using a confusion matrix. In addition, a qualitative evaluation was performed on the spatial distribution of mismatches.

2.5.3. Quality of national forecasts

We evaluated prediction skill of machine learning and past MCYFS forecasts at national level by calculating the NRMSE and MAPE using Eurostat national yields as the ground truth. Wilcoxon p-value was used to evaluate the statistical significance of NRMSE differences between machine learning models and the MCYFS. Regional forecasts were aggregated to successive NUTS levels and to national level based on modeled crop area weights (Cerrani and López Lozano, 2017). In addition, we plotted the time series of machine learning forecasts and MCYFS forecasts together with reported yields to see how well they capture the temporal variability during test years.

2.6. Implementation

We used Apache Spark (Zaharia et al., 2016) for data preprocessing and feature design, and the scikit-learn python package (Pedregosa et al., 2011) for machine learning. Bayesian optimization for hyper-parameter search was based on the scikit-optimize package (Scikit-optimize Contributors, 2021). Our implementation is available through the pypi repository (pypi.org) as *cypml* pkg. Version 1.0.* include the machine learning baseline; version 1.1.* include the improvements made to the baseline as modular options that can be turned on or off; and version 1.2.* replace grid-search with Bayesian

optimization to find optimal hyperparameters.

3. Results

We executed the same workflow for all thirty-five case studies. This reusability made regional crop yield forecasting scalable to all major crop growing countries of Europe. In the following analysis we primarily focus on forecasts 60 days before harvest. In terms of metrics, we report normalized RMSE here. MAPE and R^2 scores are included in the supplementary results (Appendix A.5, Appendix B).

3.1. Prediction skill and uncertainty of regional forecasts

In general, workflow updates improved the performance of machine learning. The optimized models had a lower normalized RMSE than the baseline in 25 out of 35 cases (~71%) for 60 days before harvest and 22 out of 35 cases (63%) for end of season (Fig. A.4). The median NRMSEs for 60 days early were 17.27% (baseline) and 16.57% (optimized), and those for the end of season were 21.67% (baseline) and 15.88% (optimized). The corresponding Wilcoxon p-values for 60 days early and end of season were 1e-3 and 2e-7 respectively, indicating significant performance improvement with the optimized configuration. Both the optimized and baseline models showed prediction skill as early as 120 days before harvest. The optimized machine learning models had a lower normalized RMSE than the trend model for all 35 cases 120 days before harvest (Table A.6), for all except potatoes (HU) (Fig. 5, Table A.6) 60 days before harvest, and for all cases at the end of season (Table A.6, Fig. A.4). The median NRMSE for the trend model was 20.35% and the Wilcoxon p-values were 2e-7 (120 days early), 3e-7 (60 days early) and 2e-7 (end of season), indicating that the optimized machine learning models were significantly better. The boxplots of prediction residuals for 60 days before harvest showed that machine learning prediction residuals had lower variance and fewer outliers than trend residuals (Fig. 6, Fig. A.6).

For test years in which yield trend residuals had low average (<= 10%) but high standard deviation (> 25%), machine learning had a lower CV in 10 out 11 instances (Table 3; Table A.7). The low CV of machine learning provided confidence on the quality of regional

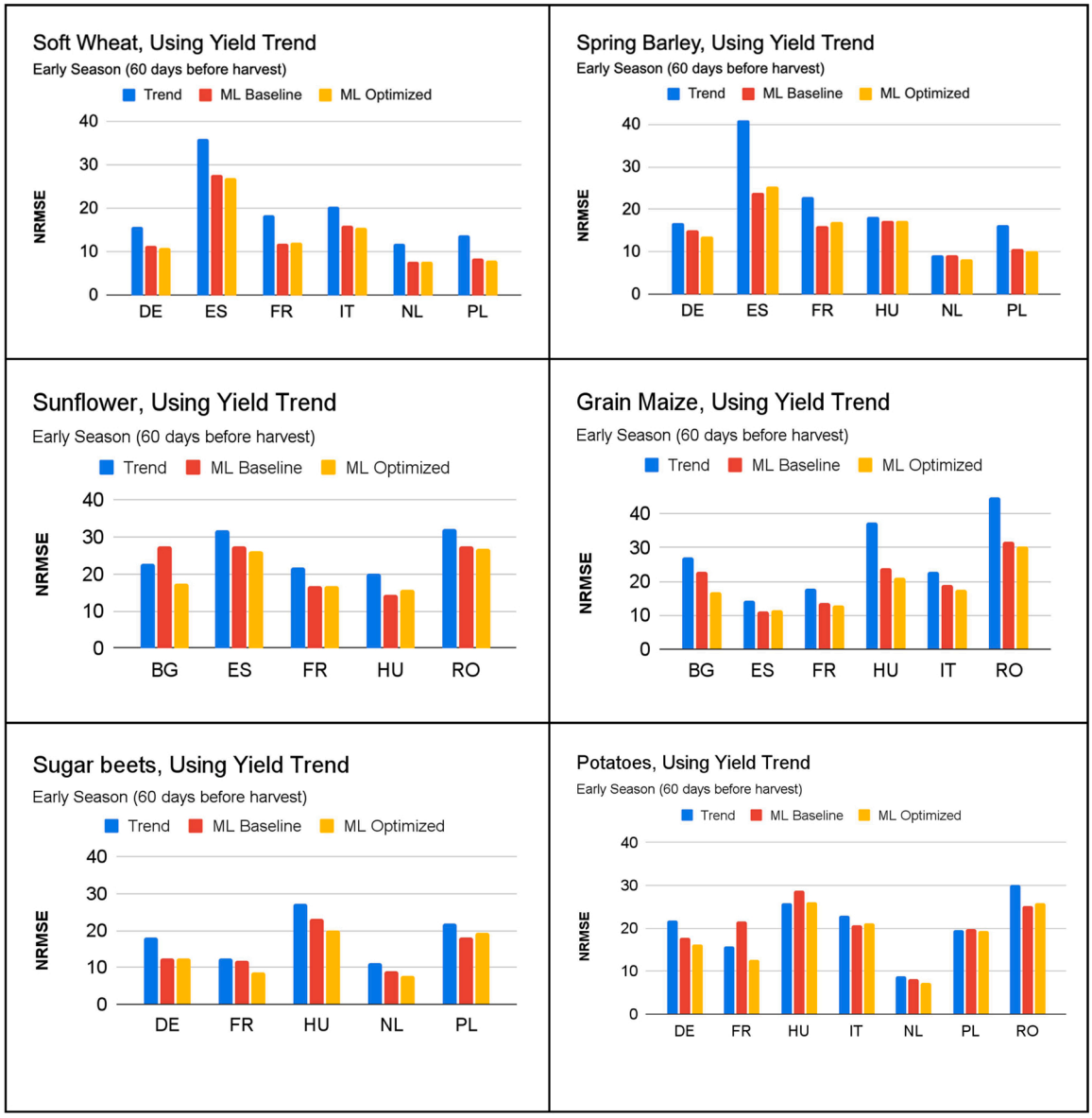

Fig. 5. Normalized RMSE of regional forecasts 60 days before harvest. Regional forecasts from machine learning (baseline and optimized) were compared with the trend model. For machine learning models, we show results for the algorithm with the lowest NRMSE.

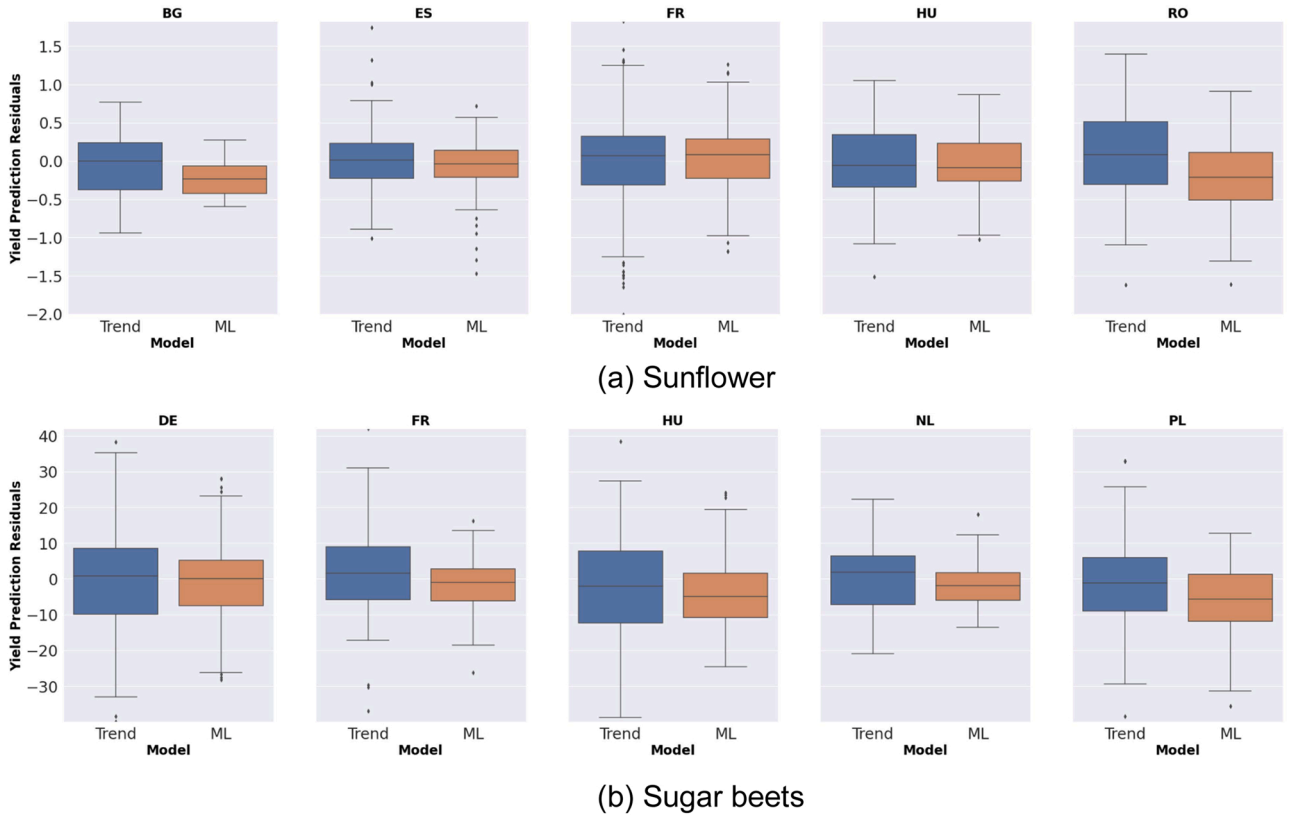

Fig. 6. Boxplots of regional yield residuals 60 days before harvest. The trend model (blue) has a higher variance than machine learning (orange). For machine learning, we show results for the algorithm with the lowest NRMSE. Fig. A.6 shows boxplots for all other crops.

Table 3
Coefficient of variation for regional prediction residuals 60 days before harvest.

Crop	Country	Test Year	Trend CV (%)	Machine learning CV (%)
Soft wheat	ES	2015	122.55	2.41
Spring barley	ES	2015	82.46	2.99
Sunflower	ES	2011	9.54	4.55
Sunflower	ES	2015	3.43	9.54
Grain maize	ES	2012	28.49	4.50
Grain maize	IT	2009	5.65	3.28
Sugar beets	HU	2010	14.09	4.26
Potatoes	DE	2016	13.94	6.02
Potatoes	IT	2012	4.59	4.28
Potatoes	IT	2013	22.68	5.31
Potatoes	IT	2014	10.34	5.21

For instances where yield trend residuals had a low average but high variance, machine learning prediction residuals almost always had a lower coefficient of variation, indicating lower uncertainty.

forecasts. In these years, national forecasts would fail to capture the regional differences due to compensating and averaging effects of residuals. We observed this compensating effect of regional yield residuals for soft wheat in Spain (in addition to 2015, shown in Table 3). Soft wheat in Spain had very high residuals at regional level and very low errors at national level. The national level MAPE and NRMSE for the optimized machine learning model were less than 5% (see Table A.8, Figs. A.10, A.11). Similarly, machine learning and MCYFS forecasts at national level follow the reported yields quite closely (Fig. 8a). However, absolute prediction residuals for many regions were high (25–50%, orange) and very high (>50%, red) in 2012, 2014, 2015 and 2016 (Fig. A.8). Such disparity between regional and national level shows the limitations of national averages and the added value of regional forecasts.

3.2. Regional differences in average and extreme years

Confusion matrices for potatoes (2013), grain maize (2015) and soft wheat (2016) showed that machine learning forecasts matched well with reported yields for an average harvest, but not so well for extreme harvests. For potatoes in 2013, considered an average harvest, predicted yield classes matched reported yield classes for ~71% of regions and the rest were mostly off by one (~28%) (Fig. 7a, Fig. A.9a). For grain maize in 2015, when there were significant yield losses in Central Europe, machine learning predicted matching yield classes for ~52% of the regions and had many mismatches (off by one: ~41%; off by 2: ~7%) (Fig. 7b, Fig. A.9b). There were fewer mismatches in ES, where close to 80% of grain maize area is irrigated (Eurostat, 2016a). On the other hand, FR had many mismatches in the north-east, where irrigation percentages are lower (see van der Velde et al., 2010). Finally, for soft wheat in 2016, machine learning predicted yield classes matched reported yield classes in ~53% of the cases and again had a large number of mismatches (off by one: ~41%; off by 2: ~5%) (Fig. 7c, Fig. A.9c). DE, with its small NUTS3 regions, had the maximum number of off-by-one mismatches, but FR had a large number of more extreme mismatches (off by 2 or more). For FR, most of the mismatches were in the north (Fig. 7c).

3.3. Quality of national forecasts

Machine learning predictions aggregated to the national level were in general comparable to the past MCYFS forecasts. For 120 days before harvest, one of the machine learning configurations (baseline or optimized) had a lower normalized RMSE than MCYFS for 25 out of 35 cases (Table A.8). The same was true for 22 out of 35 cases 60 days before harvest. The median NRMSEs for 60 days early were 8.81% for MCYFS, 8.54% for the baseline and 8.41% for the optimized models. The Wilcoxon p-values for the machine learning models 60 days early were 0.64

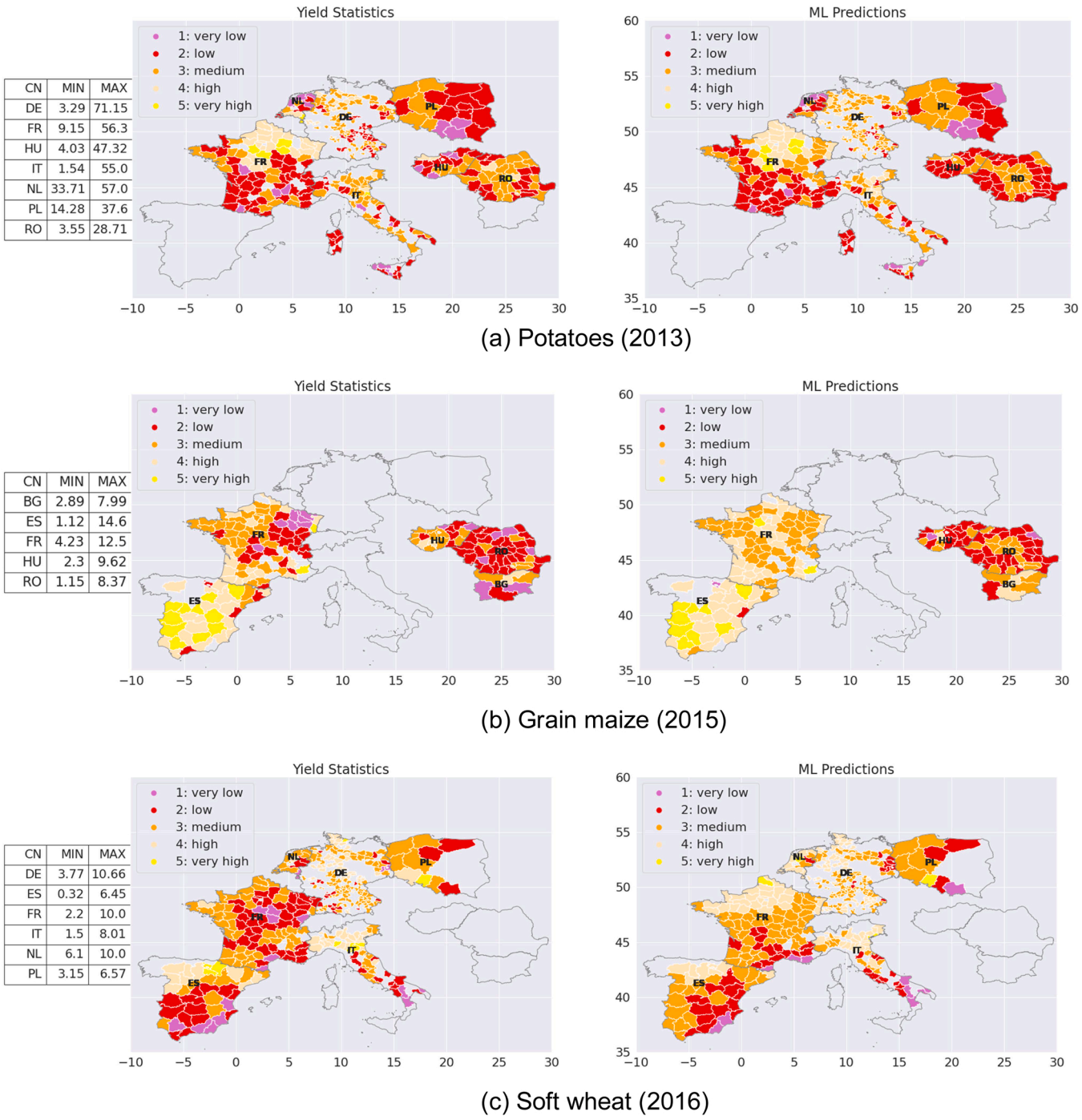

Fig. 7. Regional forecasts 60 days before harvest vs reported yields. (a) 2013 - an average harvest for potatoes. (b) 2015 - an extreme harvest for grain maize. (c) 2016 - an extreme harvest for soft wheat (mainly in the north of FR). The machine learning models and yield classes are per country. Very low: up to 20% above the min yield; Low: 20–40%; Medium: 40–60%; High: 60–80%; Very high > 80%.

and 0.95, indicating no significant differences compared to MCYFS. Although their overall performance was similar, machine learning and MCYFS had lower NRMSEs for different case studies. For example, machine learning had significantly lower NRMSEs for soft wheat (ES) (4.15 vs 10.44), spring barley (ES) (9.83 vs 17.8) and spring barley (PL) (4.73 vs 11.77). Similarly, MCYFS performed significantly better for sunflower (BG) (5.16 vs 16.18) and sunflower (RO) (13.22 vs 24.34). These examples show potential benefits of combining the expertise-driven approach of MCYFS with the data-driven approach of machine learning. At the end of season, normalized RMSE for machine learning were lower than MCYFS for 13 out of 35 cases (Fig. A.10). The median NRMSEs for end of season were 6.74% for MCYFS, 8.18% for the baseline and 7.49% for the optimized models. The corresponding Wilcoxon p-values for machine learning models were 3e-4 and 1e-3, indicating that MCYFS had significantly better performance. Evidently, MCYFS forecasts improve as the season progresses. This is expected since MCYFS analysts update the forecasts using expertise as well as information from additional sources such as farmer magazines and news reports (López-Lozano and Baruth, 2019).

Machine learning and MCYFS captured the year-to-year variability of national crop yields in some cases (e.g. soft wheat (ES), soft wheat (PL), grain maize (HU)), but not others (e.g. soft wheat (DE), spring barley

(NL), sunflower (FR), sugar beets (HU) (Fig. 8a,b; Fig. A.12a, b). Machine learning followed the reported yield better than MCYFS in certain cases, such as soft wheat (NL), spring barley (ES) and spring barley (PL). Similarly, MCYFS captured the variability better than ML in others, such as sunflower (BG), sunflower (RO), grain maize (FR) and potatoes (FR) (Fig. 8c,d; Fig. A.12c, d). Overall, we could see the added value of machine learning in some cases and its limitations in others.

4. Discussion

Crop yield forecasts at higher spatial resolutions provide additional information about yield variability not present in national forecasts. In Europe, official crop yield forecasts are available only at the national level (Van der Velde and Nisini, 2019; Lecerf et al., 2019). The MARS (Monitoring Agricultural Resources) unit of European Commission's Joint Research Centre publishes agro-meteorological analyses, areas of concerns and the outlooks for crop yields in the MARS bulletins (van der Velde et al., 2019; Seguini et al., 2019). However, there are no official regional forecasts and very few studies have attempted to predict regional crop yields in Europe (e.g. Pagani et al., 2019; Bussay et al., 2015). We attempted to fill this gap by building a generic machine learning workflow that scales to different crops and countries, with very little extra time and effort. With our workflow, systems such as MCYFS could use machine learning for cases where it typically performs well and switch to expertise-based methods for others. We found cases (for example, soft wheat (ES) and spring barley (PL); see Table A.8) in which machine learning performs significantly better than MCYFS early in the season. Our results indicate that large-scale regional crop yield forecasting is a viable goal and machine learning can help with scaling the task as well as producing reliable forecasts at both regional and national levels. Overall, access to regional forecasts would provide additional information to explain national and provincial yields based on constituent regions and to design targeted agricultural policies.

In this paper, we improved and optimized the machine learning baseline from Paudel et al. (2021), both in terms of scaling and prediction skill. The optimized configuration had better normalized RMSEs for 60 days before harvest than the baseline according to the Wilcoxon signed rank test. Even then, the median NRMSEs were only marginally better (17.27% for baseline vs. 16.57% for optimized). Despite small improvements in forecast errors, our workflow updates have practical significance. For example, data cleaning is a standard preprocessing step; dynamic calendars (i.e. per-region, per year calendars) provide more accurate growing season information; and Bayesian optimization is more robust than grid-search. In this work, we used the same configuration options for all case studies to keep the experiment setup simple and generic. We expect case study-specific configuration options and optimizations to help when paired with contextual knowledge of, for example, how many models to build per country, how to group

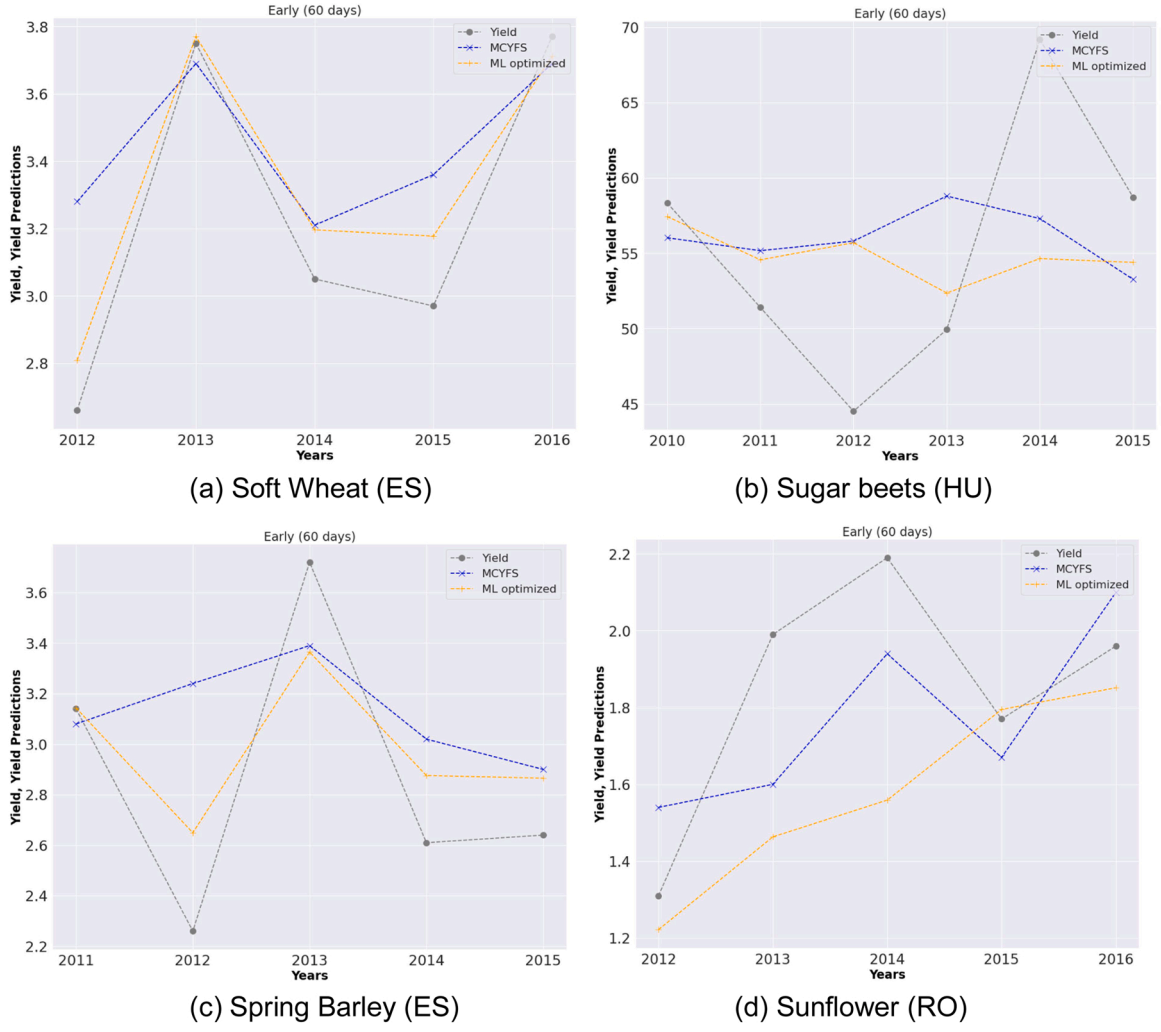

Fig. 8. National forecasts 60 days before harvest compared with reported yields. For machine learning, we selected the algorithm with the lowest NRMSE. (a) Both machine learning and MCYFS capture the temporal variability. (b) Both do not capture temporal variability. (c) Machine learning performs better than MCYFS. (d) MCYFS outperforms machine learning.

regions, and what yield trend window to use for the selected regions. The analyst-driven approach used by MCYFS will provide an ideal setting for crop and country specific choices and optimizations. Similarly, we did not delve into explaining machine learning predictions even though the workflow produces feature importance that can provide some explainability. Feature importance and explainability would also be useful when selecting and analyzing case study-specific configurations or optimizations.

Our per-country models based on regional data have room for improvement in capturing spatial and temporal variability. We pooled data from possibly very different regions to have a large enough data size for machine learning. Machine learning requires a sufficiently long time series to split the data into training, validation and test sets. For example, if we were to use 30% of the data for testing and 5-fold sliding validation for model selection, we would need at least 15 years of data. Due to regional differences in data size and agro-climatic variables, there were cases among the 35 crop-country combinations in which machine learning struggled to learn meaningful relationships. Comparison of predicted and reported yields showed that machine learning forecasts captured regional differences for average harvests but not so well for extreme harvests. Boxplots of prediction residuals also indicated that machine learning forecasts were quite conservative and stayed close to the trend or the average (Fig. 6, A.6). We attempted to capture weather extremes using z-score features, but they were not always effective (for example, Grain Maize (2015); Fig. 7b). Similarly, our input data did not account for yield extremes related to diseases, pests or farm management practices. Nevertheless, machine learning forecasts showed lower uncertainty than trend forecasts and comparable performance with MCYFS forecasts early in the season.

From cases with low agreement between forecasted and reported yields, we extracted insights about data quality and potential overfitting. Spring barley (FR) and sunflower (FR) had near identical reported yields for many data points (Fig. A.7). Although forecasts errors are quite low for these cases, concerns remain about reliability of the data. In cases where the baseline outperformed the optimized model (e.g. spring barley (ES), sunflower (HU), grain maize (ES)), we found lower validation set errors and higher test set errors. Such instances indicate overfitting or large differences between validation and test set distributions. Our workflow does include safeguards against overfitting, such as 5-fold sliding validation. Because all optimizations rely on validation set performance to select the optimal configuration (e.g. hyperparameters, configuration options), they can still lead to overfitting.

We identified five areas that could help improve regional crop yield forecasting going forward. *First*, the reported yield statistics would have to be more reliable. Forecasting models work with the assumption that reported yield statistics are objective ground-truths that are consistent across space and time (Van der Velde and Nisini, 2019). The collection and curation protocols for these statistics vary across countries (López-Lozano et al., 2015). Standard data collection and curation protocols would help improve their quality. *Second*, machine learning or statistical models can only learn relationships between predictors and yield that are present in the data. The input data used to create features does not capture all factors contributing to yield variability. For example, signals from meteorological variables may not always be spatially and temporally coherent (Lecerf et al., 2019). In addition, remote sensing features were not crop-specific and there were no features to account for farm management practices. Data sources that capture additional factors contributing to yield variability would be useful, especially when they are consistent (from the same or similar sources) and complete (matching the time series of other data sources). *Third*, machine learning takes advantage of larger data sizes at regional level. However, machine learning models trained on data from widely different regions have to learn spatial and temporal yield variability simultaneously. Such models will struggle when relationships between predictors and yields are different for different regions. We believe grouping regions based on agro-climatic similarities would help and defer this for future work. Similarly, per region models could be built when regional time-series are sufficiently long. *Fourth*, crop yield prediction at NUTS2 or NUTS3 still has to deal with errors associated with aggregation of predictors from smaller spatial units. Reliable crop areas and aggregation methods play an important role in reducing such errors. High-resolution remote sensing data could provide a more accurate way to estimate crop areas in the future. However, it will take some time to produce a consistent and long time series of reliable crop areas. *Finally*, we have not delved into outliers detection in this paper. A systematic approach to identify outliers and to impute missing or outlier data points would improve the data at regional level. Unsupervised machine learning methods (e.g. clustering) would prove helpful in outliers detection.

In an ideal setting, we would have measurements, statistics and crop yield forecasts from farm level all the way up to national and global levels. There are studies that have combined remote sensing data with crop modeling and statistical methods to predict farm-level crop yields (e.g., Lobell et al., 2015; Zhao et al., 2020; Deines et al., 2021). However, there is not enough public data and long time series to build large-scale farm or field level models. In this paper, we focused on regional forecasts at NUTS2 or NUTS3 level. These regional forecasts were aggregated to national level for comparison with MCYFS forecasts. The same approach can be used to get forecasts for intermediate NUTS levels (e.g. NUTS2 and NUTS1 in FR, NUTS1 in NL). Our work will motivate crop yield forecasting at higher spatial resolutions and the adoption of a consistent forecasting method across multiple spatial levels.

5. Conclusions

We highlighted two main limitations of national level crop yield forecasts that motivate the need for regional crop yield forecasting. *First*, the aggregation of predictors from small spatial units to larger ones accumulates errors associated with crop areas and interpolation methods. *Second*, national level yield forecasts often hide regional differences, especially when they cancel out each other. Regional crop yield forecasts limit the aggregation errors and provide information about spatial variability. At the same time, regional forecasts can be aggregated to produce national forecasts. We showed that machine learning can take advantage of larger data sizes at regional level and provide a scalable way to produce regional forecasts. Based on our evaluation, machine learning forecasts had lower uncertainty than a trend model. These forecasts aligned quite well with reported yields for an average harvest, but less so for extreme harvests. Similarly, machine learning forecasts aggregated to national level compared well with past MCYFS forecasts, especially early in the season. Machine learning models did not perform significantly better than MCYFS at national level, but provided insights about uncertainty of regional forecasts and spatial variability of yields. Our work motivates the adoption of a consistent crop yield forecasting method across multiple spatial levels based on regional forecasts. Machine learning could be a tool to help make that transition.

Data and software availability

Sample data for the Netherlands are available at DOI: https://doi.org/10.5281/zenodo.5561113 courtesy of the European Commission's Joint Research Centre (JRC). The software implementation is available at: https://github.com/BigDataWUR/MLforCropYieldForecasting. The main branch has the baseline implementation and the mlopt branch has the optimized implementation.

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.

Acknowledgements

This work was partially supported by the European Union's Horizon 2020, EU 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 MCYFS data and to provide open access to MCYFS data for the Netherlands. Similarly, we would like to thank M. van der Velde, L. Nisini and I. Cerrani from JRC for sharing with us past MCYFS forecasts, Eurostat regional and national yield statistics and crop areas. We would also like to thank Hiske Overweg from the Geo-Information and Remote Sensing Group of Wageningen University and Research for help on using Bayesian optimization.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fcr.2021.108377.

References

Aha, D.W., Kibler, D., Albert, M.K., 1991. Instance-based learning algorithms. Mach. Learn. 6, 37–66. https://doi.org/10.1007/BF00153759.

Allen, R.G., Pereira, L.S., Raes, D., Smith, M., et al., 1998. Crop evapotranspiration – guidelines for computing crop water requirements. In: Irrigation and Drainage. FAO, Rome.

Bartholome, E., Belward, A.S., 2005. GLC2000: a new approach to global land cover mapping from Earth observation data. Int. J. Remote Sens. 26, 1959–1977. https://doi.org/10.1080/01431160412331291297.

Ben-Ari, T., Boé, J., Ciais, P., Lecerf, R., van der Velde, M., Makowski, D., 2018. Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. Nat. Commun. 9, 1–10. https://doi.org/10.1038/s41467-018-04087-x.

Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM New York, NY, USA, pp. 144–152.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brochu, E., Cora, V.M., De Freitas, N., 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599 doi:arXiv:1012.2599[cs.LG].

Bussay, A., van der Velde, M., Fumagalli, D., Seguini, L., 2015. Improving operational maize yield forecasting in Hungary. Agric. Syst. 141, 94–106. https://doi.org/10.1016/j.agsy.2015.10.001.

Buttner, G., Feranec, J., Jaffrain, G., Mari, L., Maucha, G., Soukup, T., 2004. The corine land cover 2000 project. EARSeL eProc. 3, 331–346. In: ⟨http://eproceedings.uni-oldenburg.de/website/vol03_3/03_3_buttner2.pdf⟩. Last accessed: May 18, 2021.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., As- Seng, S., Zhang, Y., You, L., et al., 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.

Ceglar, A., Toreti, A., Lecerf, R., Van der Velde, M., Dentener, F., 2016. Impact of meteorological drivers on regional inter-annual crop yield variability in France. Agric. For. Meteorol. 216, 58–67. https://doi.org/10.1016/j.agrformet.2015.10.004.

Cerrani, I., López Lozano, R., 2017. Algorithm for the disaggregation of crop area statistics in the MARS crop yield forecasting system. ⟨https://agri4cast.jrc.ec.europa.eu/DataPortal/Resource_Files/PDF_Documents/31_rationale.pdf⟩, (Accessed 8 October 2020).

Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., Warren, R., Qian, B., Daneshfar, B., Bedard, F., et al., 2015. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agric. For. Meteorol. 206, 137–150. https://doi.org/10.1016/j.agrformet.2015.03.007.

Copernicus GLS, 2020. Fraction of Absorbed Photosynthetically Active Radiation. ⟨https://land.copernicus.eu/global/products/fapar⟩, Last accessed: Oct 19, 2020.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/BF00994018.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13, 21–27. https://doi.org/10.1109/TIT.1967.1053964.

Crane-Droesch, A., 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ. Res. Lett. 13, 114003 https://doi.org/10.1088/1748-9326/aae159.

DE-RegionalStatistiks, 2020. Regionaldatenbank deutschland. ⟨https://www.regionalstatistik.de/genesis/online/data⟩, Last accessed: May 11, 2020.

De Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., Supit, I., van der Wijngaart, R., van Diepen, K., 2019. 25 years of the WOFOST cropping systems model. Agric. Systems 168, 154–167. https://doi.org/10.1016/j.agsy.2018.06.018.

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 Sen. Environ. 253, 112174 https://doi.org/10.1016/j.rse.2020.112174.

Demsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30. ⟨https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf⟩. Last accessed: Sept 20, 2021.

EC-JRC, 2021a. JRC Agri4Cast Data Portal. ⟨https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx⟩, (Accessed 11 May 2021).

EC-JRC, 2021b. JRC MARS Bulletins. ⟨https://ec.europa.eu/jrc/en/mars/bulletins⟩, (Accessed 11 May 2021).

Eurostat, 2016a. Irrigated area of semi-intensive crops, updated 2016. ⟨https://ec.europa.eu/eurostat/statistics-explained/images/f/f7/Irrigated_area_of_semi-intensive_crops_%28maize_and_cereals_excluding_maize_and_rice%29%2C_2010_%28%25_of_total_area_of_each_crop%29.png⟩.(Accessed 16 June 2021).

Eurostat, 2016b. Nomenclature of territorial units for statistics. ⟨https://ec.europa.eu/eurostat/web/nuts/background⟩, Last accessed: May 11, 2020.

Eurostat, 2021a. Eurostat - Agricultural Production - crops. ⟨https://ec.europa.eu/eurostat/statistics-explained/index.php/Agricultural_production_-_crops⟩, (Accessed 11 May 2021).

Eurostat, 2021b. Eurostat database. ⟨https://ec.europa.eu/eurostat/web/agriculture/data/database⟩, (Accessed 28 April 2021).

FR-Agreste, 2020. Agreste web data portal. ⟨https://agreste.agriculture.gouv.fr/agreste-web/⟩,(Accessed 11 May 2020).

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232. ⟨https://www.jstor.org/stable/2699986⟩. Last accessed: May 11, 2020.

García-León, D., López-Lozano, R., Toreti, A., Zampieri, M., 2020. Local-scale cereal yield forecasting in Italy: lessons from different statistical models and spatial aggregations. Agronomy 10, 809. https://doi.org/10.3390/agronomy10060809.

Gouache, D., Bouchon, A.S., Jouanneau, E., Le Bris, X., 2015. Agrometeorological analysis and prediction of wheat yield at the departmental level in France. Agric. For. Meteorol. 209, 1–10. https://doi.org/10.1016/j.agrformet.2015.04.027.

Han, J., Zhang, Z., Cao, J., Luo, Y., Zhang, L., Li, Z., Zhang, J., 2020. Prediction of winter wheat yield based on multi-source data and machine learning in China. Remote Sens. 12, 236. https://doi.org/10.3390/rs12020236.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67. https://doi.org/10.1080/00401706.1970.10488634.

Kadra, A., Lindauer, M., Hutter, F., Grabocka, J., 2021. Regularization is all you need: Simple neural nets can excel on tabular data. arXiv preprint arXiv:2106.11189 ⟨https://arxiv.org/pdf/2106.11189.pdf⟩, Last accessed: Sept 20, 2021.

Lecerf, R., Ceglar, A., Lopez-Lozano, R., Van Der Velde, M., Baruth, B., 2019. Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe. Agric. Syst. 168, 191–202. https://doi.org/10.1016/j.agsy.2018.03.002.

Lesiv, M., Laso Bayas, J.C., See, L., Duerauer, M., Dahlia, D., Durando, N., Hazarika, R., Kumar Sahariah, P., Vakolyuk, M., Blyshchyk, V., et al., 2019. Estimating the global distribution of field size using crowdsourcing. Glob. Ch. Biol. 25, 174–186. https://doi.org/10.1111/gcb.14492.

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.

López-Lozano, R., Baruth, B., 2019. Anevaluationframeworktobuilda cost-efficient crop monitoring system. experiences from the extension of the european crop monitoring system. Agric. Syst. 168, 231–246. https://doi.org/10.1016/j.agsy.2018.04.002.

López-Lozano, R., Duveiller, G., Seguini, L., Meroni, M., Garcıa-Condado, S., Hooker, J., Leo, O., Baruth, B., 2015. Towards regional grain yield forecasting with 1 km-resolution EO biophysical products: strengths and limitations at pan-European level. Agric. For. Meteorol. 206, 12–32. https://doi.org/10.3390/s18082674.

MARSWiki, 2021. MARS Crop Yield Forecasting System. ⟨https://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Welcome_to_WikiMCYFS⟩, (Accessed 11 May 2021).

NL-CBS, 2020. CBS Open Data Portal. ⟨https://opendata.cbs.nl/statline/%23/CBS/nl/?fromstatweb⟩, (Accessed 11 May 2020).

Pagani, V., Guarneri, T., Busetto, L., Ranghetti, L., Boschetti, M., Movedi, E., Campos-Taberner, M., Garcia-Haro, F.J., Katsantonis, D., Stavrakoudis, D., et al., 2019. A high-resolution, integrated system for rice yield forecasting at district level. Agric. Syst. 168, 181–190. https://doi.org/10.1016/j.agsy.2018.05.007.

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. Agric. Syst. 187, 103016 https://doi.org/10.1016/j.agsy.2020.103016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit- learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Porwollik, V., Müller, C., Elliott, J., Chryssanthacopoulos, J., Iizumi, T., Ray, D.K., Ruane, A.C., Arneth, A., Balkovič, J., Ciais, P., et al., 2017. Spatial and temporal uncertainty of crop yield aggregations. Eur. J. Agron. 88, 10–21.

Scikit-optimize Contributors, 2021. Scikit-optimize: Sequential model-based optimization. ⟨https://scikit-optimize.github.io/stable/getting_started.html⟩, (Accessed 20 September 2021).

Seguini, L., Bussay, A., Baruth, B., 2019. From extreme weather to impacts: the role of the areas of concern maps in the JRC MARS bulletin. Agric. Syst. 168, 213–223. https://doi.org/10.1016/j.agsy.2018.07.003.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., De Freitas, N., 2015. Taking the human out of the loop: a review of Bayesian optimization. Proc. IEEE 104, 148–175. https://doi.org/10.1109/JPROC.2015.2494218.

Statistics Canada, 2021. Statistics canada - surveys and statistical programs, Field Crop Reporting Series. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3401, Last accessed: April 28, 2021.

Supit, I., Hooijer, A., Van Diepen, C., 1994. System description of the WOFOST 6.0 crop simulation model implemented in CGMS. Vol. 1. Theory and Algo- rithms. Office for Official Publications of the European Communities, Luxembourg, p. 146.

USDA-NASS, 2012. The Yield Forecasting Program of NASS. United States Department of Agriculture (USDA) (Technical Report). ⟨https://www.nass.usda.gov/Education_and_Outreach/Understanding_Statistics/Yield_Forecasting_Program.pdf⟩. Last accessed: May 11, 2020.

USGS-EROS, 2021. USGS EROS Archive - Digital Elevation - Global 30 Arc-Second Elevation (GTOPO30). ⟨https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-global-30-arc-second-elevation-gtopo30?qt-science_center_objects=0#qt-science_center_objects⟩, (Accessed 11 May 2021).

Van Diepen, C., Wolf, J., Van Keulen, H., Rappoldt, C., 1989. WOFOST: a simulation model of crop production. Soil Use Manag. 5, 16–24. https://doi.org/10.1111/j.1475-2743.1989.tb00755.x.

van der Velde, M., Biavetti, I., El-Aydam, M., Niemeyer, S., Santini, F., van den Berg, M., 2019. Use and relevance of European Union crop monitoring and yield forecasts. Agric. Syst. 168, 224–230.

Van der Velde, M., Nisini, L., 2019. Performance of the MARS-crop yield forecasting system for the European Union: assessing accuracy, in-season, and year-to-year improvements from 1993 to 2015. Agric. Syst. 168, 203–212. https://doi.org/10.1016/j.agsy.2018.06.009.

van der Velde, M., Wriedt, G., Bouraoui, F., 2010. Estimating irrigation use and effects on maize yield during the 2003 heat wave in France. Agric. Ecosyst. Environ. 135, 90–97. https://doi.org/10.1016/j.agee.2009.08.017.

You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep Gaussian process for crop yield prediction based on remote sensing data, in: Thirty-First AAAI Conference on Artificial Intelligence. ⟨https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14435/14067⟩, (Accessed 11 May 2020).

Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., et al., 2016. Apache spark: a unified engine for big data processing. Commun. ACM 59, 56–65. https://doi.org/10.1145/2934664.

Zhao, Y., Potgieter, A.B., Zhang, M., Wu, B., Hammer, G.L., 2020. Predicting wheat yield at the field scale by combining high-resolution sentinel-2 satel- lite imagery and crop modelling. Remote Sens. 12, 1024. https://doi.org/10.3390/rs12061024.