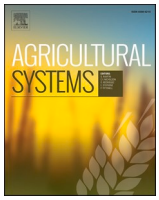
Agricultural Systems 187 (2021) 103016

Contents lists available at ScienceDirect

Agricultural Systems

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



Machine learning for large-scale crop yield forecasting

Dilli Paudel [a,*], Hendrik Boogaard [b], Allard de Wit [b], Sander Janssen [b], Sjoukje Osinga [a], Christos Pylianidis [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, Wageningen University and Research, PO Box 4e7, 6700 AA Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Crop yield prediction
Machine learning
Modularity
Reusability
Large-scale crop yield forecasting

ABSTRACT

Many studies have applied machine learning to crop yield prediction with a focus on specific case studies. The data and methods they used may not be transferable to other crops and locations. On the other hand, operational large-scale systems, such as the European Commission's MARS Crop Yield Forecasting System (MCYFS), do not use machine learning. Machine learning is a promising method especially when large amounts of data are being collected and published. We combined agronomic principles of crop modeling with machine learning to build a machine learning baseline for large-scale crop yield forecasting. The baseline is a workflow emphasizing correctness, modularity and reusability. For correctness, we focused on designing explainable predictors or features (in relation to crop growth and development) and applying machine learning without information leakage. We created features using crop simulation outputs and weather, remote sensing and soil data from the MCYFS database. We emphasized a modular and reusable workflow to support different crops and countries with small configuration changes. The workflow can be used to run repeatable experiments (e.g. early season or end of season predictions) using standard input data to obtain reproducible results. The results serve as a starting point for further optimizations. In our case studies, we predicted yield at regional level for five crops (soft wheat, spring barley, sunflower, sugar beet, potatoes) and three countries (the Netherlands (NL), Germany (DE), France (FR)). We compared the performance with a simple method with no prediction skill, which either predicted a linear yield trend or the average of the training set. We also aggregated the predictions to the national level and compared with past MCYFS forecasts. The normalized RMSE (NRMSE) for early season predictions (30 days after planting) were comparable for NL (all crops), DE (all except soft wheat) and FR (soft wheat, spring barley, sunflower). For example, NRMSE was 7.87 for soft wheat (NL) (6.32 for MCYFS) and 8.21 for sugar beet (DE) (8.79 for MCYFS). In contrast, NRMSEs for soft wheat (DE), sugar beet (FR) and potatoes (FR) were twice as much compared to MCYFS. NRMSEs for end of season were still comparable to MCYFS for NL, but worse for DE and FR. The baseline can be improved by adding new data sources, designing more predictive features and evaluating different machine learning algorithms. The baseline will motivate the use of machine learning in large-scale crop yield forecasting.

1. Introduction

Crop yield prediction is an important but complex problem, necessary for sustainable intensification and efficient use of natural resources (Phalan et al. 2014; Tilman et al. 2011). Crop yield forecasts are valuable to many stakeholders in the agri-food chain, including farmers, agronomists, commodity traders and policymakers (Basso and Liu 2019; Chipanshi et al. 2015). Crop yield is influenced by many crop-specific parameters, environmental conditions and management decisions (Fischer 2015), and it is difficult to build a reliable and explainable prediction model.

Field surveys, crop growth models, remote sensing, statistical models and their combinations have been commonly used to predict crop yield. On their own, these methods address slightly different aspects of crop yield forecasting. Field surveys try to capture the ground truth. Crop growth models simulate crop growth and development according to

* Corresponding author.
E-mail addresses: dilli.paudel@wur.nl (D. Paudel), hendrik.boogaard@wur.nl (H. Boogaard), allard.dewit@wur.nl (A. de Wit), sander.janssen@wur.nl (S. Janssen), sjoukje.osinga@wur.nl (S. Osinga), christos.pylianidis@wur.nl (C. Pylianidis), ioannis.athanasiadis@wur.nl (I.N. Athanasiadis).

https://doi.org/10.1016/j.agsy.2020.103016
Received 9 June 2020; Received in revised form 27 October 2020; Accepted 4 December 2020
Available online 14 December 2020

agronomic principles of plant, environment and management interactions (Basso et al. 2013; Chipanshi et al. 2015). Remote sensing methods rely on satellite imagery to capture the current state of crops and then to estimate the final yield (Lopez-Lozano et al., 2015). Statistical models use weather variables and the outputs of the three previous methods as predictors to derive linear relationships between the predictors and crop yield (e.g. Bussay et al. 2015). Recent studies have combined different methods in innovative ways to build yield forecasting models. For example, Lobell et al. (2015) and Zhao et al. (2020) used high-resolution remote sensing data and crop modeling to build statistical models to forecast the actual yield. Similarly, Newlands et al. (2014) developed a probabilistic yield forecasting framework for Canada using remote sensing, crop modeling, Bayesian inference and statistical models.

Machine learning takes a data-driven or empirical modeling approach to learn useful patterns and relationships from input data (Willcock et al., 2018) and provides a promising avenue for improving crop yield predictions. Machine learning algorithms approximate a function that relates features or predictors to labels, such as crop yield. Similar to statistical models, machine learning algorithms can utilize the outputs of other methods as features. In addition, machine learning algorithms have some distinct benefits: they can model non-linear relationships between multiple data sources (Chlingaryan et al. 2018); their performance generally improves when more training data is available (Goodfellow et al. 2016); and they can become robust to noisy data by using regularization techniques that help decrease the variance and the generalization error (James et al. 2013; Goodfellow et al. 2016). Therefore, machine learning could combine the benefits of other methods, such as crop growth models and remote sensing, with data-driven modeling to make reliable crop yield predictions.

Many studies have applied machine learning to predict yields of certain crops in specific locations, but it is unclear whether their data and methods are transferable to other crops and locations. Some of them used empirical data collected for specific purposes that may not be available for other crops or locations (e.g. Pantazi et al. (2016)). Some others used generally available climate and satellite data, but made crop and location-specific design choices that limit their reusability (e.g. Cai et al. (2019)). In this paper, we seek to address the need for modular and reusable workflows that would help understand the usefulness of various data sources, predictors or features and machine learning algorithms for different crops across spatial and temporal settings. Reusable workflows would allow researchers to run repeatable experiments, such as early season or end of season predictions, for different crops and countries with standard input data and obtain reproducible results. The models could be improved for specific crops and locations using new data sources, more advanced features and other optimizations.

Large-scale crop yield forecasting systems, such as the MARS Crop Yield Forecasting System (MCYFS) of the European Commission's Joint Research Centre (JRC) and the National Agricultural Statistics Service (NASS) of US Department of Agriculture (USDA), have the infrastructure and historical data to build and assess crop yield prediction models for different crops and locations. However, the operational systems we know of do not use machine learning. They build statistical models from weather observations, field survey results, crop growth model outputs, remote sensing indicators and yield statistics (MARSWiki, 2020; USDA-NASS, 2012). Van der Velde and Nisini (2019) evaluated the performance of MCYFS from 1993 to 2015 and found that there is no significant improvement in MCYFS performance from 2006 onwards. Machine learning is a promising method especially when a large amount of data is being collected and made public (Lokers et al. 2016; GODAN 2020; EC-JRC 2020). A reusable and extensible workflow based on inputs similar to MCYFS would motivate the adoption of machine learning in large-scale crop yield forecasting.

We present a machine learning baseline for large-scale early and end of season crop yield forecasts. The baseline is a general machine learning workflow emphasizing three principles: (i) correctness, (ii) modularity, and (iii) reusability. First, our methodology focuses on how to create features that can explain crop growth and development based on agronomic principles of crop modeling, and how to apply machine learning without leaking information from the test set. Second, a modular design permits the workflow to be improved or extended by adding new data sources, designing more advanced features and evaluating different machine learning methods. Third, reusability addresses the transferability of the workflow to different crops and countries with small configuration changes. The results obtained can be a starting point for further optimizations.

We tested the machine learning baseline on three countries (the Netherlands (NL), Germany (DE), France (FR)) and five crops (soft wheat, spring barley, sunflower, sugar beet, potatoes) using MCYFS (MARSWiki, 2020; EC-JRC, 2020) and Eurostat data (Eurostat, 2020a, Eurostat, 2020b). We ran experiments to predict early season and end of season crop yield at NUTS2 or NUTS3 level (see Eurostat (2016), *Section E* of *Supplement 1*). We compared the regional predictions with a simple method with no prediction skill, which we call the "null" method. The null method either predicted a linear yield trend or the average of the training set. We also aggregated the predictions to the national (NUTS0) level and compared the results with past MCYFS forecasts.

The remainder of the paper is organized as follows: *Section 2* reviews related work in the field; *Section 3* describes the methodology and the case studies; *Section 4* presents the results; *Section 5* discusses our findings and areas for further research; and *Section 6* summarizes our conclusions. *Supplement 1* provides a brief introduction to MCYFS and machine learning, and the workflow details not included in *Section 3* (*Methodology*). *Supplement 2* includes a Jupyter notebook implementation (see https://jupyter.org/) of the machine learning baseline, a sample data set and supporting materials for *Section 4* (*Results*) and *Section 5* (*Discussion*).

2. Related work

Four methods or combinations thereof have been commonly used to predict crop yield: (i) field surveys, (ii) crop growth models, (iii) remote sensing, and (iv) statistical models. These methods have their strengths and weaknesses. Field surveys try to capture the ground truth by means of grower-reported surveys and objective measurement surveys (USDA-NASS, 2012). These surveys suffer from declining responses (Schnepf 2017), resource restrictions and reliability concerns due to sampling and non-sampling errors (Chipanshi et al. 2015). Process-based crop models simulate crop growth and development by using crop parameters, environmental conditions and management practices as input. They apply agronomic principles of crop growth and development that apply across space and time (Basso and Liu 2019). However, they do not account for all yield-reducing factors and have considerable data and calibration requirements (De Wit et al., 2019). Remote sensing tries to capture current information about crops by using satellite images. Remote sensing data are globally available under open data policies and they do not suffer from human errors (Chipanshi et al. 2015). However, remote sensing observations only provide indirect measurements of crop yield, namely observed radiance (Dorigo et al., 2007; Jones and Vaughan, 2010), and therefore rely on biophysical or statistical models to convert satellite observations into a yield prediction (e.g. Lopez-Lozano et al., 2015). Statistical models use meteorological indicators and the outputs of the three previous methods as predictors. These models estimate the yield trend attributable to technological advancements in genetics and management (Basso et al. 2013) and fit linear models between predictors and yield residuals (e.g. Bussay et al. 2015). They provide reasonable accuracy and explainability but cannot be extrapolated to other spatial and temporal contexts (Basso et al. 2013).

Machine learning has gained popularity in agricultural applications due to its success in other fields, such as medicine (e.g. Kang et al. (2015)), bioinformatics (e.g. Mackowiak et al. (2015)) and natural language processing (e.g. Socher et al. (2012)). Recent reviews

(Chlingaryan et al., 2018; Kamilaris and Prenafeta-Boldu, 2018; Liakos et al., 2018) have looked at the applications of machine learning in agriculture. Many studies (included in the reviews and others) have applied traditional (or shallow) machine learning and deep learning to crop yield prediction. Among applications of shallow methods, Shahhosseini et al. (2019) built machine learning metamodels from outputs of the APSIM crop model (Holzworth et al., 2015) to predict maize yield and nitrogen loss in the US; Jeong et al. (2016) applied Random Forests (Breiman 2001) to predict wheat yield globally and maize and potato yield in the US; and Gonzalez Sanchez et al. (2014) compared the performance of four machine learning algorithms on ten crops in Mexico. Among applications of deep learning, Crane-Droesch (2018) applied semiparametric deep neural networks to predict corn yield in the US; You et al. (2017) leveraged representation learning ideas to predict soybean yield in the US; and Pantazi et al. (2016) used self-organizing maps (Von der Malsburg 1973; Kohonen 2001) to predict within-field variation of wheat yield in the UK. These examples show that both shallow and deep methods can predict crop yield. However, they focus on optimizing performance for specific case studies. Some studies (e.g. Pantazi et al. (2016)) use empirical data collected for a specific location. Others use generally available data (e.g. You et al. (2017)), but focus on novel methods to improve performance. Some of them cover different crops (e.g. Jeong et al. (2016); Gonzalez Sanchez et al. (2014)) and locations (e.g. Jeong et al. (2016)), but their emphasis is again on performance compared to statistical methods, not on reusable methods. Therefore, it is unclear whether their data and methods are transferable to other crops and locations.

Large-scale crop yield forecasting systems, such as MCYFS, NASS and Statistics Canada, have historical data, infrastructure, expertise, evaluation frameworks and dissemination channels to build and assess crop yield prediction models for different crops and locations (see *Section A* of *Supplement 1*; USDA-NASS (2012); Statistics Canada 2019). To our knowledge, these systems do not use machine learning. They build statistical models using weather observations, field survey results, crop growth model outputs, remote sensing indicators and yield statistics. NASS uses survey results and linear statistical models to forecast crop yields (USDA-NASS, 2012). MCYFS provides a control board for human experts to run analyses and to build crop yield prediction models using two methods. The first method estimates the trend related to technological improvements and applies a simple or multiple linear regression on the yield residuals using crop growth model outputs and meteorological indicators (MARSWiki, 2020; Lecerf et al. 2019). The second method applies principal component analysis (Wold et al., 1987) and cluster analysis to identify similar years and forecast the yield based on similarities (MARSWiki, 2020; Lecerf et al. 2019). In addition, MCYFS experts use their judgment based on information from other sources, such as farming magazines. No previous work has applied machine learning to MCYFS data. A generic workflow based on MCYFS data would motivate the use of machine learning in large-scale crop yield forecasting.

Common applications of statistical models estimate the yield trend and detrend yield values before building regression models between predictors and yield residuals (e.g. Lecerf et al. 2019; Bussay et al. 2015). The yield trend for later years includes information from the earlier years. Evaluating such models by including earlier years in the test set and later years in the training set would cause information leakage. Some applications of machine learning to crop yield prediction have also used yield trend or other information from previous year(s). However, not all of them have avoided information leakage. For instance, Cai et al. (2017) ran cross-validation to train and optimize their prediction models. During cross-validation, the test fold can be in a bin earlier than the training folds, thus leading to information leakage. To avoid this leakage, Shahhosseini et al. (2019) adopted a time-based look-forward validation that always put the training data before the test data. We designed a machine learning workflow for crop yield prediction emphasizing the application of machine learning without information leakage.

The need for modularity and reusability in agricultural modeling has been stressed by Janssen et al. (2017) and Holzworth et al. (2014). In the case of crop yield prediction, modular design makes it possible to run experiments to test alternative configurations, such as early or end of season prediction. Similarly, modularity is crucial to minimize and diagnose unexpected outcomes when one part of the workflow is updated (Janssen et al. 2017). Reusability has not been a design goal in agricultural system modeling; more emphasis has been placed on the underlying science (Holzworth et al., 2014). Example applications of machine learning to crop yield prediction show a similar pattern. Reusability or transferability of methods has not been emphasized. We have designed the machine learning baseline focusing on modularity and reusability.

3. Methodology

We designed a machine learning workflow for crop yield prediction using MCYFS data. We evaluated the workflow by predicting crop yield at NUTS2 or NUTS3 levels for five crops and three countries. For each crop and country, we ran experiments to predict early season (30 days after planting) and end of season crop yield with and without using the estimated yield trend from previous years. For each experiment, we compared the regional predictions with a simple method with no prediction skill (the "null" method) and also aggregated the predictions to national (NUTS0) level and compared them with past MCYFS forecasts.

The overall workflow has two parts (Fig. 1). The first part consists of preprocessing and feature design, which are specific to data sources, and splitting data into training and test sets. The second part, focusing on machine learning, is independent of data sources. Data from various sources, such as crop growth simulation outputs, weather observations and yield statistics, were homogenized and aligned to the same spatial and temporal resolutions. The data was split into training and test sets before designing features (see *Section 3.1.2*). Some data sources required feature design; others were directly used as features. Once we had features and labels, machine learning algorithms were trained and optimized on the training set and evaluated on the test set.

We designed the workflow emphasizing three principles: correctness, modularity and reusability.

The overall workflow has two parts. The first part includes preprocessing and feature design. The second part includes machine learning.

3.1. Workflow design: correctness

For correctness, we focused on how to design explainable features and how to apply machine learning without information leakage.

3.1.1. Explainable feature design

We incorporated agronomic principles from crop modeling to design features with physical meaning in terms of their impact on crop growth and development. Based on the outputs of the WOFOST crop model (Supit et al., 1994; Van Diepen et al., 1989), we selected 3 dekads (10-day periods) when significant changes occur in the crop's development stage (DVS): (i) START_DVS (DVS $\geq$ 0) is when the crop emerges from the soil, (ii) START_DVS1 (DVS $\geq$ 100) is the middle of the flowering phase, and (iii) START_DVS2 (DVS $\geq$ 200) is when the crop becomes ripe. (See De Wit et al. (2019) for a summary of how DVS is calculated.) Using these 3 dekads, we divided the crop season into 6 periods: (i) pre-planting window, (ii) planting window, (iii) vegetative phase, (iv) flowering phase, (v) yield formation phase, and (vi) harvest window (Table 1).

For each period of the crop calendar, we identified the weather indicators, crop growth model outputs and remote sensing indicators that affect or capture the state of crop growth and development (Table 2). Using these indicators, we designed 3 types of features: (i) maximum

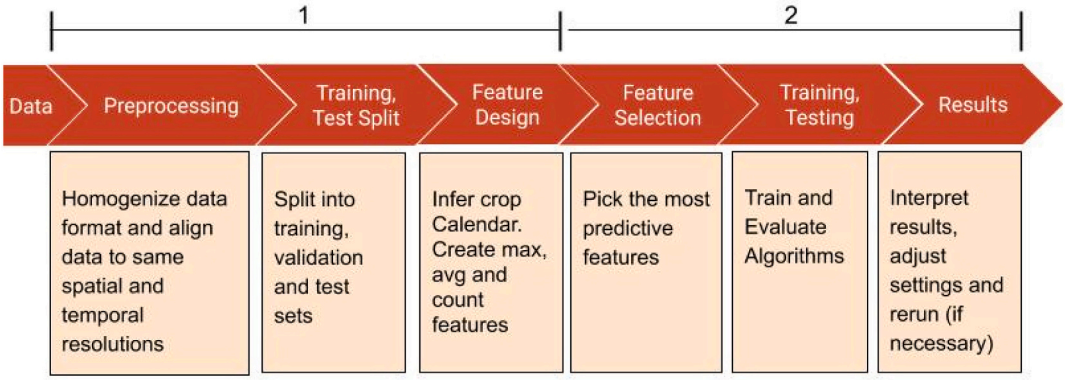


Fig. 1. The high-level workflow.

Table 1
Crop calendar definition.

Period	Start Dekad	End Dekad
Pre-planting window (p0)	min(1, avg START_DVS - 11)	avg START_DVS
Planting window (p1)	avg START_DVS - 1	avg START_DVS + 1
Vegetative phase (p2)	avg START_DVS	avg START_DVS1
Flowering phase (p3)	avg START_DVS1–1	avg START_DVS1 + 1
Yield Formation phase (p4)	avg START_DVS1	avg START_DVS2
Harvest window (p5)	avg START_DVS2–1	avg START_DVS2 + 1

We inferred the crop calendar from WOFOST outputs by selecting 3 dekads that signified important development stage changes. START_DVS is when the crop emerges from the soil. START_DVS1 is the middle of the flowering phase. START_DVS2 is when the crop becomes ripe. The pre-planting window was restricted to a maximum of 12 dekads or 4 months.

Table 2
Feature design using crop modeling principles.

Period	Maximum values	Average values	Counts of days or dekads with extreme values
Pre-planting window		TAVG, PREC, CWB	
Planting window		TAVG, PREC	RSM, TMIN, PREC
Vegetative phase	WLIM_YB, TWC, WLAI	RSM, TAVG, CWB, FAPAR	RSM
Flowering phase		PREC	RSM, PREC, TMAX
Yield Formation phase	WLIM_YB, WLIM_YS, TWC, WLAI	RSM, CWB, FAPAR	RSM
Harvest window		PREC	PREC

We identified indicators affecting crop growth and development during different crop calendar periods. Weather indicators included average temperature (TAVG), precipitation (PREC), climate water balance (CWB = precipitation - evapotranspiration), minimum temperature (TMIN) and maximum temperature (TMAX). WOFOST outputs included water-limited yield biomass (WLIM_YB), water-limited yield storage (WLIM_YS), water-limited leaf area index (WLAI), relative soil moisture (RSM) and total water consumption (TWC). Remote sensing indicators included the fraction of absorbed photosynthetically active radiation (FAPAR).

values for accumulative indicators, such water-limited yield biomass, (ii) counts of days or dekads for indicators related to extreme conditions, such as maximum temperature, and (iii) average values for other indicators. *Section E* of *Supplement 1* includes details about the data sources and the indicators used in feature design. Features for extreme conditions counted days or dekads with values +/− 1 standard deviation and +/− 2 standard deviations from the average. By taking the averages and standard deviations of indicators, we made the workflow generic and reusable. Similarly, by creating a large number of features, we explored the space of thresholds for extreme conditions and leveraged feature selection (see *Section C.2.2* of *Supplement 1*) to identify the features with the appropriate thresholds.

Some studies have experimented with crop calendar periods for one crop (e.g. Han et al. (2020) for winter wheat, Shahhosseini et al. (2019) for maize), but they did not explore the transferability of their approach to other crops. Lopez-Lozano et al. (2015) identified the optimal period for the correlation between fraction of absorbed photosynthetically active radiation (FAPAR) and yield statistics for three crops. We did not calculate optimal periods; instead, we devised a generic method that could be reused for different crops and countries.

3.1.2. Machine learning without information leakage

We applied supervised learning (see *Section B* of *Supplement 1*), specifically supervised regression, to crop yield prediction. Supervised learning relies on training examples that include features as well as labels, such as yield statistics, to learn a function that relates features to labels. We split the full dataset into training and test sets. When using the yield trend, we added the last few years for each region to the test set (Fig. 2a). This restriction was necessary because later years would contain yield trend estimated from earlier years and having earlier years in the test set would cause information leakage. When not using the yield trend, we could have used random splits. However, we needed the same test years for all regions to compare the predictions with MCYFS (see *Section 3.5*). Therefore, we added every *n*th year to the test set, with *n* determined by the test fraction. In both cases, we allocated 70% of the data for training and 30% for testing. We used the training set to train and optimize a model and the test set for the final evaluation. We split the data into training and test sets before feature design because feature design relied on crop calendar information (see Table 1) and the averages and standard deviations of the indicators shown in Table 2. We inferred the crop calendar and calculated indicator statistics only using the training set.

We optimized the hyperparameters of feature selection (the number of features to select) and prediction algorithms (e.g. the number of neighbors for *k*-nearest neighbors) by dividing the training set into validation folds. When using the yield trend, we could not run cross-validation because the test fold could end up in a bin earlier than the training folds and that would cause information leakage. Therefore, we used a time-based *k*-fold sliding validation (Fig. 2b). For example, NL data was available from 1994 to 2018 and the training years included 1994 to 2011. Assuming 5-folds, we trained the first iteration of *k*-fold sliding validation on data from 1994 to 2007, the second iteration on 1995 to 2008 and so on until the fifth iteration, which we trained on 1998 to 2011. When not using the yield trend, we applied regular *k*-fold cross-validation.

We created pipelines consisting of feature scaling, feature selection

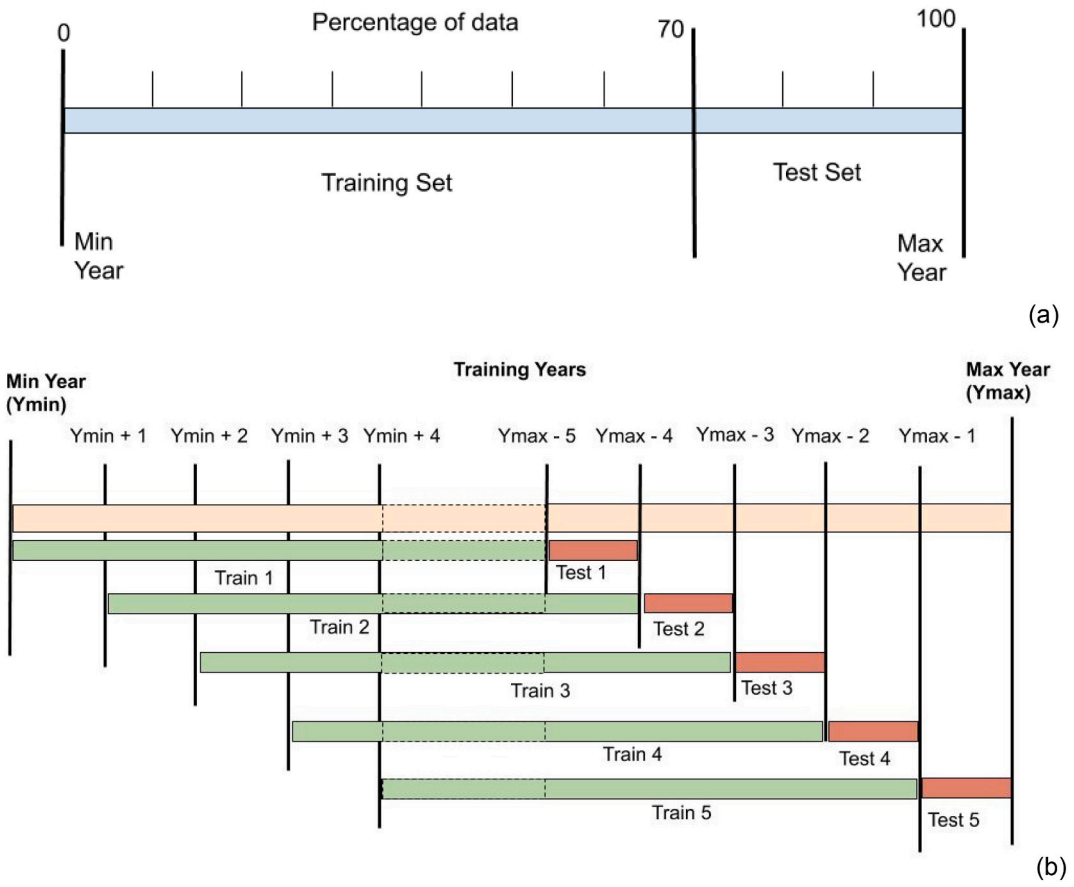


Fig. 2. Training, validation and test splits when using yield trend.
(a) For each region, we split the full dataset into training and test sets.
(b) We further divided the training set into validation training and test sets for feature selection and hyperparameter optimization using a time-based 5-fold sliding validation.

and training stages (see *Section C.2* of *Supplement 1*) to avoid information leakage during feature selection and training (Muller and Guido 2016). The pipelines ensured each stage of training and optimization used only the training data. In effect, the parameters for scaling features (e.g. mean and standard deviation), the number of features to select and the feature weights for the trained model were learned from the training set. Furthermore, we optimized the hyperparameters using only the training set. When optimizing the hyperparameters, the pipeline was run for each iteration of 5-fold sliding validation or 5-fold cross-validation. Therefore, all stages of the pipeline (feature scaling, feature selection and training) were run using the training folds and the trained model was evaluated using the corresponding test fold.

3.2. Workflow design: Modularity

For modularity, we focused on making the baseline relatively easy to improve and extend. We minimized the dependencies between successive stages of the workflow. We chose extensible data structures to allow the indicators selected for feature design to change without affecting the workflow (*Fig. 3*). The goal was to simplify the process of designing new features or improving existing features with new data. For example, features for extreme conditions count days or dekads with values $+/-$ 1 standard deviation and $+/-$ 2 standard deviations from the average. The use of the averages and standard deviations of indicators makes the workflow generic and reusable. However, when crop-specific thresholds for different indicators are available, such data can be used to manually define more accurate and predictive features (see *Section C.1.3* of *Supplement 1* for examples).

We defined configuration options to control data flow when running various experiments (Fig. 4). For example, geographical information about region centroids was not included by default, but could be used if desired. Different experiments could be run by updating the configuration options and running the workflow; the workflow itself did not change. In addition, the generated features could be saved in a file and loaded later for machine learning, making the machine learning part of the workflow independent of preprocessing and feature design. Similarly, predictions of machine learning algorithms could be saved to a file and loaded later for comparison with MCYFS (*Section 3.5*).

We defined feature selection and prediction algorithms in a modular and extensible manner to enable experimentation with different algorithms (Fig. 4). Feature selection algorithms could be added by specifying the number of features to select. Similarly, prediction algorithms could be added by setting certain hyperparameters to default values and specifying the values of other hyperparameters to be optimized. We defined the range of values of hyperparameters as lists that could be extended or shortened.

3.3. Workflow design: Reusability

We designed the workflow to be reusable for different crops and countries. We applied data homogenization to standardize the filenames, file formats and data columns, thereby minimizing the amount of input required to run the workflow. We reused the same feature design principles for different case studies (see Section 3.1.1). Data homogenization and configuration options for crop name, country (two letter code, e.g. NL) and NUTS level made it possible to run the workflow

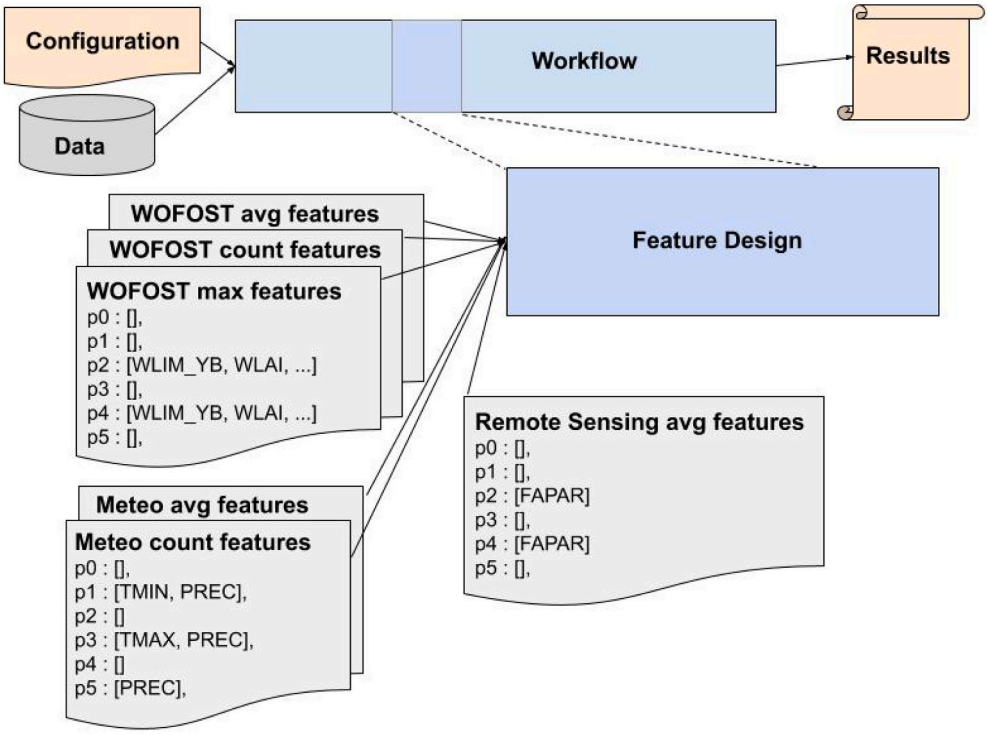


Fig. 3. Modularity and extensibility in feature design.
Features were designed using extensible lists of indicators for each crop calendar period. Lists of indicators correspond to entries in Table 2.

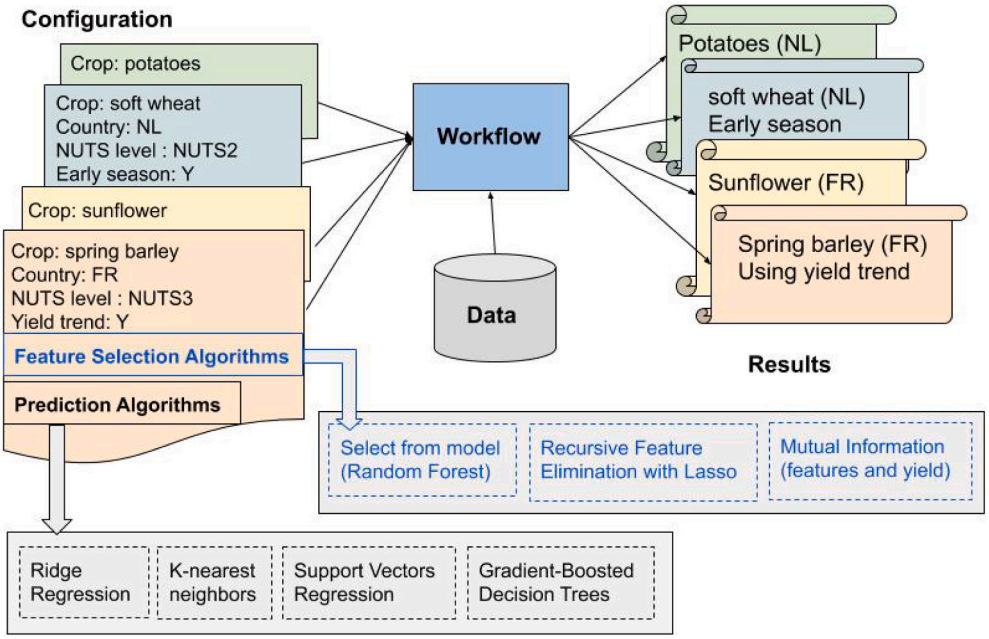


Fig. 4. Configuration Options.
Configuration options were used to select the case study and the experiment being run. Feature selection algorithms and prediction algorithms were defined using extensible data structures. Therefore, different algorithms could be added or removed to study their benefits without affecting the workflow. (see *Section C.2.3* of *Supplement 1* for more details about the algorithms.)

for different crops, countries and NUTS levels (Fig. 4). We set most configuration options to reasonable defaults to avoid specifying all of them for every experiment.

3.4. Data, case studies and experiments

We used WOFOST crop growth model outputs, weather observations, remote sensing data, soil data, region centroids, modeled crop area fractions and yield statistics for the Netherlands (NL), Germany (DE) and France (FR) to evaluate the workflow. We had NL data for 12 NUTS2 regions from 1994 to 2018, FR data for 101 NUTS3 regions from 1989 to 2018 and DE data for 401 NUTS3 regions from 1999 to 2018. As described in *Section 3.1.2*, we used 70% of the data for training and 30% for testing. *Section E* of *Supplement 1* provides more details about the data and the NUTS regions. We did not use region centroids by default because it was unclear whether they provided additional information not included in WOFOST outputs and weather observations.

We used thirteen case studies and ran four experiments for each case

study to verify correctness, modularity and reusability of the machine learning workflow. First, to verify the explainability of features, we counted the frequencies of selected features for each crop across different countries and algorithms. We deferred a detailed analysis of feature importance for future research. Second, to verify modularity of the workflow, we ran four experiments for each crop and country with options for using yield trend (*Yes* or *No*) and early season prediction (*Yes* or *No*). For early season prediction, we used current season information up to 30 days after planting. For end of season prediction, we used current season information up to the end of the harvest window. Third, to verify reusability, we ran the four experiments for thirteen case studies: soft wheat (NL, DE, FR), spring barley (NL, DE, FR), sunflower (FR), sugar beet (NL, DE, FR) and potatoes (NL, DE, FR). We tested the optional components of the workflow (e.g. using centroids, saving and loading features) on soft wheat (NL). For NL, predictions were made at NUTS2; for DE and FR, predictions were made at NUTS3. Overall, we tested the workflow with two NUTS levels, five crops and three countries.

We evaluated the performance of four machine learning algorithms in predicting crop yield: (i) Ridge Regression (Hoerl and Kennard 1970), (ii) *K*-nearest Neighbors Regression (Cover and Hart 1967; Aha et al. 1991), (iii) Support Vector Machines Regression (Boser et al. 1992; Cortes and Vapnik 1995), and (iv) Gradient Boosted Decision Trees Regression (see Friedman 2001; Hastie et al. 2009). These methods represent different classes of algorithms based on how they learn the relationships between features and labels. *Section C.2.3* of *Supplement 1* provides a brief description of these algorithms. The predictions of machine learning algorithms were compared with those of a simple method with no skill (the "null" method). When yield trend was not used, the null method was equivalent to the ZeroR algorithm (see Baskin et al. 2017), which predicts the average of the training set. When yield trend was used, the null method predicted the linear yield trend estimated from a 5-year window. All algorithms were evaluated using mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and the coefficient of determination or R^2. MAE and RMSE were compared using their normalized counterparts. The normalized errors were calculated by dividing the mean error with the mean yield of the test set. *Section C.2.3* of *Supplement 1* provides the details about the evaluation metrics used.

3.5. Comparison with MCYFS forecasts

We aggregated the predictions of the machine learning baseline from NUTS2 (NL) or NUTS3 (DE, FR) to national (NUTS0) level to compare with past MCYFS forecasts. NUTS2 or NUTS3 predictions were aggregated to NUTS0 by weighting them on the modeled crop area. Cerrani and Lopez Lozano (2017) have described in detail the algorithm used to model crop areas for different NUTS levels. Predictions at NUTS3 were aggregated to NUTS2 based on crop area weights for NUTS3 regions, and predictions at NUTS2 were further aggregated to NUTS1 using crop area weights for NUTS2 regions, and so on. We compared the aggregated NUTS0 predictions and the actual MCYFS forecasts (see Van der Velde and Nisini (2019)) using the official Eurostat national yield statistics (Eurostat, 2020a) as the reference. We evaluated the two sets of predictions using mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and the coefficient of determination or R^2.

We had to make an adjustment to training and test splits to aggregate the crop yield predictions from NUTS3 or NUTS2 to NUTS0: the test set had to include the same set of years for all regions. (Note this restriction is necessary only when aggregating the predictions to NUTS0 level.) When we made the test years the same, some regions and test years were missing predictions. We filled the missing predictions in two ways. First, if the region had predictions for other test years, we filled the missing value with the average of the remaining years. Second, if the region had no predictions at all, we ignored the region and adjusted the area fractions of other sibling regions (with the same parent NUTS region).

3.6. Implementation

We used Apache Spark dataframes (Zaharia et al. 2016) for data preprocessing and feature design, and applied machine learning using the scikit-learn python package (Pedregosa et al. 2011). We developed and tested the workflow in Google Colaboratory (https://colab.research.google.com/) and ran the different experiments in Google Dataproc cluster (https://cloud.google.com/dataproc) and Microsoft Azure Databricks (https://azure.microsoft.com/en-us/services/databricks/).

4. Results

To verify explainability of features, we looked at feature selection frequencies for each crop across different countries and algorithms. To demonstrate modularity and reusability, we ran four experiments with options to use yield trend (*Yes* or *No*) and to predict early in the season (*Yes* or *No*) for all thirteen crop and country combinations: soft wheat (NL, DE, FR), spring barley (NL, DE, FR), sunflower (FR), sugar beet (NL, DE, FR) and potatoes (NL, DE, FR). Predictions for NL were made at NUTS2 and predictions for DE and FR were made at NUTS3. All results were aggregated to national level and compared with past MCYFS forecasts. In this section, we present the normalized RMSE for different case studies. MAPE results are included in *Section D* of *Supplement 1*, and all results including normalized MAE, normalized RMSE, MAPE and R^2 for all case studies, experiments and algorithms are provided in *Supplement 2*. Results of optional experiments (e.g. using region centroids data, saved features and saved predictions) are also included in *Supplement 2*.

4.1. Feature selection frequencies

Feature selection counts for potatoes show that soil water holding capacity was always selected (Table 3). Similarly, all the features for the pre-planting window were frequently selected. For the planting window, averages and extremes of temperature and precipitation were important. Similarly, most frequently selected features for the vegetative phase were the fraction of absorbed photosynthetically active radiation (FAPAR), water-limited yield biomass, leaf area index and average temperature. Precipitation and maximum temperature extremes were important for the flowering phase. For the yield formation phase, FAPAR and WOFOST indicators such as total water consumption, water-limited yield biomass and yield storage were important. Finally, average and extremes of precipitation were important during the harvest window. Feature selection frequencies are generally consistent with the factors affecting crop growth and development during these periods. For example, temperature extremes during the flowering phase and precipitation extremes during planting and harvest windows (see Van der Velde et al. 2018) are known to influence crop yield. Feature selection frequencies for other crops are included in *Supplement 2*.

4.2. Yield trend vs. no yield trend

We compared the end of season predictions of the Gradient Boosted Decision Trees (GBDT) algorithm with the option of using yield trend (*Yes* or *No*) to those of the null method (Fig. 5*; Fig. 13*). We chose GBDT because its performance was better than other algorithms in most cases. Except for a few instances (e.g. normalized RMSE for sugar beet (NL) and sugar beet (DE) "No Yield Trend" (Fig. 5); MAPE for potatoes (FR) "Yield Trend" (*Fig. 13*)), machine learning performed better than the null method. Because of the differences in training and test sets (see *Section 3.1.2*), we cannot directly compare "Yield Trend" and "No Yield Trend". Nevertheless, the two sets of error values were quite similar, indicating that machine learning could be applied with or without yield trend. When using the yield trend, the test set included the tail end of available

Table 3
Feature selection frequencies for potatoes (No Yield Trend).

	Static Features (Frequency)
	Soil water holding capacity (12)
Period	Features (Frequency)
Pre-planting window	avg TAVG (9), avg CWB (8), avg PREC (8)
Planting window	avg TAVG (4), avg PREC (6), TMIN >1 STD (5), PREC >1 STD (4), TMIN <2 STD (3), TMIN <1 STD (3), RSM < 2 STD (1), TMIN >2 STD (1)
Vegetative phase	max WLIM_YB (11), max TWC (7), max WLAI (7), avg RSM (4) avg FAPAR (12), avg TAVG (11), avg CWB (9), RSM > 2 STD (3)
Flowering phase	avg PREC (8), TMAX >1 STD (4), TMAX <1 STD (4), RSM < 1 STD (3), PREC >1STD (3), PREC >2 STD (3), TMAX >2 STD (1), TMAX <2 STD (1)
Yield formation phase	avg FAPAR (12), max WLIM_YB (11), max WLIM_YS (8), max TWC (8), max WLAI (6), avg RSM (8), avg CWB (7), RSM > 2 STD (4), RSM < 1 STD (4)
Harvest window	PREC >2 STD (4), avg PREC (3)

Selection frequencies were aggregated for three countries (NL, DE, FR) and four algorithms. Weather indicators included average temperature (TAVG), precipitation (PREC), climate water balance (CWB = precipitation - evapotranspiration), minimum temperature (TMIN) and maximum temperature (TMAX). WOFOST outputs included water-limited yield biomass (WLIM_YB), water-limited yield storage (WLIM_YS), water-limited leaf area index (WLAI), relative soil moisture (RSM) and total water consumption (TWC). Remote sensing indicators included the fraction of absorbed photosynthetically active radiation (FAPAR). Other abbreviations: avg= average, max = maximum, min = minimum, STD = standard deviation.

years. Therefore, using the yield trend would be useful to make predictions for the future. The "No Yield Trend" approach could be useful to make predictions for missing years.

4.3. Early season vs. end of season predictions

Early season predictions using yield trend (*Fig. 6*; *Fig. 14*) indicated that the baseline could make early season predictions better than the null method. We selected GBDT for comparison because its performance was better than other algorithms in most cases. The normalized RMSE and MAPE values for machine learning were lower than those for the null method in all instances except MAPE for potatoes (FR) (*Fig. 14*). The null method predicted the yield using a linear 5-year trend. Early season predictions were made 30 days (or 3 dekads) after planting. End of season predictions were made at the end of the harvest window. Both early season and end season predictions used the yield values of 5 previous years, soil data and the current season information up to the prediction dekad. Except for Spring Barley (NL), error values for the machine learning baseline improved slightly over the course of the season.

4.4. Comparison with MCYFS forecasts

We aggregated the predictions of the machine learning baseline to NUTS0 and compared them with past MCYFS forecasts using Eurostat national yield statistics as the reference. Because the MCYFS method performs trend analysis, we compared the predictions of machine learning algorithms using the yield trend. For comparison, we used predictions from the best machine learning algorithm and the selected algorithm varied by case study. The details are included in *Supplement 2*. For early season, we compared the predictions of machine learning for 30 days after planting with MCYFS forecasts from the closest dekad (Fig. 7a*; Fig. 15a*). We also compared machine learning predictions at the end of the harvest window with the final MCYFS prediction of the year (Fig. 7b*; Fig. 15b*). The machine learning baseline performed similar to MCYFS early in the season. Predictions were comparable for NL (all four crops) and DE (spring barley, sugar beet, potatoes) and FR (soft wheat, spring barley, sunflower). For example, the Normalized RMSE was 7.87 for soft wheat (NL) (6.32 for MCYFS), 8.21 for sugar beet (DE) (8.79 for MCYFS) and 10.63 for sunflower (FR) (10.91 for MCYFS). On the other hand, predictions for DE (soft wheat) and FR (sugar beet and potatoes) were much worse; the Normalized RMSE was 16.38 for soft wheat (DE) (6.21 MCYFS), and 14.34 sugar beet (FR) (MCYFS 7.42). As the season progressed, MCYFS forecasts improved significantly while machine learning predictions did not improve as much (Fig. 7a,b*; Fig. 15a,b*). Predictions for NL were still comparable to MCYFS (e.g. Normalized RMSE was 3.05 for soft wheat (NL) (MCYFS 5.48)), but worse for DE and FR. The baseline used the same data sources throughout the season: WOFOST outputs, weather observations, remote sensing indicators and soil data. On the other hand, MCYFS uses other sources of information, such as media reports and farming magazines, to update their predictions. Moreover, the role of MCYFS analysts is key as they investigate the underlying feature data, identifying the ones that better explain crop growth and yields, and select the appropriate statistical models to produce reliable yield forecasts (Lopez-Lozano and Baruth, 2019).

5. Discussion

Previous studies (e.g. Shahhosseini et al. 2019; Cai et al. 2019; You et al. 2017; Jeong et al. 2016) have demonstrated that machine learning can play an important role in crop yield prediction and the same was confirmed by our results. Likewise, machine learning has the potential to build on other methods of yield prediction, such as field surveys, crop growth models and remote sensing. Prior applications of machine learning to crop yield prediction focused on optimizing performance for specific case studies. We focused on a generic workflow that could be used to investigate the potential of machine learning across different crops and locations. The machine learning baseline covers the methodological aspects of applying machine learning and acts as a baseline in terms of performance. Future applications of machine learning could investigate in more detail the advantages of combining machine learning with other methods, such as crop growth models and remote sensing, and compare their results with the baseline.

We designed the machine learning baseline emphasizing three principles: correctness, modularity and reusability. First, we focused on *correctness* to design explainable features and to apply machine learning without information leakage. When working with time series data, such as crop yield, features designed using values from previous years, such as the yield trend, are used. Whenever information from previous years is included in features, particular attention is required to avoid information leakage. The baseline presents a time-based training and test split and a *k*-fold sliding validation to ensure that information from the test set is not used during training. Second, we emphasized *modularity* to let the workflow evolve and to run experiments with alternative configurations. The workflow supports incremental changes to extend and optimize the baseline for specific case studies. Third, we focused on *reusability* to enable the same workflow to run for different crops and locations. The emphasis on modularity and reusability will encourage model and software reuse and prevent a proliferation of monolithic and duplicate software implementations (Janssen et al. 2017; Holzworth et al., 2014).

A key innovation of the baseline is the feature design method followed by feature selection later in the workflow. We designed features based on agronomic principles from crop modeling. We identified indicators that affect crops during different crop calendar periods. We also included features to account for extreme conditions. Features for extreme conditions were based on averages and standard deviations of indicators, making the workflow generic and reusable. By creating a large number of features, we explored the space of thresholds for extreme conditions and leveraged feature selection to identify the appropriate thresholds. Similarly, instead of having experts hand pick features, we generated a large number of features and applied feature

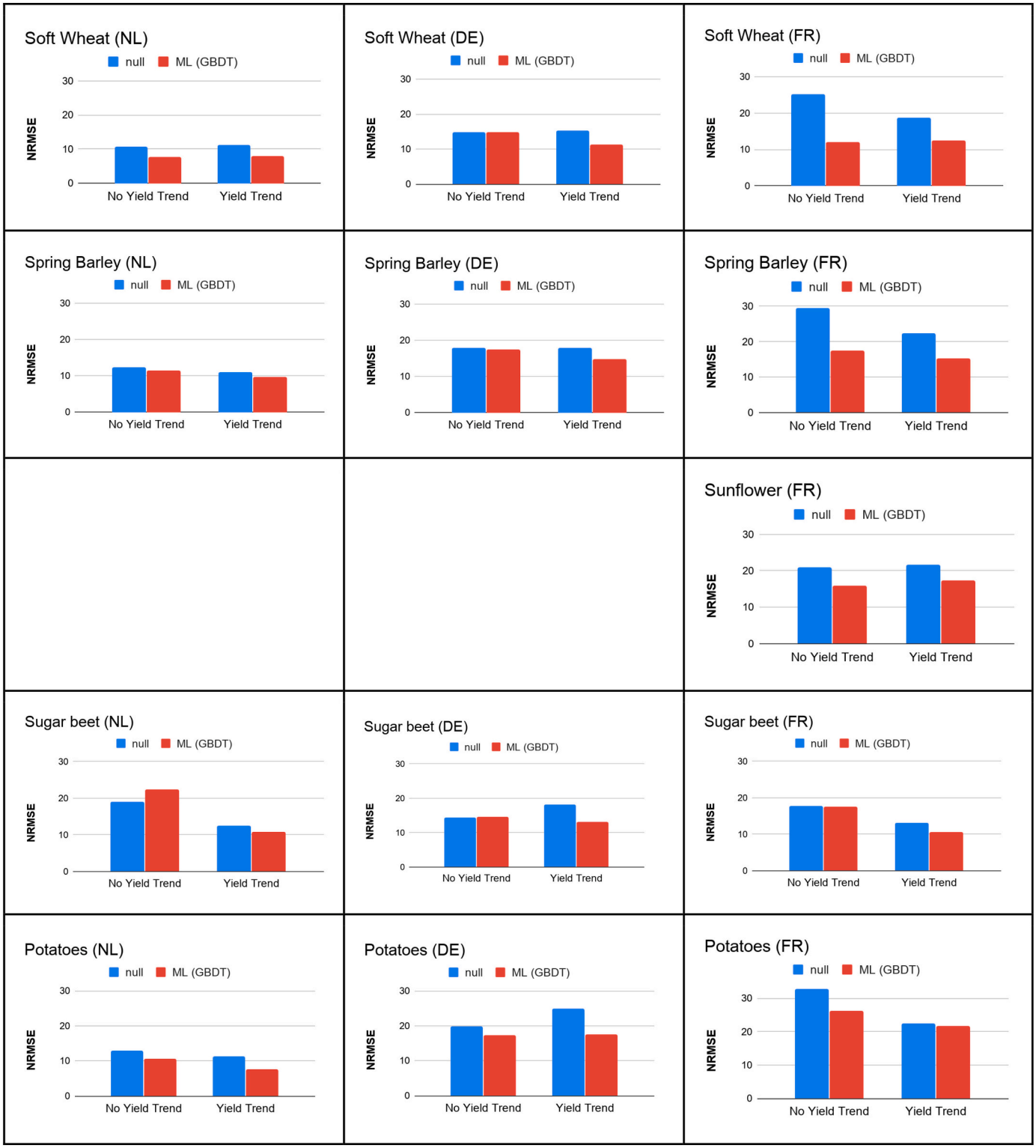


Fig. 5. Yield Trend vs. No Yield Trend.
The normalized RMSE of Gradient Boosted Decision Trees was compared with the null method.

selection to identify the most predictive ones. In this respect, we take a data-driven approach to learn the features that explain yield variability for each crop and country.

We ran the baseline to predict crop yield by applying supervised machine learning, which relies heavily on the size and quality of the data. In particular, a supervised learning algorithm is a good predictor when training labels are reliable and the training set is representative of the full dataset. We decided to predict crop yield at the sub-national level and combined data from different regions to ensure a sizable dataset. MCYFS forecasts are made at the national level and rely on crop yield statistics reported by European Union countries to Eurostat following the guidelines set out in the Annual Crop Statistics Handbook (Eurostat, 2019). Yield statistics at sub-national levels are not curated as often and vary across countries and crops (Lopez-Lozano et al., 2015).

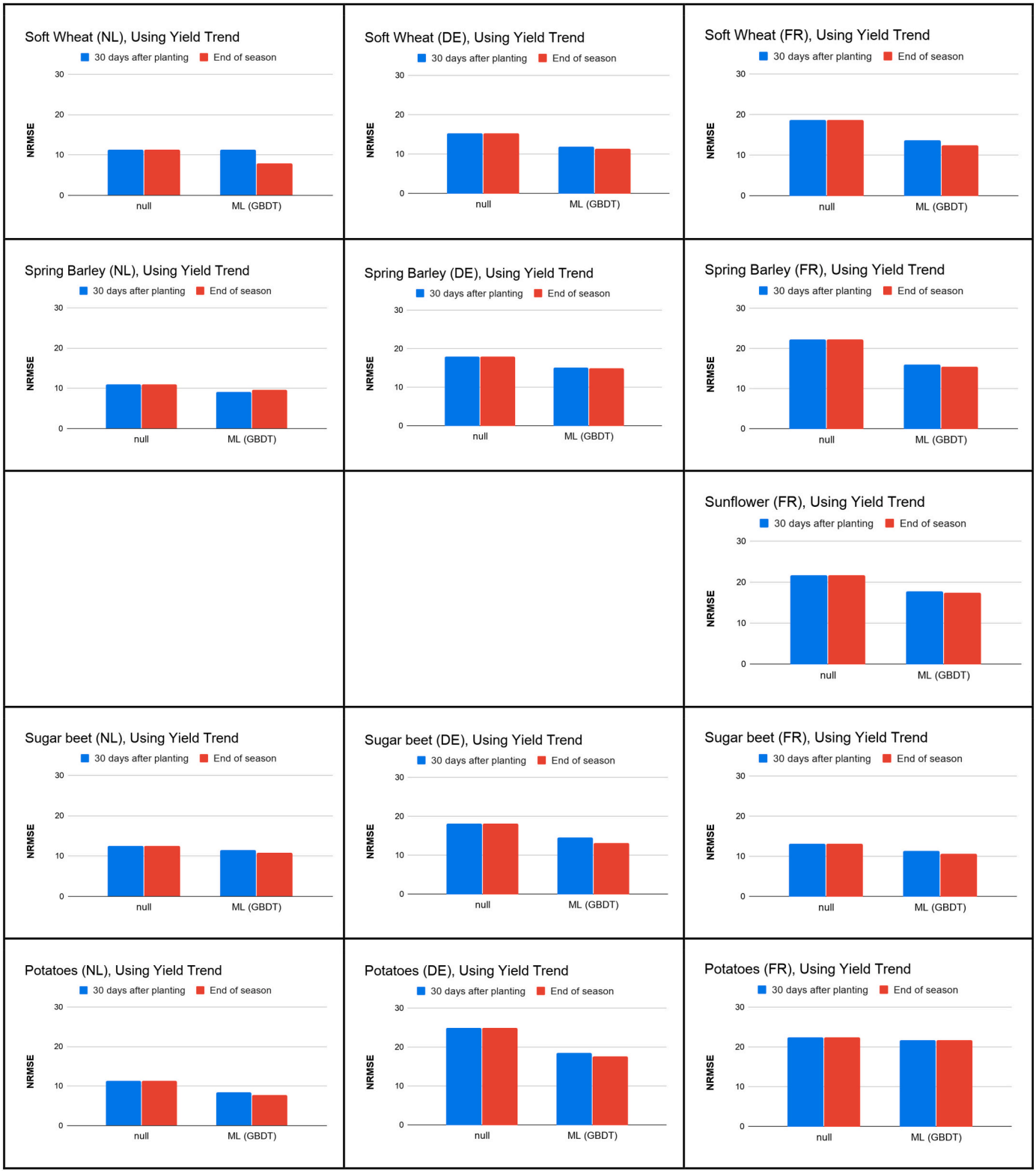


Fig. 6. Early season prediction using a 5-year yield trend.
The normalized RMSE of Gradient Boosted Decision Trees for early and end of season predictions.

Some regions have missing data and others have data copied from previous years. Thus, regional crop yield prediction illustrates the data size vs. data quality trade-off (e.g. see MAPE for potatoes (FR), *Fig. 14*). Nevertheless, the aggregated NUTS0 predictions of machine learning were promising, especially early in the season. In the case of NL (all four crops) and DE (spring barley, sugar beet, potatoes) and FR (soft wheat, spring barley, sunflower), the baseline's performance was comparable to MCYFS (see Fig. 7a; *Fig. 15a*). In terms of methodology, MCYFS uses data from all previous years to train models for the upcoming year (see Van der Velde and Nisini (2019)). In contrast, the machine learning baseline was trained with data up to 2011 or 2012, with predictions extrapolating up to 2018. Such differences in data and methods should be

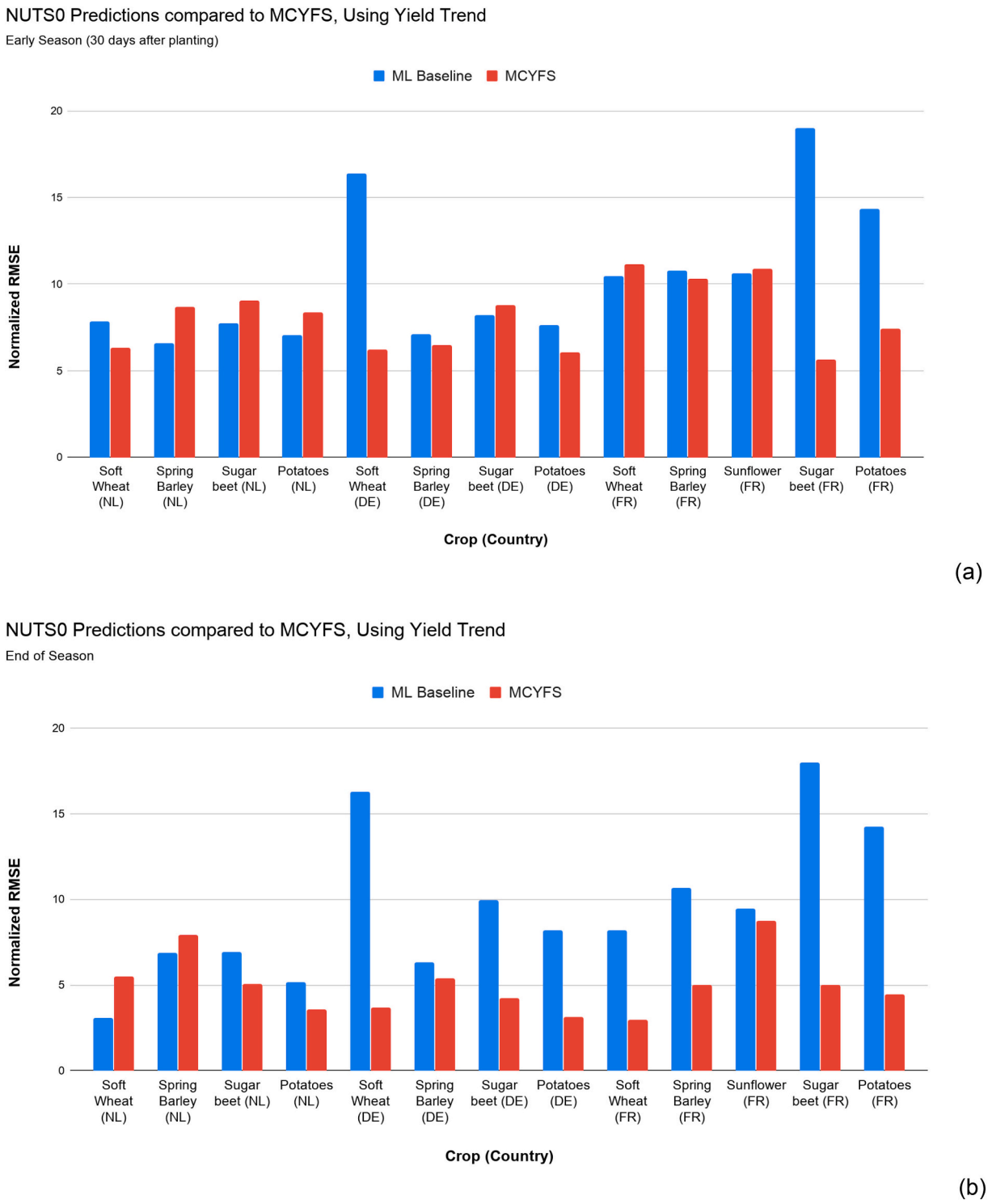


Fig. 7. Comparing machine learning baseline with past MCYFS forecasts.
Normalized RMSE for a) Early season predictions (30 days after planting), and b) end of season predictions, both using a 5-year yield trend.

considered when comparing the performance between the baseline and the MCYFS forecasts. Future research could investigate methods to address data quality and analyze the impact of different features, algorithms, hyperparameters and regularization methods to shed light into the potential of machine learning to improve crop yield predictions. Crop yield prediction at sub-national level may be a better approach for certain crops and countries where regional data is reliable. On one hand, the aggregated national yield forecasts could be more accurate and, on the other, the sub-national yield forecasts could also be useful for regional analysis. The machine learning baseline would serve as a starting point for such research.

As the present implementation of the baseline is based on MCYFS data, it can be directly used for crops and countries covered by MCYFS. Similarly, the baseline can be extended to scenarios where equivalent crop development and crop yield indicators (e.g. dry-weight yield biomass, leaf area, development stage) are available from other crop simulation models. Furthermore, Lopez-Lozano and Baruth (2019) have proposed a framework to extend MCYFS-style data and infrastructure to

the rest of the world. The machine learning baseline would be useful when data for the rest of the world is available in a similar format to MCYFS.

The baseline has ample room for improvement both in terms of the general design principles as well as fit-for-purpose optimizations. From our experience, the baseline could be improved in at least five ways. First, detection of outliers and duplicate data (particularly for yield statistics) could help improve the quality of training data. Second, the impact of different features, algorithms, hyperparameters and regularization methods could be analyzed to build a better optimized machine learning model. Third, new data sources could be added by applying appropriate data homogenization and preprocessing. Another consideration is feature design. Some data sources can be directly used as features; others require careful feature design. Fourth, certain additional data could make feature design more accurate. In the baseline, we infer the crop calendar for the whole country using WOFOST outputs. Crop calendar could be made per region, especially when the country covers multiple agro-ecological zones. More accurate sowing and harvest dates, phenological databases or remote sensing (see Alemu and Henebry 2016) could be used to define the crop calendar. Similarly, crop-specific thresholds could be used to define extreme conditions. Fifth, more advanced features could be designed to include weather or soil information from the previous years and to capture changes in cropping patterns.

The machine learning baseline has some technical limitations as well. First, the baseline does not have a generic method for data preprocessing. Data for certain crops and countries may need extensive preprocessing to fit the requirements of the baseline. Second, the baseline is not implemented for very big data analyses. Although we used Spark data frames for distributed preprocessing and feature design, we employed scikit-learn for feature selection and machine learning. Scikit-learn does not distribute data and computations when running multiple algorithms or when optimizing hyperparameters. The main reason for using scikit-learn instead of Spark machine learning library (Spark MLlib, https://spark.apache.org/mllib/) was feature selection. In the future, Spark MLlib may evolve to support the required functionality. In any case, future research could focus on running the machine learning part of the workflow in a distributed environment.

6. Conclusions

We designed a modular and reusable machine learning workflow for crop yield prediction and tested the workflow on thirteen case studies. Overall, we found that explainable features designed using principles of crop modeling can be used to predict crop yield at sub-national level. For early season predictions, the machine learning baseline performed similar to MCYFS in most cases. There was room for improvement as the season progressed. For crops and countries where regional data is reliable, sub-national yield prediction using machine learning is a promising approach going forward. Apart from addressing data quality issues, the baseline could be improved in three main ways: adding new data sources, designing more predictive features and evaluating different algorithms. The machine learning baseline serves as a starting point to explore the potential of machine learning for large-scale crop yield forecasting.

Data and software availability

Sample data for the Netherlands are available at DOI: https://doi.org/10.5281/zenodo.4312941 courtesy of the European Commission's Joint Research Centre (JRC).

The software implementation is available at: https://github.com/BigDataWUR/MLforCropYieldForecasting.

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 from 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 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 and Eurostat national yield statistics. We acknowledge D. Tuia from Geo-Information and Remote Sensing Group of Wageningen University and Research for insights on application of machine learning to crop yield prediction. We would like to thank M. van der Velde from JRC, P. Griffiths from Wageningen Into Languages and R. Fletcher from Wageningen School of Social Sciences for feedback on the manuscript text. We are thankful to Yiqing Cai from Gro Intelligence for the clarification on their method of crop yield prediction.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.agsy.2020.103016.

References

EC-JRC, 2020. JRC Agri4Cast Data Portal. https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx (Last accessed: May 11, 2020).

Eurostat, 2020a. Eurostat - agricultural production - crops. https://ec.europa.eu/eurostat/statistics-explained/index.php/Agricultural_production_-_crops (Last accessed: May 11, 2020).

Eurostat, 2020b. Eurostat - geographical information and maps. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts (Last accessed: May 11, 2020).

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

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

Alemu, W.G., Henebry, G.M., 2016. Characterizing cropland phenology in major grain production areas of Russia, Ukraine, and Kazakhstan by the synergistic use of passive microwave and visible to near infrared data. Remote Sens. 8, 1016. https://doi.org/10.3390/rs8121016.

Baskin, I.I., Marcou, G., Horvath, D., Varnek, A., 2017. Benchmarking machine- learning methods. In: Tutorials in Chemoinformatics, pp. 209–222. https://doi.org/10.1002/9781119161110.ch13.

Basso, B., Liu, L., 2019. Seasonal crop yield forecast: methods, applications, and accuracies. In: Advances in Agronomy, 154. Elsevier, pp. 201–255. https://doi.org/10.1016/bs.agron.2018.11.002.

Basso, B., Cammarano, D., Carfagna, E., 2013. Review of crop yield forecasting methods and early warning systems. In: Report Presented to the First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics, FAO Headquarters, Rome, Italy, 18–19 July.

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.

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.

Cai, Y., Moore, K., Pellegrini, A., Elhaddad, A., Lessel, J., Townsend, C., Solak, H., Semret, N., 2017. Crop yield predictions-high resolution statistical model for intra-season forecasts applied to corn in the US. In: 2017 Fall Meeting. Gro Intelligence Inc.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, 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.

Cerrani, I., Lopez 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 (Last accessed: Oct 8, 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. Agri- cultural and Forest Meteorology 206, 137–150. https://doi.org/10.1016/j.agrformet. 2015.03.007.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. Comput. Electron. Agric. 151, 61–69. https://doi.org/10.1016/j.compag.2018.05.012.

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 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. Syst. 168, 154–167. https://doi.org/10.1016/j.agsy.2018.06.018.

Dorigo, W.A., Zurita-Milla, R., de Wit, A.J., Brazile, J., Singh, R., Schaepman, M.E., 2007. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. Int. J. Appl. Earth Obs. Geoinf. 9, 165–193. https://doi.org/10.1016/j.jag.2006.05.003.

Fischer, R., 2015. Definitions and determination of crop yield, yield gaps, and of rates of change. Field Crop Res. 182, 9–18. https://doi.org/10.1016/j.fcr.2014. 12.006.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of Statistics 1189–1232. https://www.jstor.org/stable/2699986 (Last accessed: May 11, 2020).

GODAN, 2020. Global open data for agriculture and nutrition. www.godan.info (Last accessed: June 2, 2020).

Gonzalez Sanchez, A., Frausto Solıs, J., Ojeda Bustamante, W., et al., 2014. Predictive ability of machine learning methods for massive crop yield prediction. Span. J. Agric. Res. 12.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org (Last accessed: May 11, 2020).

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 Learn- Ing: 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.

Holzworth, D.P., Huth, N.I., deVoil, P.G., Zurcher, E.J., Herrmann, N.I., McLean, G., Chenu, K., van Oosterom, E.J., Snow, V., Murphy, C., et al., 2014. APSIM–evolution towards a new generation of agricultural systems simulation. Environ. Model Softw. 62, 327–350. https://doi.org/10.1016/j.envsoft.2014.07.009.

Holzworth, D.P., Snow, V., Janssen, S., Athanasiadis, I.N., Donatelli, M., Hoogenboom, G., White, J.W., Thorburn, P., 2015. Agricultural produc- tion systems modelling and software: current status and future prospects. Environ. Model Softw. 72, 276–286. https://doi.org/10.1016/j.envsoft. 2014.12.013.

Eurostat, 2019. Annual Crop Statistics Handbook. https://ec.europa.eu/eurostat/cache/metadata/Annexes/apro_cp_esms_an1.pdf (Last accessed: May 11, 2020).

Eurostat, 2016. Nomenclature of territorial units for statistics. https://ec.europa.eu/eurostat/web/nuts/background (Last accessed: eMay 11, 2020).

MARSWiki, 2020. MARS Crop Yield Forecasting System. https://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Welcome_to_WikiMCYFS (Last accessed: May 11, 2020).

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

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning, 112. Springer.

Janssen, S.J., Porter, C.H., Moore, A.D., Athanasiadis, I.N., Foster, I., Jones, J.W., Antle, J.M., 2017. Towards a new generation of agricultural system data, models and knowledge products: information and communication technology. Agric. Syst. 155, 200–212. https://doi.org/10.1016/j.agsy.2016.09. 017.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R., et al., 2016. Random forests for global and regional crop yield predictions. PLoS One 11, e0156571. https://doi.org/10.1371/journal.pone.0156571.

Jones, H.G., Vaughan, R.A., 2010. Remote Sensing of Vegetation: Principles, Techniques, and Applications. Oxford University Press.

Kamilaris, A., Prenafeta-Boldu, F.X., 2018. Deep learning in agriculture: a survey. Comput. Electron. Agric. 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016.

Kang, J., Schwartz, R., Flickinger, J., Beriwal, S., 2015. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. International Journal of Radiation Oncology*Biology*Physics 93, 1127–1135. https://doi.org/10.1016/j.ijrobp.2015.07.2286.

Kohonen, T., 2001. Self-Organizing Maps. Springer.

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. Agricultural Systems 168, 191–202.

Liakos, K., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: a review. Sensors 18, 2674. https://doi.org/10.3390/s18082674.

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.

Lokers, R., Knapen, R., Janssen, S., van Randen, Y., Jansen, J., 2016. Analysis of big data technologies for use in agro-environmental science. Environ. Model Softw. 84, 494–504. https://doi.org/10.1016/j.envsoft.2016.07.017.

Lopez-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. In: Agricultural and Forest Meteorology, 206, pp. 12–32. https://doi.org/10.3390/s18082674.

Mackowiak, S.D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., et al., 2015. Extensive identification and analysis of conserved small orfs in animals. Genome Biol. 16, 179. https://doi.org/10.1186/s13059-015-0742-x.

Muller, A.C., Guido, S., 2016. Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media, Inc.

Newlands, N.K., Zamar, D.S., Kouadio, L.A., Zhang, Y., Chipanshi, A., Toure, S., Hill, H. S., 2014. An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. Frontiers in Environmental Science 2, 17. https://doi.org/10.3389/fenvs. 2014.00017.

Pantazi, X.E., Moshou, D., Alexandridis, T., Whetton, R.L., Mouazen, A.M., 2016. Wheat yield prediction using machine learning and advanced sensing techniques. Comput. Electron. Agric. 121, 57–65. https://doi.org/10.1016/j.compag.2015.11.018.

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.

Phalan, B., Green, R., Balmford, A., 2014. Closing yield gaps: perils and possibilities for biodiversity conservation. Philosophical Transactions of the Royal Society B: Biological Sciences 369, 20120285. https://doi.org/10.1098/rstb.2012. 0285.

Schnepf, R., 2017. NASS and US Crop Production Forecasts: Methods and Issues. Technical Report. Congressional Research Service (Last accessed: May 11, 2020).

Shahhosseini, M., Martinez-Feria, R.A., Hu, G., Archontoulis, S.V., 2019. Maize yield and nitrate loss prediction with machine learning algorithms. Environ- mental Research Letters 14, 124026. https://doi.org/10.1088/1748-9326/ab5268.

Socher, R., Huval, B., Manning, C.D., Ng, A.Y., 2012. Semantic compositionality through recursive matrix-vector spaces. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Association for Computational Linguistics, pp. 1201–1211. https://doi.org/10.5555/2390948.2391084.

Statistics Canada, 2019. An Integrated Crop Yield Model Using Remote Sensing, Agroclimatic Data and Crop Insurance Data. https://www.statcan.gc.ca/eng/statistical-programs/document/3401_D2_V1 (Last accessed: Oct 8, 2020).

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 Algorithms., in: EUR Publication No. 15959 EN. Office for Official Publications of the European Communties, Luxembourg, p. 146.

Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. In: Proceedings of the National Academy of Sciences, National Academy of Sciences of the US, pp. 20260–20264. https://doi.org/10.1073/pnas.1116437108.

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., Baruth, B., Bussay, A., Ceglar, A., Condado, S.G., Lecerf, R., Lopez, R., Maiorano, A., Nisini, L., et al., 2018. In-season performance of European Union wheat forecasts during extreme impacts. Sci- entific Reports 8, 1–10. https://doi.org/10.1038/s41598-018-33688-1.

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.

Von der Malsburg, C., 1973. Self-organization of orientation sensitive cells in the striate cortex. Kybernetik 14, 85–100. https://doi.org/10.1007/BF00288907.

Willcock, S., Hooftman, D.A., Bagstad, K.J., Balbi, S., Marzo, A., Prato, C., Sciandrello, S., Signorello, G., Voigt, B., et al., 2018. Machine learning for ecosystem services. Ecosystem services 33, 165–174. https://doi.org/10.1016/j.ecoser.2018.04.004.

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemom. Intell. Lab. Syst. 2, 37–52. https://doi.org/10.1016/0169-7439(87)80084-9.

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 (Last accessed: May 11, 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 satellite imagery and crop modelling. Remote Sens. 12, 1024. https://doi.org/10.3390/rs12061024.