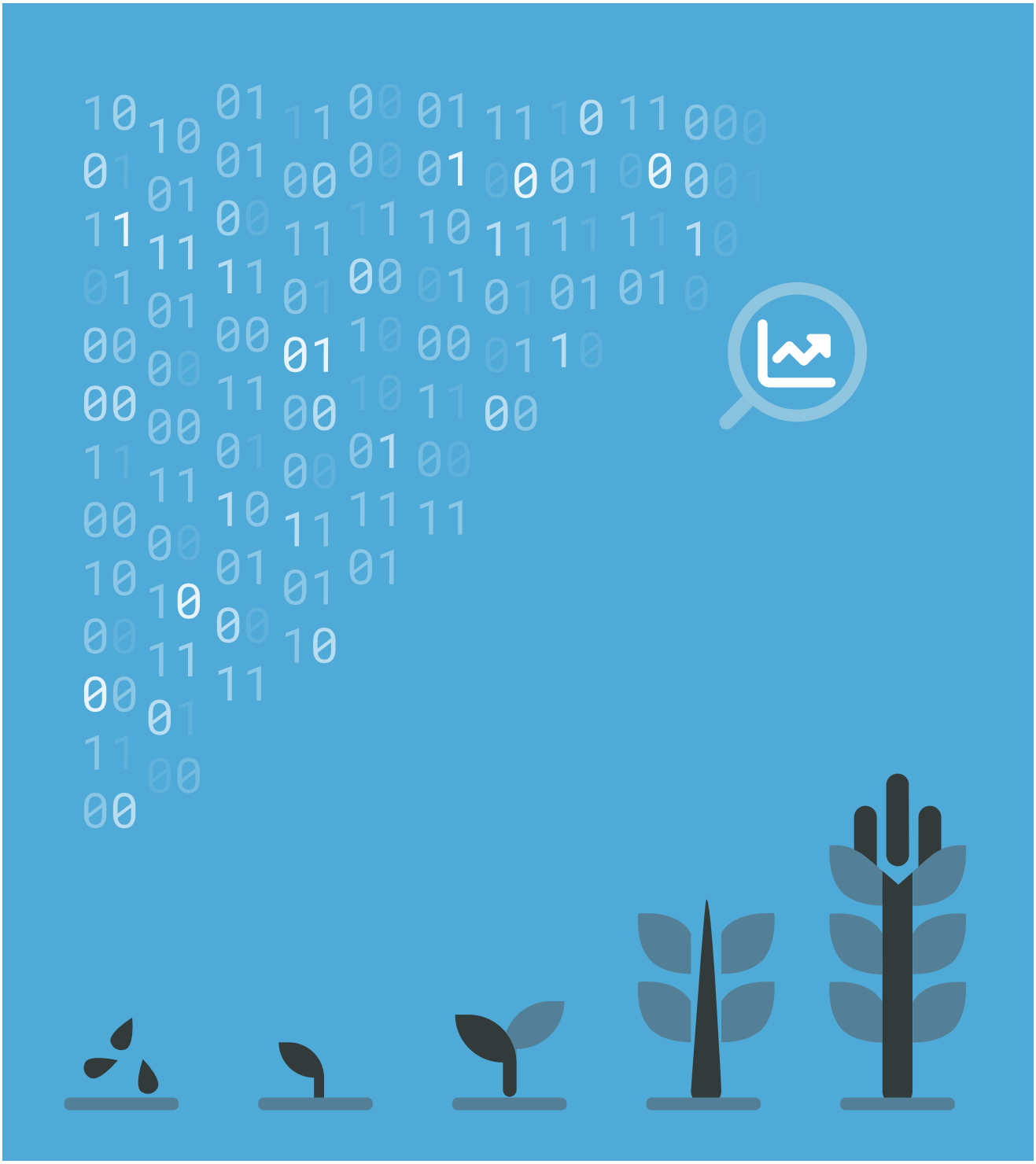
Dilli Raj Paudel

Machine learning for large-scale crop yield forcasting



Propositions

1. Artificial intelligence will not have ChatGPT-like success in agriculture.
   (this thesis)

2. Generic solutions and optimal solutions are conflicting goals in large-scale crop yield forecasting.
   (this thesis)

3. Increased automation is necessary for reproducibility of research.

4. Most models are more useful to modelers themselves than to end users.

5. Big data is a resource, but not a solution.

6. In nature, an individual is less important than the collective.

7. Sustainable choices will have a low priority until they become profitable.

8. Complacency is worse than embarrassment.

Propositions belonging to the thesis, entitled

Machine learning for large-scale crop yield forecasting

Dilli Raj Paudel
Wageningen, 13 June 2023

Machine learning for large-scale crop yield forecasting

Dilli Raj Paudel

Thesis committee

Promotors:
Prof. Dr I.N. Athanasiadis
Personal chair, Wageningen Data Competence Center
Wageningen University & Research

Prof. Dr B. Tekinerdogan
Professor of Information Technology
Wageningen University & Research

Co-promotors:
Dr S.A. Osinga
Assistant professor, Information Technology
Wageningen University & Research

Dr S.J.C. Janssen
Team leader, Earth Observation and Environmental Informatics
Wageningen University & Research

Other members:
Prof. Dr K.K.E. Descheemaeker, Wageningen University & Research
Prof. Dr B. Ganapathysubramanian, Iowa State University, USA
Prof. Dr R. Roscher, University of Bonn, Germany
Prof. Dr D. Cammarano, Aarhus University, Denmark

This research was conducted under the auspices of Wageningen School of Social Sciences (WASS).

Machine learning for large-scale crop yield forecasting

Dilli Raj Paudel

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 13 June 2023
at 1:30 p.m. in the Omnia Auditorium.

Dilli Raj Paudel
Machine learning for large-scale crop yield forecasting
210 pages.

PhD thesis, Wageningen University, Wageningen, NL (2023)
With references, with summary in English

ISBN 978-94-6447-599-9
DOI https://doi.org/10.18174/588095

Contents

Chapter 1

Introduction

1.1 Artificial intelligence and its role in agriculture

We are in the midst of an artificial intelligence revolution (Walsh, 2017; Makridakis, 2017). Artificial Intelligence, machine learning and deep learning power many aspects of our society (LeCun et al., 2015). Significant improvements have been made in computer vision, natural language processing and other complex tasks, such as autonomous driving, which require human-level intelligence (Ma & Sun, 2020). Artificial Intelligence (AI) means making machines behave in ways that would be considered intelligent if humans behaved in the same way (McCarthy et al., 1955). Machine learning is a subfield of AI and deep learning a subfield of machine learning (*Figure 1.1*, Goodfellow et al. (2016)). Four key drivers are responsible for the AI revolution: (i) easy access to computing power, (ii) growing amounts of data, (iii) research in machine learning and deep learning algorithms, and (iv) large amounts of funds pouring into the field (Walsh, 2017). Advances in AI and data-driven learning are transforming research and decision-making in many fields, such as healthcare (Bajwa et al., 2021), businesses (Dirican, 2015) and agriculture (Ruiz-Real et al., 2020).

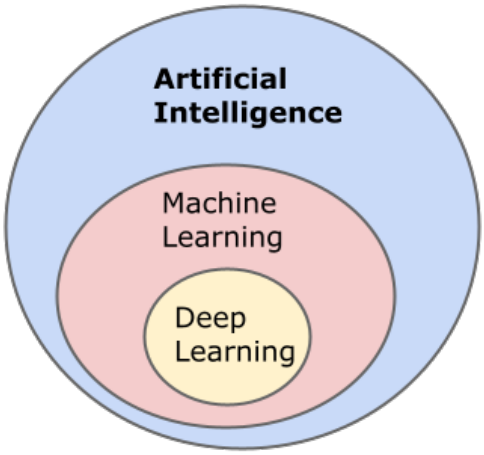


Figure 1.1: AI, machine learning and deep learning.

AI is considered crucial for monitoring and mitigating the impact of human activities on the environment, and making agriculture more efficient and resilient. Therefore, AI features prominently in the European Commission's Destination Earth (DestinE) initiative (European Commission, 2019) and the European Green Deal (European Commission, 2021). DestinE aims to build a highly accurate digital model of the Earth (called a digital twin) to monitor and predict the interaction between natural phenomena and human activities. The Green Deal aims to transform the European Union into a resource-efficient and climate-neutral economy. To facilitate data-driven innovation, the EU data strategy proposes the creation of European data spaces, including agricultural data space (European Commission, 2020). Similarly, the National Science Foundation (NSF) and the National Institute for Food and Agriculture (NIFA) of the US Department of Agriculture have funded AI institutes (NIFA, 2021) to address production, resilience and sustainability challenges in agriculture.

Challenges in agriculture stem from complex influences of agro-environmental and economic factors. AI technologies, such as robotics, computer vision, machine learning and digital twins, will be useful to find solutions to all aspects of agricultural production, including

optimal resource allocation and full or partial automation of farm processes (Smith, 2018; Dengel, 2013). Precision agriculture aims to maximize crop yields and farm profits while reducing environmental costs (Chlingaryan et al., 2018; Schieffer & Dillon, 2015) by ensuring optimal use of water, fertilizers, and phytosanitary products (Ruiz-Real et al., 2020). Smart farming goes one step further and incorporates informed decision making based on data and context-awareness (Sundmaeker et al., 2016). Smart farming is expected to bridge the gap between farming and AI (Chung et al., 2015). With the help of AI, food of higher nutritional value could be produced more efficiently in more stable supplies and with less environmental costs (Osinga et al., 2022). Therefore, AI and machine learning will play a key role in the twin goals of feeding a growing population while making food systems sustainable (Walter et al., 2017) and climate neutral.

Both precision farming and smart farming seek to improve crop yields, and accurate yield forecasts are crucial to achieve that. Horie et al. (1992) list three advantages of early season crop yield forecasts. First, yield forecasts are indispensable for food security planning. Second, farmers can adapt farm management practices, such as irrigation and fertilization, based on estimates of the final yield. Third, reliable yield forecasts enable farmers to make better marketing plans for their products. For these advantages to be realized, all stakeholders need access to consistent and unbiased yield forecasting models. Expected yields strongly influence the price of produce, and public availability of forecasting models will reduce information asymmetry among players in the commodity markets (Jiang et al., 2020). Machine learning and other AI technologies can help democratize access to data and yield forecasting models useful for farm management, market pricing, logistics and food security planning.

1.2 Machine learning

Mitchell (1997) provides a concise definition of machine learning: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." The experience can be a set of training data points, the task can be crop yield prediction and the performance measure can be accuracy, root mean square error (RMSE), etc. Machine learning makes predictions using predictors or features. Features are aggregations or summaries of input data for specific spatial and temporal windows. Feature design or feature engineering is the process of using prior (domain) knowledge to create features that influence or explain the variability in the labels or prediction targets, such as crop yield.

1.2.1 Supervised and unsupervised learning

Supervised machine learning (*Figure 1.2*) uses a training set of examples which includes features or predictors as well as the output label (e.g. crop yield) to learn a function which relates feature values to the labels. When each data point has a corresponding label, the model gets strong supervision from the label. When label data is not available for all data points at high resolution, learning is still possible using high resolution features and low resolution labels. High resolution forecasts can be aggregated to low resolution and compared with the labels. Such learning is an example of *weakly supervised* learning (Zhou, 2018).

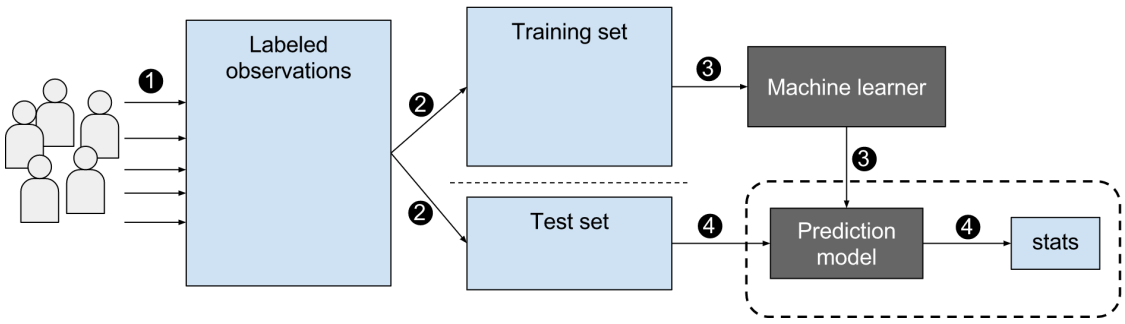


Figure 1.2: Supervised machine learning. Supervised machine learning identifies relationships between inputs and targets and learns to predict the target based on the inputs. 1. Collection of labeled data. Labeling often requires human experts. 2. Splitting data into training and test sets. 3. Training a model. 4. Evaluation of the learned model. *Image source*: WikiMedia Commons.

Unsupervised learning extracts some information from a set of examples which do not have output labels. One approach is to provide a meaningful grouping of examples (e.g. clustering). Another approach is to extract the most important factors of variation (e.g. dimensionality reduction). This thesis covers crop yield forecasting using supervised and weakly supervised methods.

1.2.2 Training and evaluation

Supervised machine learning algorithms are trained and evaluated on two disjoint sets. The common split is 70%-30% or 80%-20%, with the larger portion used as the training set and the smaller portion used as the test set. During training, it is important to avoid information leakage from the test set. Information leakage refers to the intentional or inadvertent use of test data (data or information about data) during the training stage. Training often involves iteratively improving the model by minimizing a loss function (e.g. mean squared error). Gradient-based methods are used to find the minimum or a reasonably low value of the loss function. Training is typically applied repeatedly to determine the optimal values of hyperparameters of an algorithm. Hyperparameters are parameters that are not learned from data by the algorithm (e.g. the number of neighbors in k-nearest neighbors algorithm). Hyperparameters are commonly optimized using k-fold cross validation or custom k-fold validation. During k-fold validation, the training set is further subdivided into training and validation sets to select the model and hyperparameters that perform best in the validation set.

After an algorithm is trained and optimized, it can be evaluated by making predictions on the training set and the test set. The error metric for predictions on the training set is called the training error; the error metric for predictions on the test set is called the test or generalization error. If the training error is high, the algorithm is said to have a high bias. If the training error is low, but the test error is high, the algorithm has a high variance. An algorithm is said to underfit if it does not fit the data well and has high training and test set errors. On the other hand, an algorithm is said to overfit if the training error is low but the test error is high. Algorithms that perform well on the test set are said to generalize well.

Regularization is a technique to improve the generalization error of an algorithm without hurting the training error. In other words, regularization decreases the variance of a model without significantly increasing the bias.

1.2.3 Model capacity

Machine learning algorithms try to approximate the function that relates predictor inputs to labels. Different algorithms have different capacities based on the set of functions they can model. Linear methods can only model linear functions. Linear Regression is the most common linear method. Support Vector Machines (Boser et al., 1992; Cortes & Vapnik, 1995) can model both linear and nonlinear functions. Decision trees (Quinlan, 1986) and their variants, e.g. Random Forests (Breiman, 2001), can model nonlinear functions.

1.2.4 Deep learning

A neural network is a simplified model of the biological neural network and typically consists of (i) an input layer where the data enters the network, (ii) one or more hidden layers where useful representation of the data is learned, and (iii) an output layer where the decision or prediction is made. Deep learning (LeCun et al., 2015) refers to using neural networks with many hidden layers to learn complex relationships between inputs and labels. Deep neural networks can automatically learn representations or features that explain the variation in labels, eliminating the need for manual or expertise-based feature design.

1.3 Crop yield forecasting

1.3.1 Definition and usefulness

Crop yield forecasting or crop yield prediction is the practice of predicting crop yields for a growing season prior to harvest based on inputs or predictors that influence crop yield. In contrast, the term crop yield estimation usually refers to estimation after harvest is completed (Basso & Liu, 2019). In precision agriculture, another similar term – crop yield monitoring – refers to monitoring spatial variability of crop yield in fields (Chung et al., 2016; Souza et al., 2016). Similarly, crop yield monitoring can also mean tracking the growth and development of crops in large geographical areas together with weather patterns and vegetation indices (Baruth et al., 2008; Rojas et al., 2005; Genovese et al., 2001). This thesis focuses on crop yield forecasting before harvest with the intent to understand the influence of predictors on crop yield and to provide useful information to decision-makers.

Reliable crop yield forecasts for a growing season are valuable to many stakeholders, such as farmers, commodity traders, logistics companies, policymakers and food aid programs (Basso & Liu, 2019; Schauberger et al., 2020). Forecasts can impact farm management practices, market information and food security measures. Farmers adjust farm management practices based on yield outlooks. Businesses and governments respond to conditions in commodity markets (USDA-NASS, 2012). Similarly, crop monitoring and yield forecasting contribute to the implementation of the European Union's Common Agricultural Policy (van der Velde et al., 2019). Furthermore, certain organizations, such as the International Grains Council

(IGC), use yield forecasts to improve cooperation in international trade and enhance market stability and food security (IGC, 2022).

Crop yield is affected by myriad factors that include crop-specific parameters, environmental conditions and management practices. The interactions among these factors are complex. Therefore, crop yield forecasting is a challenging task. Even then, crop yield forecasts are used by many stakeholders for strategic decisions. To support informed decision-making, forecasting models must be reliable. Model reliability depends heavily on interpretability, which is defined as the degree to which humans can understand the causes of a decision (Doshi-Velez & Kim, 2017; Miller, 2019). Therefore, yield forecasting models must be accurate and also interpretable in terms of causal relationships between predictors and crop yield.

1.3.2 Crop yield forecasting methods

Crop yields are commonly forecasted using five methods and their combinations: (i) field surveys, (ii) process-based simulation models, (iii) remote sensing (iv) statistical models, and (v) machine learning.

Field surveys try to capture the state of crops in the field and the expected production by means of phone interviews, farmer reports and objective yield measurements (USDA-NASS, 2012). The National Agricultural Statistics Service (NASS) of the US Department of Agriculture (USDA) relies on field surveys to produce forecasts of many crops, including corn and soybean. Farmers are selected to participate in surveys based on farm size and intensity, with large and high intensity farms sampled with higher frequency (USDA-NASS, 2012). Objective yield measurements employ professional enumerators to visit selected fields and prepare samples for data collection. These surveys face many reliability concerns related to sampling methods, response rates and data processing errors (Schnepf, 2017; Chipanshi et al., 2015). Statistics Canada is gradually moving away from surveys to modeling that uses remote sensing, agro-climatic and crop insurance information to forecast crop yields (Statistics Canada, 2020).

Biophysical process-based crop models take crop parameters, weather, soil conditions and farm management practices as input to simulate crop growth, development and yield. The outputs of crop models include above-ground biomass, storage organs biomass and leaf area index. The agronomic principles that underlie crop models apply across space and time (Basso et al., 2013). As a result, crop models, such as WOFOST (van Diepen et al., 1989; de Wit et al., 2019), DSSAT (Jones et al., 2003), STICS (Brisson et al., 2003), APSIM (Holzworth et al., 2014), LINTUL (Kooman & Haverkort, 1995), and CERES (see Ritchie et al. (1998)), are commonly used in yield forecasting. The MARS Crop Yield Forecasting System (MCYFS) of the European Commission relies on WOFOST outputs. APSIM simulations have been used to predict crop yields in the US and Australia (Shahhosseini et al., 2019; Feng et al., 2019). Crop models have been criticized because of their considerable data and calibration requirements (Basso et al., 2013; Grassini et al., 2015; García-León et al., 2020). Furthermore, these models do not always capture the impact of certain factors, such as pests, diseases and nutrient limitations.

Remote sensing provides indirect indicators of crop yield by measuring observed radiance

or reflectance using satellites, airplanes and unmanned vehicles or drones. Remote sensing products need to be processed to extract vegetation indices, such as the normalized difference vegetation index (NDVI) (Doraiswamy et al., 2004; Johnson, 2014), and passed to statistical methods to predict crop yield. Vegetation indices express mathematical relations among mainly red, green and infrared spectral bands and capture functional relationships between crop characteristics and remote sensing observations (Basso et al., 2013; Wiegand et al., 1991). Remote sensing data is openly available and covers large areas (Chipanshi et al., 2015). Similarly, data is increasingly available at higher resolution. Therefore, remote sensing is becoming popular in high-resolution crop yield forecasting (Donohue et al., 2018; Lobell et al., 2015).

Statistical models generally model linear relationships between predictors (independent variables) and crop yield (dependent variable) and identify the most important predictors that explain the variance in yield. The predictors could include survey results, crop model simulation outputs, soil characteristics, weather variables and remote sensing products. Statistical analyses often account for a yield trend that captures effects of technological advancements in genetics and changes in farm management, such as mechanization and fertilization (Basso et al., 2013). For example, the MCYFS method assumes that yield variability is composed of a yield trend and the residual from the trend, depending on inter-annual climate variability (Lobell, 2010; Lecerf et al., 2019). MCYFS analysts run multiple linear regression to fit models to determine the impact of crop model outputs, weather observations and remote sensing indicators on crop yield (Lecerf et al., 2019; van der Velde & Nisini, 2019).

Machine learning takes a data-driven approach to crop yield forecasting. Similar to statistical models, machine learning methods commonly use agro-environmental variables together with outputs of surveys, crop models and remote sensing as inputs. Standard machine learning involves an expertise-based step to summarize input data spatially and temporally into predictors or features. Machine learning algorithms can learn nonlinear relationships between features from diverse data sources and crop yield.

1.4 Large-scale crop yield forecasting

In this thesis, large-scale crop yield forecasting means scaling yield forecasting to large and diverse areas in two dimensions (*Figure 1.3*): from one country and crop to multiple countries and crops at continental scale, and from country-level forecasts to sub-national and grid-level forecasts.

1.4.1 State of the art systems

Large-scale yield forecasting systems around the world use some combination of methods described in *Section 1.3.2*, but not machine learning. Here we briefly describe three systems from Europe and North America. Other large-scale systems, such as the China CropWatch (Wu et al., 2014) and the Mahalanobis National Crop Forecast Centre of India (Mahalanobis Centre, 2022), share some similarities with those described.

The European Commission's Joint Research Centre Directorate for Sustainable Resources,

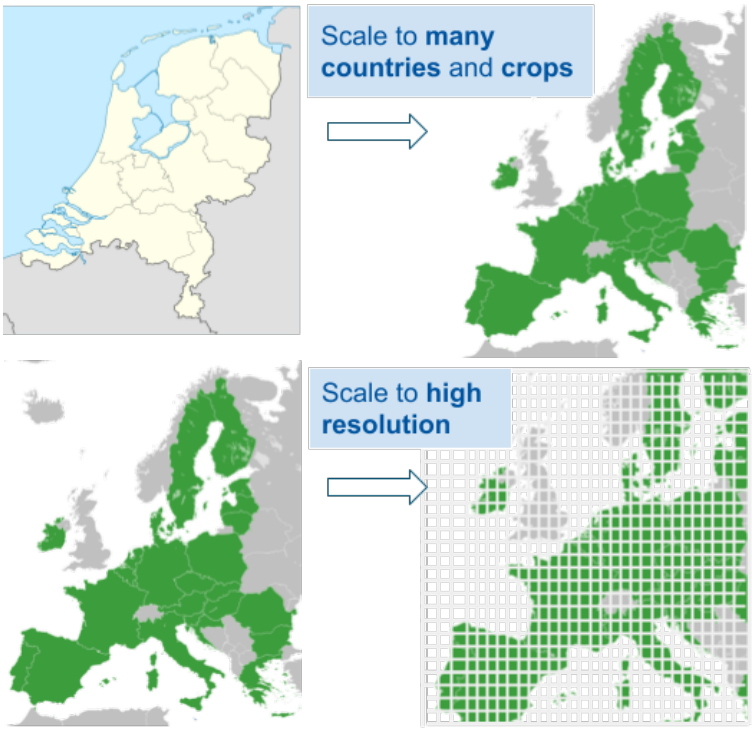


Figure 1.3: Large-scale crop yield forecasting. Large scale means can mean application to many crops and locations and to high resolution. *Map source*: WikiMedia Commons.

and specifically the Food Security Unit, uses the MARS Crop Yield Forecasting System (MCYFS) to monitor crop growth and weather events and to forecast crop yields. MCYFS includes two main components: i) the Crop Growth Monitoring System or CGMS (Supit & van der Goot, 2003) which contains a suite of crop models and a tool-box for data processing, and ii) the statistical Control Board or CoBo (Genovese & Bettio, 2004) which facilitates the analysts to run statistical analysis and to output crop yield forecasts. MCYFS analysts build interpretable models to produce forecasts for many crops and countries, but the forecasts are limited to national level. The analyst-driven approach is difficult to scale to higher resolutions.

The National Agricultural Statistics Service of the US Department of Agriculture conducts grower-reported surveys, aka Agricultural Yield Surveys and Objective Yield Surveys to collect data for yield forecasting (USDA-NASS, 2012). Sampling of farms is based on probability proportional to size (PPS), which essentially means farms with large size and high contribution to total state yield are sampled with higher probability. During objective yield surveys, trained enumerators verify the reported acreage, select plants for precise measurements and finally collect counts and measurements of plant characteristics related to yield, such as number of fruits and weight per fruit. Linear statistical models are used to calculate the county, state and national level yields. One limitation of the NASS method is its reliance on surveys. In addition to the dependence on human observations and judgment, surveys suffer from reliability concerns related to sampling methods, declining responses and data processing errors (Schnepf, 2017; Chipanshi et al., 2015). Another limitation of NASS forecasts is that early season forecasts are made monthly at state level, but not county level; county yields are published only at the end of season.

Integrated Crop Yield Model of Statistics Canada relies on remote sensing, agroclimatic indicators and crop insurance data passed to a multivariate linear model to forecast crop yields (Statistics Canada, 2020). The model is constructed using the historical relationships between the yields reported to Manitoba Agricultural Services Corporation (MASC) by the farm operator at the end of the growing season and the NDVI and agroclimatic measurements taken at different times during the growing season as well as the yield trend. Data from the previous ten growing seasons are used in deriving the model. Crop yield forecasts are produced at Census Administrative Region (CAR) or province level three times a year (early season, mid season and end of season) (Statistics Canada, 2020).

1.4.2 Challenges and opportunities

Large-scale crop yield forecasting has several challenges and opportunities. First, steps to build and validate models need to be automated. Methods that rely heavily on human interventions do not work well at large scale due to potential for errors, reproducibility concerns and time limitations. Second, the tools and workflows used must be modular and reusable. The steps involved need to be reproducible to verify their correctness. Modularity and reusability allow the processes and workflows to evolve and improve over time. Third, models must still be interpretable. The stakeholders of large-scale crop yield forecasting are still human decision-makers. They need to understand model behavior to make informed decisions. Large scale does provide opportunities to learn from growing amounts of data. However, good quality data must be available for models to work. Otherwise, models must be able to handle missing inputs or labels.

With the increasing availability of large amounts of data, technologies exist to process them and analyze them. These technologies make large-scale crop yield forecasting a viable goal. Many companies, such as Amazon, Microsoft and Google, provide easy-access computing resources in the cloud. Computing instances or clusters can be equipped with distributed data processing tools, such as Spark (Zaharia et al., 2016) and MapReduce (Dean & Ghemawat, 2008), and machine learning or deep learning packages, such as scikit-learn (Pedregosa et al., 2011) and PyTorch (Paszke et al., 2019). These technologies continue to develop further, as more experience, algorithms and best practices become available (Osinga et al., 2022). In general, we are not restricted so much by technology, but by our ability to extract insights about how different predictors influence crop yield despite limitations of data and models.

1.4.3 Machine learning for large-scale crop yield forecasting

Operational large-scale yield forecasting systems commonly build linear statistical models using outputs of crop models, remote sensing and surveys. Similarly, Lobell et al. (2015) combined crop model simulations, remote sensing data and linear regression to build a scalable crop yield mapping tool for corn and soybean in the US. While linear models are simple and interpretable, they do not always capture complex relationships between predictors and yield. Combining the strengths of crop models, remote sensing and machine learning is a promising solution to forecast crop yields at large scale. Crop models provide agronomic information, remote sensing provides crop state with increasing detail and machine learning

captures complex relationships among diverse predictor inputs and yield (Chlingaryan et al., 2018).

Apart from the capacity to model complex relationships, machine learning algorithms have additional benefits. First, their performance generally improves when more training data is available. Second, almost all steps of a machine learning workflow can be automated. In standard machine learning, designing features from input data requires expert knowledge. Deep learning methods can automate feature extraction as well. This increased automation will enable machine learning and deep learning methods to produce high resolution crop yield forecasts for multiple crops and countries. Higher resolution forecasts are desirable for targeted application of agricultural policies (García-León et al., 2020; You et al., 2014). In addition, machine learning workflows can standardize the process of producing forecasts where expertise-based approaches vary from person to person. Finally, data-driven learning also has the potential to produce insights that human analysis may have missed.

Machine learning models and forecasts need to be both accurate and interpretable for stakeholders to make strategic decisions. Therefore, existing knowledge and expertise can be incorporated into certain steps of machine learning workflows, such as feature design. Some machine learning methods, such as linear regression and algorithms based on decision trees, are inherently interpretable. In the case of deep learning, feature attribution methods (Montavon et al., 2018; Ancona et al., 2018; Lundberg & Lee, 2017) can be used to learn explanations based on a post-hoc analysis of model predictions. Despite expected benefits, accuracy and uncertainty of machine learning predictions will depend heavily on data quality and information content of predictor inputs. Regularization techniques and cross-validation schemes can be used to prevent overfitting when data quality is a concern. Deep learning combined with weak supervision can help when labels are missing at high resolution but are available at low resolution. Overall, machine learning and deep learning methods provide a set tools that have not been explored fully in large-scale crop yield forecasting.

The challenges of large-scale crop yield forecasting have not been addressed sufficiently by existing applications of machine learning. Most current studies focus on specific crops or locations (Pantazi et al., 2016; Cai et al., 2019), and their work may not generalize to other settings. Solutions that work at large scale will have to be generic, modular and reusable. The data-driven and automated approach of machine learning, paired with modular design and configuration options, will be crucial in scaling crop yield forecasting to many locations and higher resolutions. Machine learning can provide benefits that complement expertise-based crop yield forecasting especially at large scale. Machine learning methods will also face challenges related to data size and quality, applicability to many crops and locations, and reliability of model predictions, both in terms of accuracy and interpretability.

1.5 Research objectives

The main objective of this thesis is to investigate the benefits and challenges of using machine learning for large-scale crop yield forecasting. This objective is further divided into four sub-objectives:

1. ***A baseline for large scale***: Design a generic explainable, modular and reusable

machine learning workflow to forecast crop yields for multiple crops and countries.

2. ***Forecasting at scale***: Evaluate the benefits and limitations of machine learning to produce regional crop yield forecasts for many crop and countries in Europe.

3. ***Feature learning at scale***: Assess the performance and interpretability of deep learning models for crop yield forecasting.

4. ***Data requirements at scale***: Design a framework to produce high resolution crop yield forecasts when high resolution yield data are unavailable.

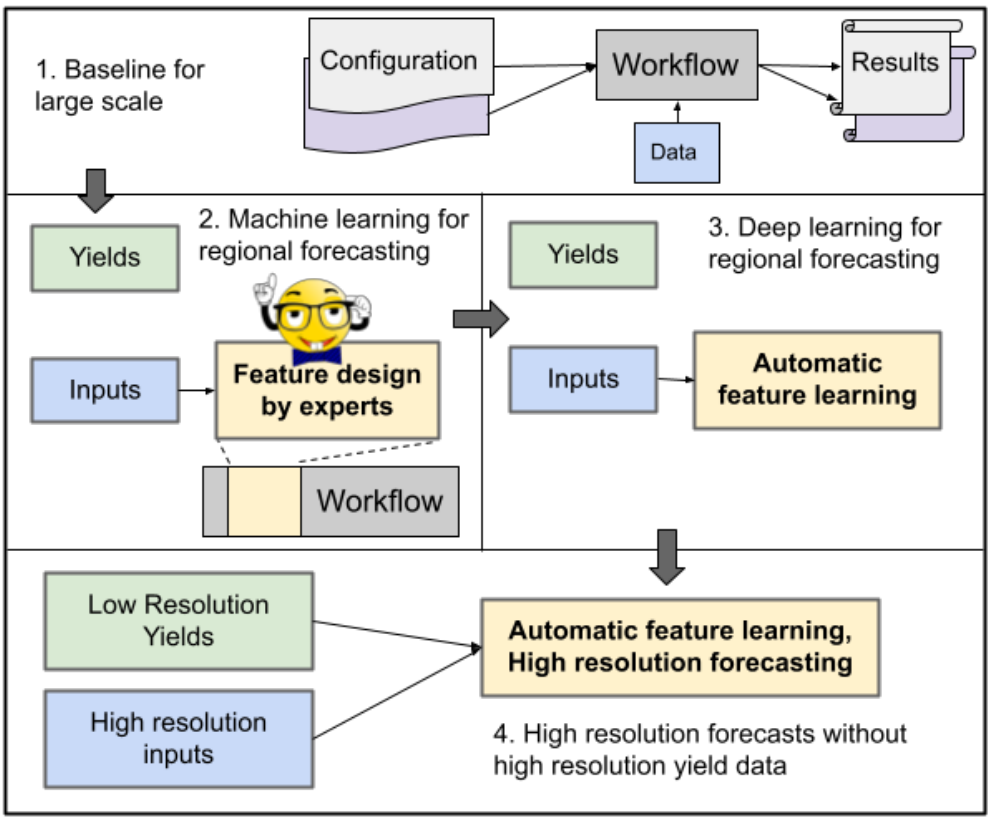


Figure 1.4: How the sub-objectives of the thesis fit together. We start with a generic workflow for machine learning. Then we focus on three other challenges of large-scale crop yield forecasting: application at regional level to multiple crops and countries; automatic feature learning and interpretability; and high resolution forecasts without high resolution labels.

Figure 1.4 shows how the sub-objectives are related to each other. Sub-objective 1 focuses on a generic, modular and reusable workflow that is based on agronomic principles and produces results that serve as the baseline for further improvements. Sub-objective 2 improves the workflow and tests its scalability to many crops and locations in Europe. Sub-objectives 1 and 2 both rely on experts for feature design. Sub-objective 3 uses deep learning to automate feature extraction and evaluates the interpretability of features learned. Sub-objective 4 addresses data requirements at high resolution by producing high resolution forecasts in the absence of high resolution yield data.

1.6 Thesis outline

The rest of this thesis is structured as follows: **Chapter 2** describes a generic machine learning workflow for large-scale crop yield forecasting emphasizing three design principles: (i) correctness, (ii) modularity, and (iii) reusability. For correctness, explainable features were designed based on agronomic knowledge, and data was split into training and test sets in a way to prevent information leakage. Different components of the workflow were kept modular to allow each component to evolve and improve without affecting other components. The workflow could be reused for multiple crops and countries by setting the crop and country in configuration options. **Chapter 3** improved the baseline from *Chapter 2* in several ways, including a dynamic (per region, per year) crop calendar and a more robust Bayesian hyperparameter optimization. The improved workflow was used to investigate the benefits and limitations of machine learning for regional crop yield forecasting in Europe. **Chapter 4** replaces expertise-based feature design with neural networks and evaluates performance and interpretability of deep learning models. Performance was compared with a standard machine learning method trained using expert-designed features. For interpretability, feature importance scores extracted from deep learning models were analyzed by a group of human experts. **Chapter 5** tackles crop yield forecasting when predictor inputs are available but yield data are unavailable at high resolution. A weakly supervised deep learning framework was designed to get supervision signals from low resolution yield data and produce forecasts for both high and low resolution. **Chapter 6** reviews and reflects on the findings from *Chapter 2* to *Chapter 5* and outlines directions for future research.

Chapter 2

A machine learning baseline for large-scale crop yield forecasting

This chapter is based on:

Dilli Paudel, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjoukje Osinga, Christos Pylianidis, and Ioannis N Athanasiadis. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187:103016, 2021. doi:10.1016/j.agsy.2020.103016

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.

2.1 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 & 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 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 (López-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 chapter, 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 & 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, 2022). 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, 2022) and Eurostat data (Eurostat, 2020a,b). We ran experiments to predict early season and end of season crop yield at NUTS2 or NUTS3 level. NUTS (Nomenclature of Territorial Units for Statistics) is a hierarchical system of dividing the territory of the European Union for statistics and policy (Eurostat, 2016b). 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 chapter is organized as follows: *Section 2.2* reviews related work in the field; *Section 2.3* describes the methodology and the case studies; *Section 2.4* presents the results; *Section 2.5* discusses our findings and areas for further research, *Section 2.6* summarizes our conclusions. *Appendix A* provides additional details about data and implementation, and also supplementary results.

2.2 Related Work

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 (Liakos et al., 2018; Kamilaris & Prenafeta-Boldú, 2018; Chlingaryan 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., 2014) 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 González Sánchez 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); González Sánchez 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.

Operational large-scale crop yield forecasting systems, such as MCYFS, NASS and Statistics Canada, build statistical models using weather observations, field survey results, crop growth model outputs, remote sensing indicators and yield statistics, but do not use machine learning. 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. (2015). 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 underlying science (Holzworth et al., 2015). 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.

2.3 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 (*Figure 2.1*). 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 2.3.1*). 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.

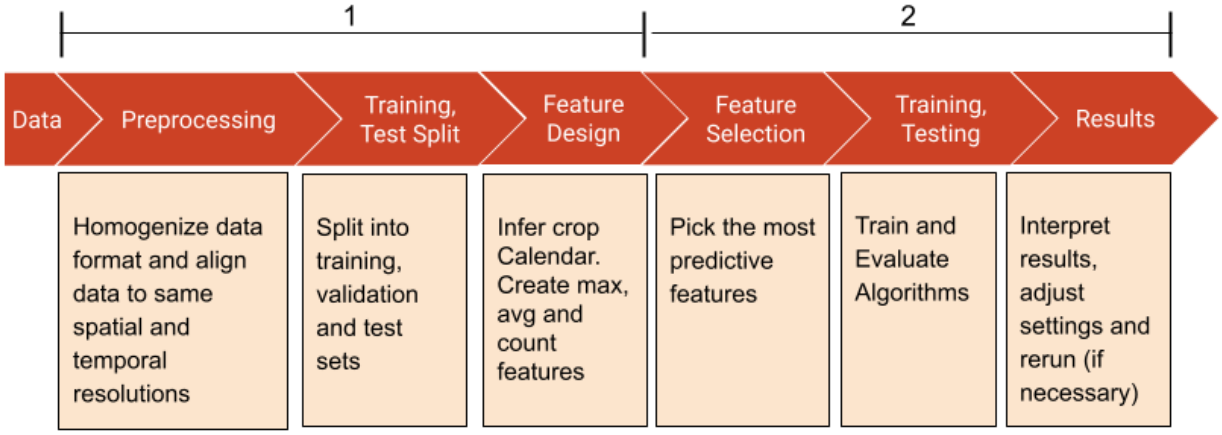


Figure 2.1: The high-level workflow. The overall workflow has two parts. The first part includes preprocessing and feature design. The second part includes machine learning.

2.3.1 Workflow Design: Correctness

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

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 (van Diepen et al., 1989; Supit et al., 1994), we selected 3 dekads (10-day periods) when significant changes occur in the crop's development stage (DVS): (i) START_DVS (DVS >= 0) is when the crop emerges from the soil, (ii) START_DVS1 (DVS >= 100) is the middle of the flowering phase, and (iii) START_DVS2 (DVS >= 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) preplanting window, (ii) planting window, (iii) vegetative phase, (iv) flowering phase, (v) yield formation phase, and (vi) harvest window (*Table 2.1*).

Table 2.1: Crop calendar definition. 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.

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

Table 2.2: Feature design using crop modeling principles. 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).

Period	**Maximum Values**	**Average Values**	**Days/dekads with extreme values**
Pre-planting		TAVG, PREC, CWB	
Planting		TAVG, PREC	RSM, TMIN, PREC
Vegetative	WLIM_YB, TWC, WLAI	RSM, TAVG, CWB, FAPAR	RSM
Flowering		PREC	RSM, PREC, TMAX
Yield Formation	WLIM_YB, WLIM_YS, TWC, WLAI	RSM, CWB, FAPAR	RSM
Harvest		PREC	PREC

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 2.2*). Using these indicators, we designed 3 types of features: (i) maximum 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 A.2* 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 A.1.2*) 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. López-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.

Machine Learning without Information Leakage

We applied supervised learning, 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 relating 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 (*Figure 2.2a*). This restriction was necessary because later years contained 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 2.3.5*). Therefore, we added every nth 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 2.1*) and the averages and standard deviations of the indicators shown in *Table 2.2*. We inferred the crop calendar and calculated indicator statistics only using the training set.

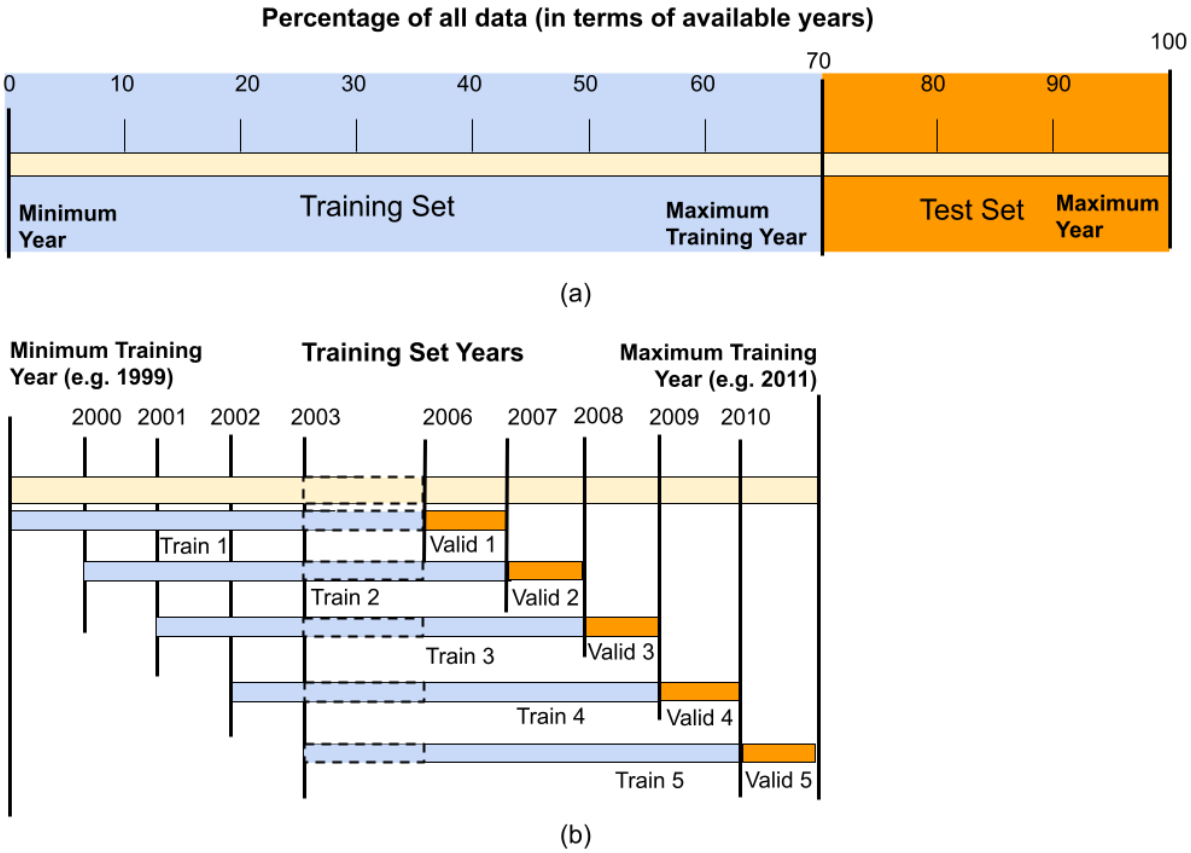


Figure 2.2: Training, validation and test splits when using yield trend (a) For each region, we split the full dataset into training and test sets. (b) We further divided the training set into validation training and test sets for feature selection and hyperparameter optimization using a time-based 5-fold sliding validation.

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 (*Figure 2.2b*). For example, NL data was available from 1999 to 2018. The training data

included 1999 to 2011, and the test data from 2012 to 2018. During 5-fold sliding validation, the first iteration used 1999 to 2006 for training and 2007 for validation, the second iteration used 2000 to 2007 for training and 2008 for validation, the third iteration used 2001 to 2008 for training and 2009 for validation, and so on. When not using the yield trend, we applied regular k-fold cross-validation.

We created pipelines consisting of feature scaling, feature selection and training stages (see *Section A.1.2*) to avoid information leakage during feature selection and training (Müller & 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.

2.3.2 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 (*Figure 2.3*). 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.

We defined configuration options to control data flow when running various experiments (*Figure 2.4*). 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 2.3.5*).

We defined feature selection and prediction algorithms in a modular and extensible manner to enable experimentation with different algorithms (*Figure 2.4*). 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.

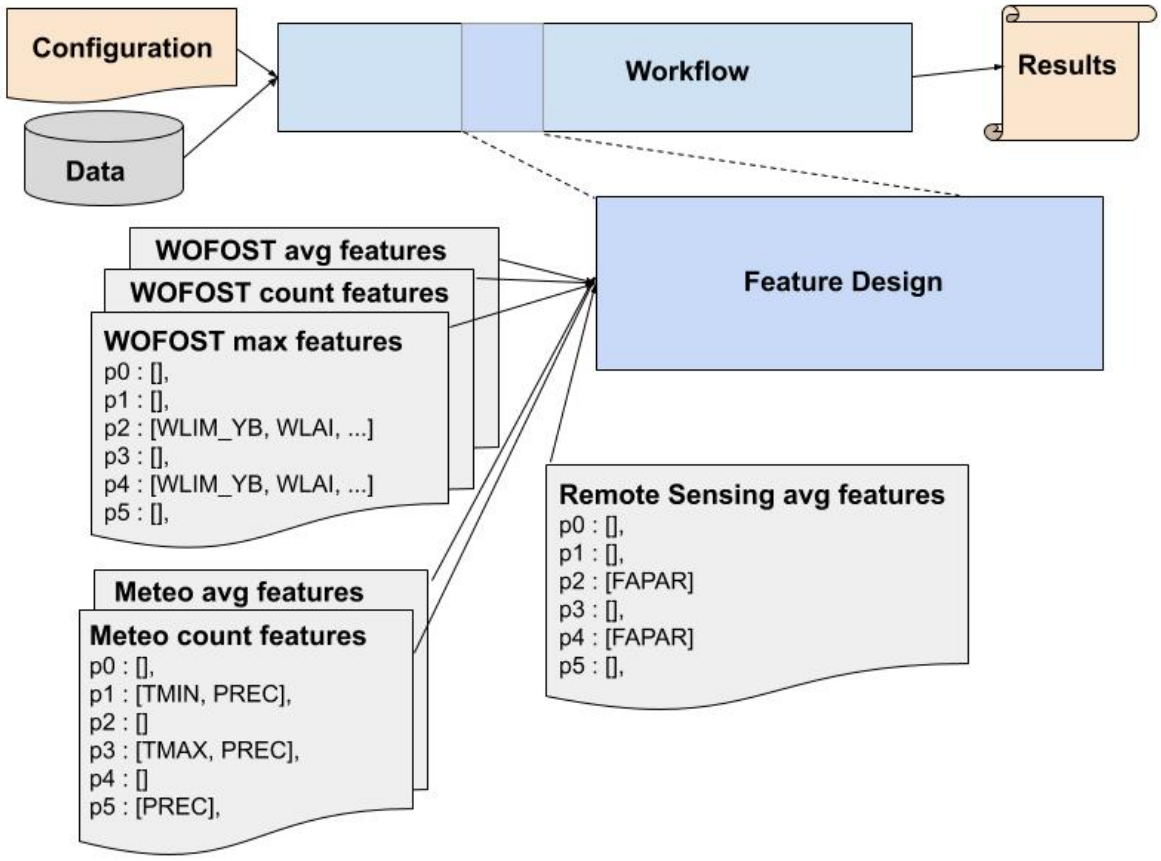


Figure 2.3: Modularity and extensibility in feature design. Features were designed using extensible lists of indicators for each crop calendar period. Lists of indicators correspond to entries in *Table 2.2*.

2.3.3 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 2.3.1*). 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 for different crops, countries and NUTS levels (*Figure 2.4*). We set most configuration options to reasonable defaults to avoid specifying all of them for every experiment.

2.3.4 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 1999 to 2018, FR data for 101 NUTS3 regions from 1999 to 2018 and DE data for 401 NUTS3 regions from 1999 to 2018. As described in *Section 2.3.1*, we used 70% of the data for training and 30% for testing. *Section A.2* provides more details about the data. We did not use region centroids by default because it was unclear whether they provided

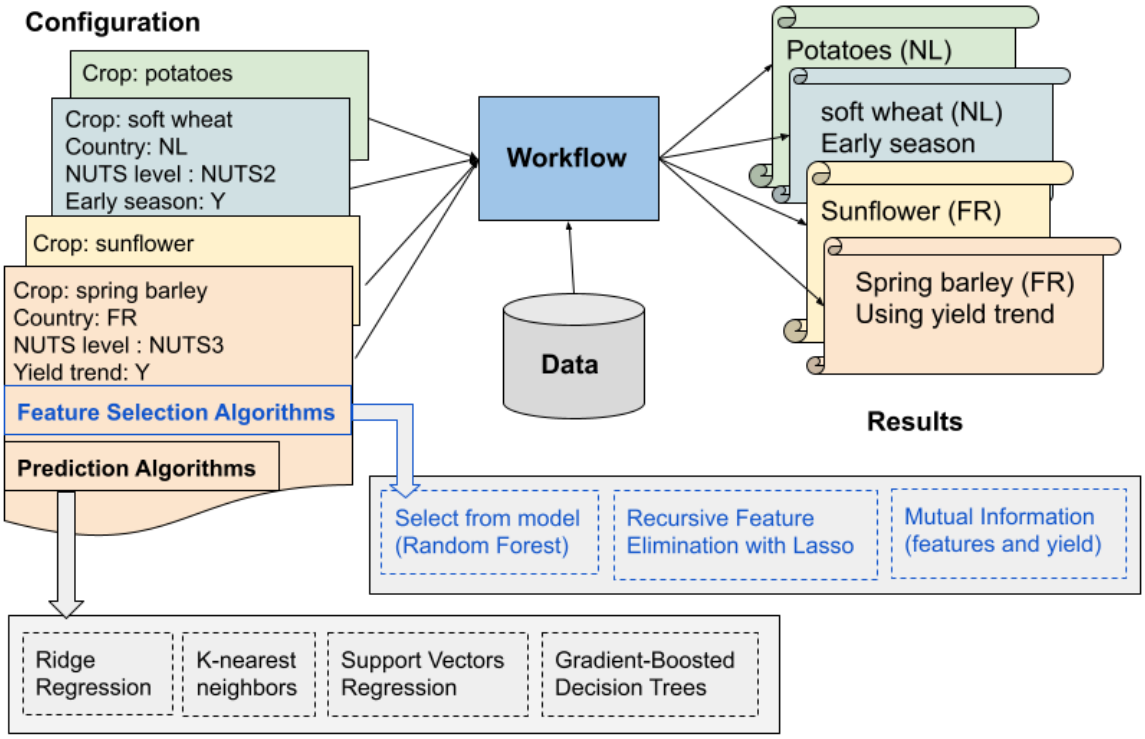


Figure 2.4: Configuration Options. Configuration options were used to select the case study and the experiment being run. Feature selection algorithms and prediction algorithms were defined using extensible data structures. Therefore, different algorithms could be added or removed to study their benefits without affecting the workflow. (see *Section A.1.2* for more details about the algorithms.)

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.

Four machine learning algorithms were used to predict the crop yield: (i) Ridge Regression (Hoerl & Kennard, 1970), (ii) K-nearest Neighbors Regression (Cover & Hart, 1967; Aha et al., 1991), (iii) Support Vector Machines Regression (Boser et al., 1992; Cortes & 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 A.1.2* 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 A.1.2* provides the details about the evaluation metrics used.

2.3.5 Comparison with MCYFS

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 & López 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 & Nisini (2019)) using the official Eurostat national yield statistics (Eurostat, 2021a) as ground truths. We compared 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).

2.3.6 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/).

Table 2.3: Feature selection frequencies for potatoes (No Yield Trend). 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.

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

2.4 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 A.4*.

2.4.1 Feature Selection Frequencies

Feature selection counts for potatoes show that soil water holding capacity was always selected (*Table 2.3*). 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.

2.4.2 Yield Trend vs. No Yield Trend

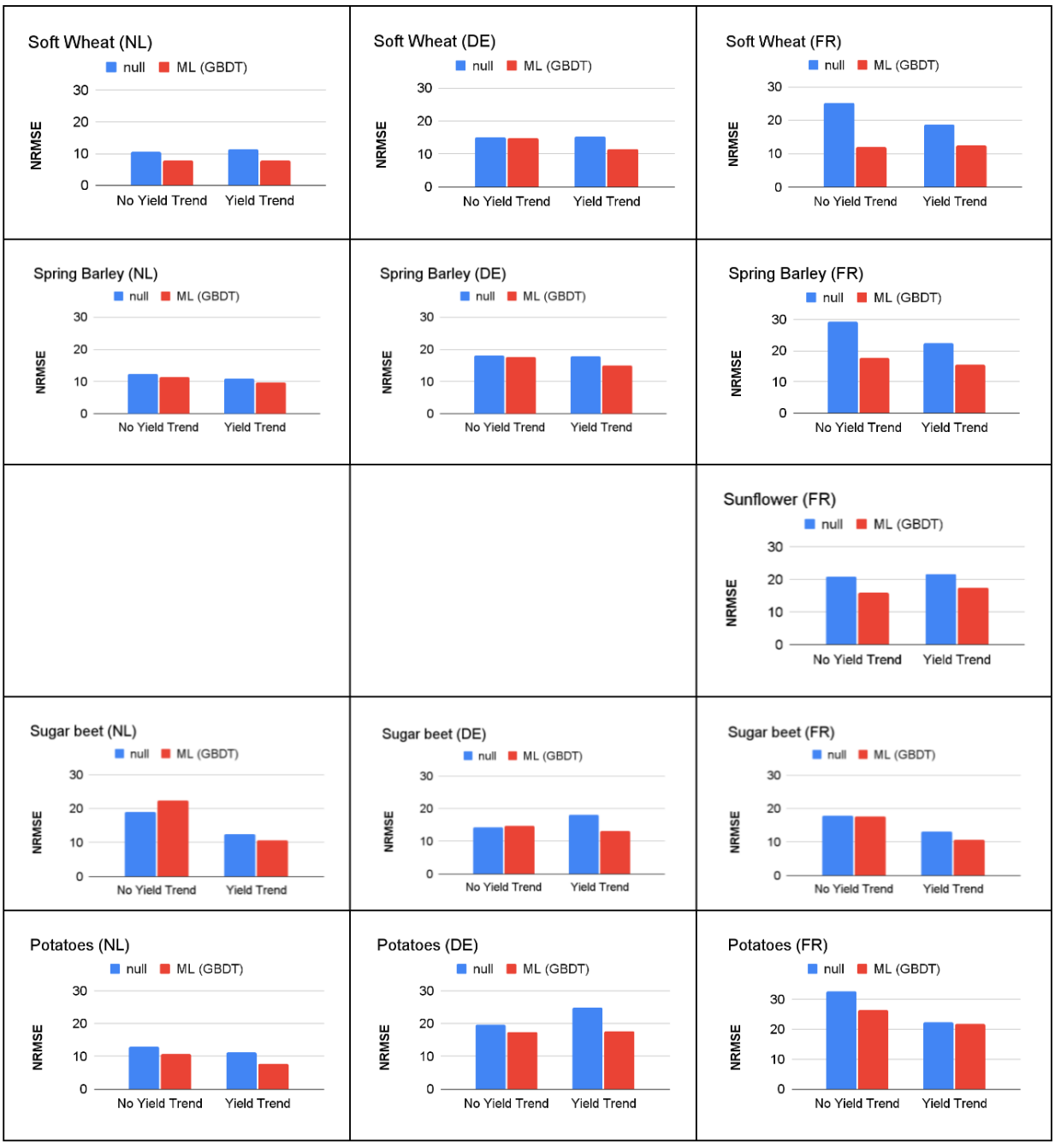


Figure 2.5: Yield Trend vs. No Yield Trend. The normalized RMSE of Gradient Boosted Decision Trees was compared with the null method.

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 (*Figure 2.5*; *Figure A.3*). 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" (*Figure 2.5*); MAPE for potatoes (FR) "Yield Trend" (*Figure A.3*)), machine learning performed better than the null method. Because of the differences in training and test sets (see *Section 2.3.1*), 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 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.

2.4.3 Early Season vs. End of Season Predictions

Early season predictions using yield trend (*Figure 2.6*; *Figure A.4*) 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) (*Figure A.4*). 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.

2.4.4 Comparison with MCYFS

We aggregated the predictions of the machine learning baseline to NUTS0 and compared them with past MCYFS forecasts using Eurostat national yield statistics as ground truths. 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. For early season, we compared the predictions of machine learning for 30 days after planting with MCYFS forecasts from the closest dekad (*Figure 2.7a*; *Figure A.5a*). We also compared machine learning predictions at the end of the harvest window with the final MCYFS prediction of the year (*Figure 2.7b*; *Figure A.5b*). 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 (potatoes and sugar beet) 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 (*Figure 2.7a,b*; *Figure A.5a,b*).

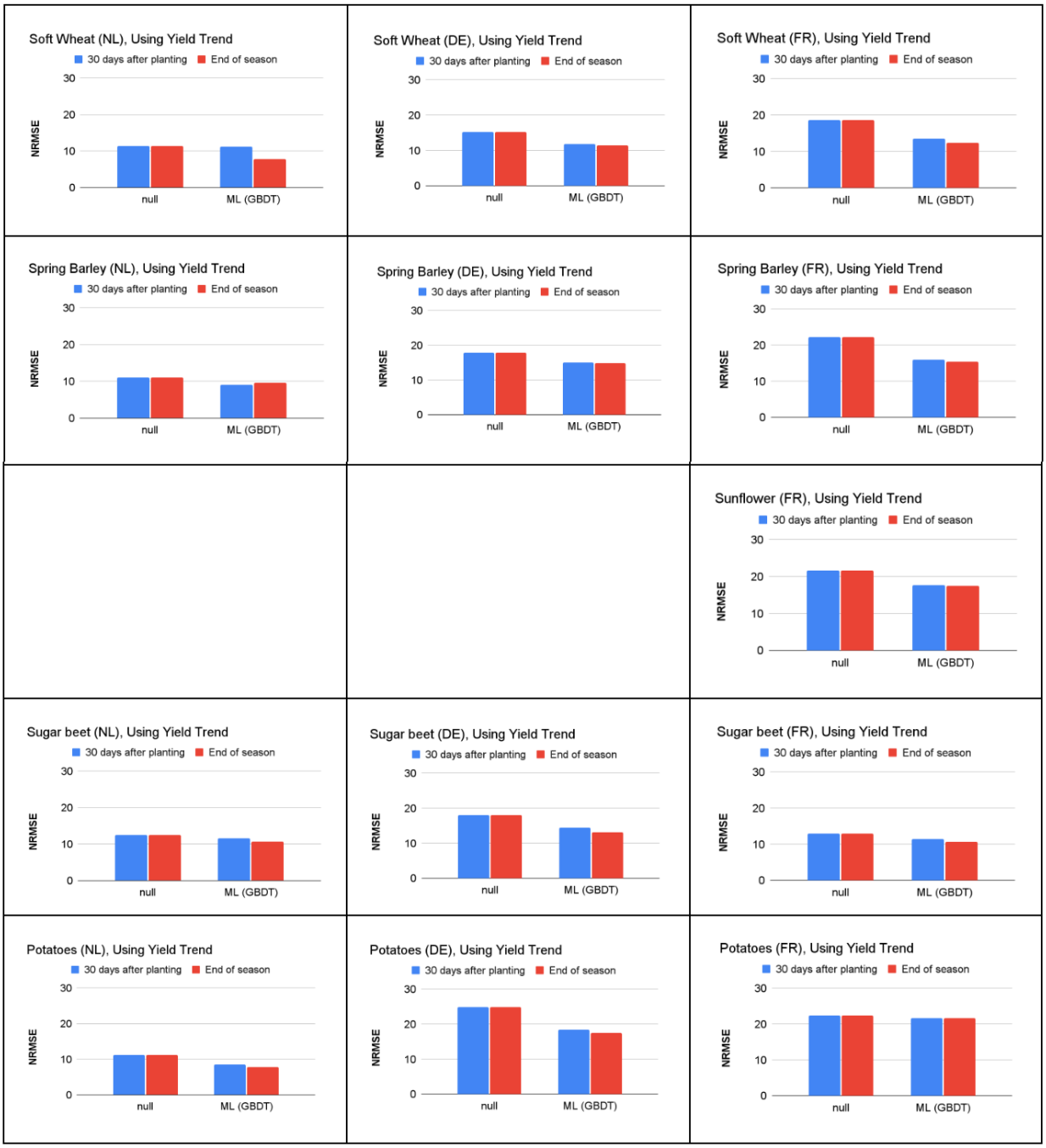
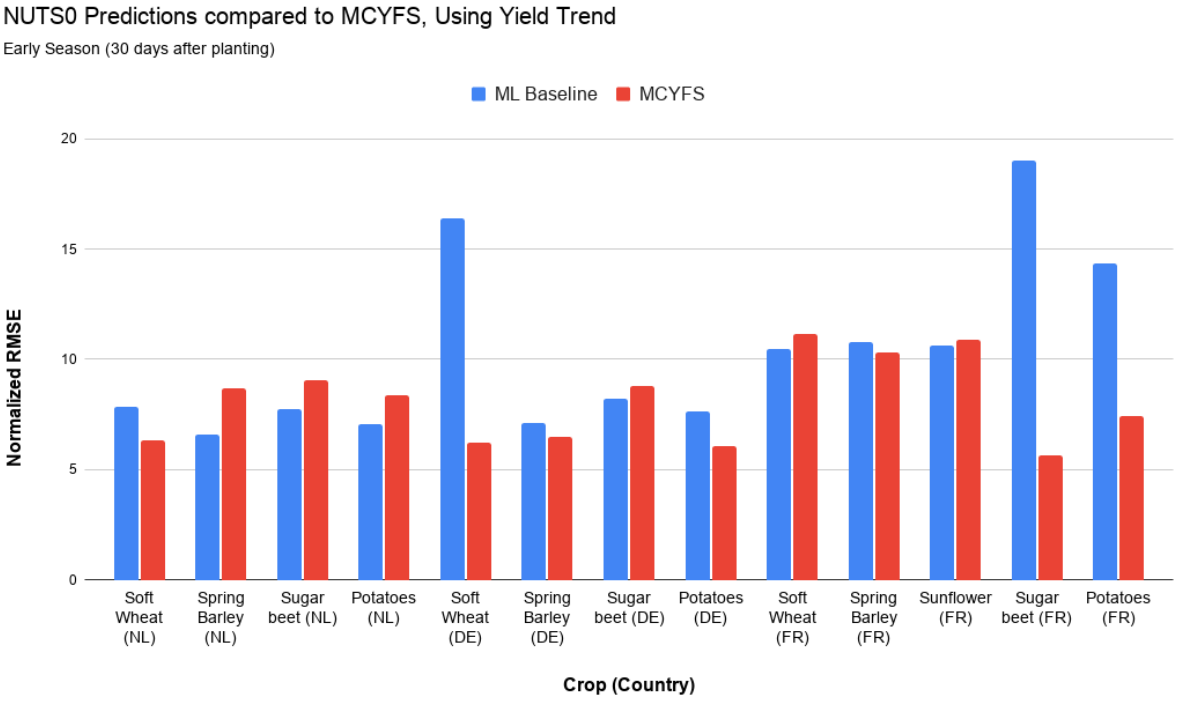
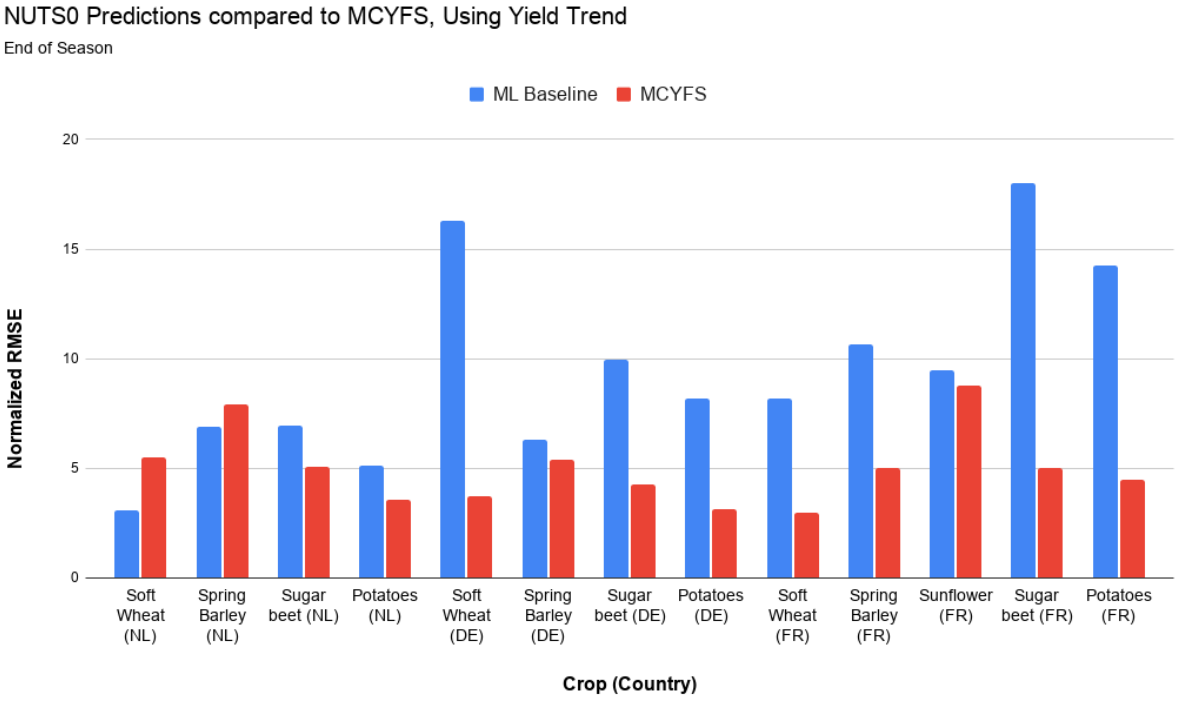


Figure 2.6: Early season prediction using a 5-year yield trend. The normalized RMSE of Gradient Boosted Decision Trees for early and end of season predictions.

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 (López-Lozano & Baruth, 2019).



(a) Early season (30 days after planting)



(b) End of season (end of harvest window)

Figure 2.7: 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.

2.5 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 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., 2015).

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 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 makes accurate predictions 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, 2020a). Yield statistics at sub-national levels are not curated as often and vary across countries and crops (López-Lozano et al., 2015). 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), *Figure A.4*). 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 *Figure 2.7a*; *A.5a*). In terms of methodology, MCYFS uses data from all previous years to train models for the upcoming year (see van der Velde & 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 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, López-Lozano & 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 & 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.

2.6 Conclusion

We designed a modular and reusable machine learning workflow for crop yield prediction and tested the workflow on ten 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.

Chapter 3

Machine learning for regional crop yield forecasting in Europe

This chapter is based on:

Dilli Paudel, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, and Ioannis N Athanasiadis. Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*, 276:108377, 2022. doi:10.1016/j.fcr.2021.108377

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.

3.1 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 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 & Belward, 2005; Büttner 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 & 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 & 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 chapter, 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 where yield and crop area statistics are available. NUTS (Nomenclature of Territorial Units for Statistics) is a hierarchical system of dividing the territory of the European Union for statistics and policy (Eurostat, 2016b). 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 chapter is structured as follows. *Section 3.2* describes the data and methods, *Section 3.3* presents the results, *Section 3.4* discusses our findings and outlines areas for future work and *Section 3.5* summarizes our conclusions. *Appendix B* provides details and supporting evidence for *Section 3.2* (*Materials and Methods*), *Section 3.3* (*Results*) and *Section 3.4* (*Discussion*).

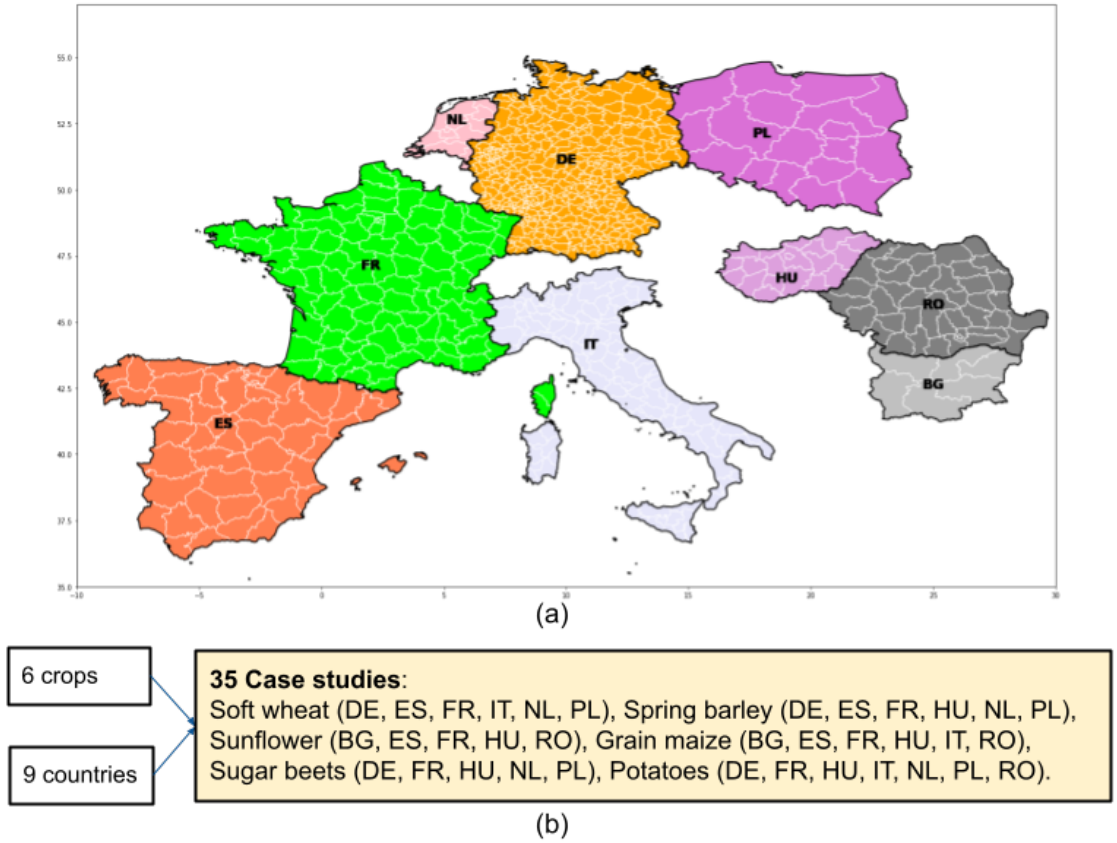


Figure 3.1: (a) Selected countries and their NUTS regions. (b) Selected case studies. Case studies included nine major crop-growing countries of Europe for soft wheat, spring barley, sunflower, grain maize, sugar beets and potatoes.

3.2 Materials and Methods

As stated in *Section 3.1* 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) (*Figure 3.1*; *Table B.1*).

3.2.1 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 (*Figure 3.2*). 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

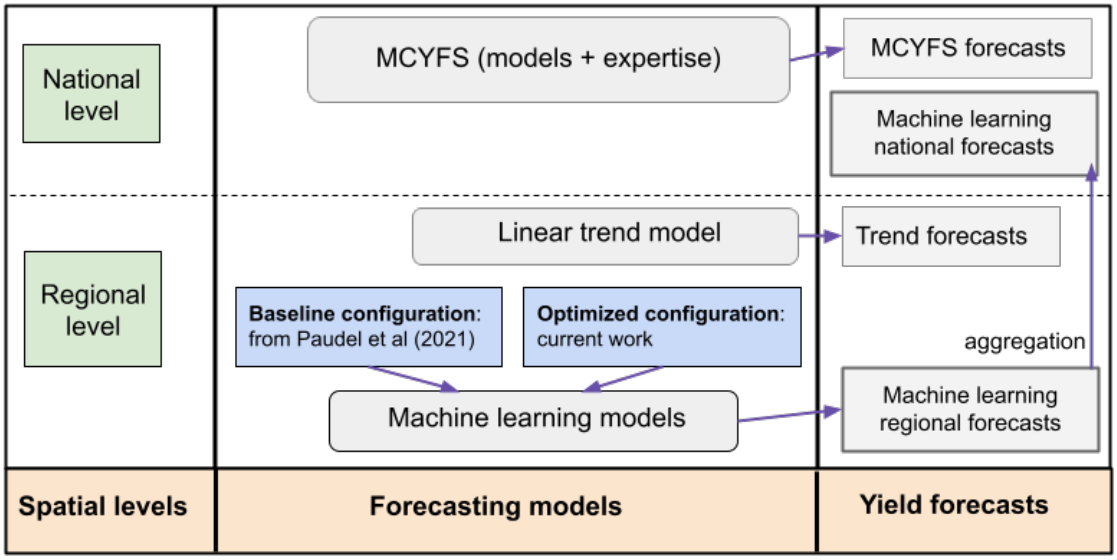


Figure 3.2: Framework to evaluate the quality of machine learning forecasts. Regional data was used to build a linear trend model and two sets of machine learning models. Regional machine learning forecasts were first compared with linear trend model forecasts and later aggregated to national level to compare with MCYFS national forecasts. MCYFS forecasts were provided by the European Commission's Joint Research Centre (JRC).

as crop, country, forecast dekad (10-day period relative to harvest), crop calendar and prediction algorithms. Models trained using the workflow configuration of this chapter were also compared with those from our previous work (Paudel et al. (2021); *Chapter 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 3.2.5*.

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 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 *Figure 3.1b*), 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 chapter, 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 3.2.2*) and the optimized configuration (*Section 3.2.3*). 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. *Section 3.2.3* and *B.2* provide more details about differences between the baseline configuration and the optimized configuration.

3.2.2 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 2.1*) 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 2.2*. 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 3.2.5*).

3.2.3 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 B*.

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 B.2*). 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 3.1*; *Table B.3*). In the baseline and the improved workflow, we capture the yield trend 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*,

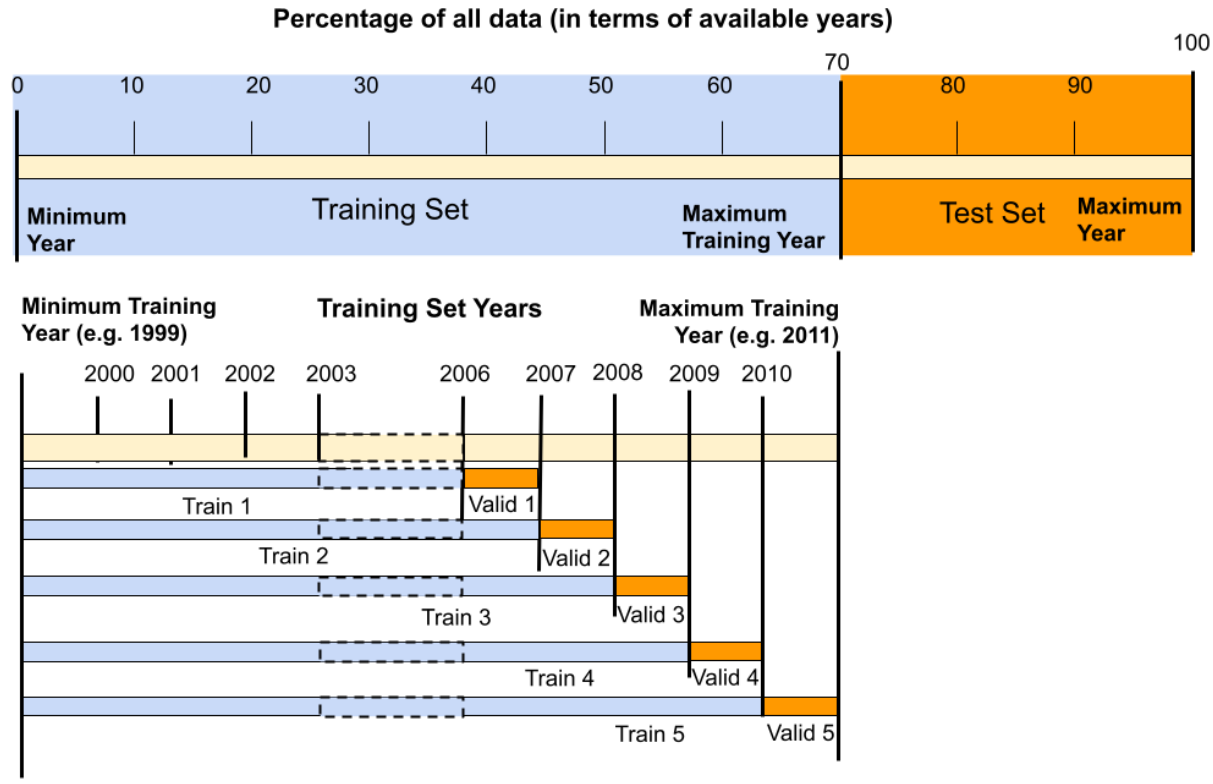


Figure 3.3: Training, validation and test splits. Data for each region was split into training and test sets using time-based ordering of available years. The training data was further split using 5-fold sliding validation for feature selection and hyperparameter optimization (the lower panel). The dashed area is not drawn to scale.

we added a robust hyperparameter search based on Bayesian optimization (Brochu et al., 2010; Shahriari et al., 2015). Bayesian optimization selects a new 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 (*Figure 3.3*). 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 to 2006 for training and 2007 for validation, the second iteration used 1999 to 2007 for training and 2008 for validation, the third iteration used 1999 to 2008 for training and 2009 for validation, and so on. During evaluation on the test set (the 30% in *Figure 3.3*), 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 (see *Figure 3.4*). This approach is comparable to how operational systems such as MCYFS work.

We designed improvements to the machine learning baseline with emphasis on reusability

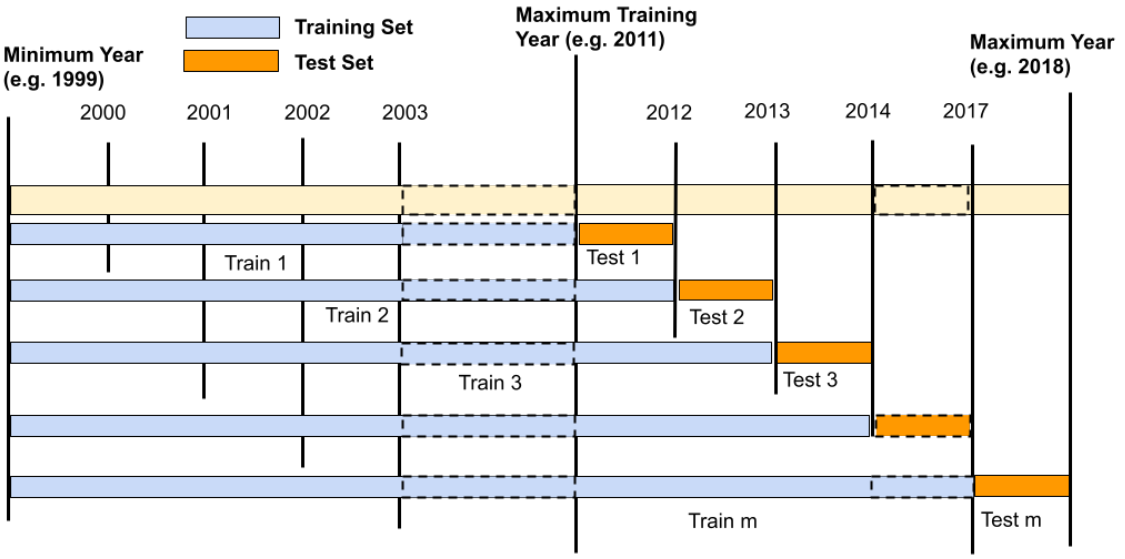


Figure 3.4: Per test year model refitting. After 5-fold sliding validation to find optimal features and hyperparameters, we fitted a model on the entire training set. For every test year, we refitted this model on the training years and the previous test years, thus utilizing additional data available. The dashed areas are not drawn to scale.

and scalability of the approach to a large-scale system such as MCYFS. In this chapter, 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.

3.2.4 Data

Our data came from MCYFS (see EC-JRC (2022); MARSWiki (2021)) 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 3.1*; *Table B.2, B.3*). 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.

3.2.5 Evaluation

We evaluated the quality and usefulness of machine learning forecasts using three steps (*Table 3.2*). *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.

Table 3.1: Data sources summary. *Section B.3* provides additional details about the data sources.

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 soil 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) (mm). **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 regions's crop area. **Source**: Eurostat (Eurostat, 2021a) and MCYFS (EC-JRC, 2022).
Irrigated area	Irrigated total area and irrigated crop-specific area (ha). **Source**: EC-JRC (2022).
Elevation, slope	Average and standard deviation of elevation (m) and slope ($degrees$). **Source**: USGS-EROS (2021).
Soil	Soil water holding capacity. **Source**: MCYFS. See Lecerf et al. (2019).
Field Size	Average and standard deviation (ha). **Source**: Lesiv et al. (2019).
Yield	Yield at regional (NUTS2 or NUTS3) and national (NUTS0) level ($t\ ha^{-1}$). **Regional Source**: NL-CBS (2020); FR-Agreste (2020); DE-RegionalStatistiks (2020); Eurostat (2021a); EC-JRC (2022). **National Source**: Eurostat (2021a); EC-JRC (2022).
MCYFS crop yield forecasts	Date and forecast value ($t\ ha^{-1}$). **Source**: MCYFS See van der Velde & Nisini (2019).

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

Table 3.2: Summary of methods to evaluate the regional and national predictions. We evaluated the quality and usefulness of machine learning predictions both at regional and national levels.

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 in cases with low average and high standard deviation of trend residuals. **Section** 3.2.5	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 the trend model 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** 3.3.1
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** 3.2.5	2 A large percentage of matching yield classes would show better prediction results. **Section** 3.3.2
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** 3.2.5	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** 3.3.3

normalized root mean squared error (NRMSE) and the coefficient of determination or R^2 for comparing model performance. 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 (Demšar, 2006; Kadra et al., 2021).

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 (predicted 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.

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 (2021)). Grain maize (2015) was selected because of high yield losses in Central Europe (see MARS bulletin for 2015 Vol 23 No. 9, EC-JRC (2021)). 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.

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 & 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.

3.2.6 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 hyperparameter search was based on the scikit-optimize package (Scikit-optimize Contributors, 2021). Our implementation is available through the *pypi* repository (`https://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.

3.3 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 is reported in the supplementary results (*Section B.5*).

3.3.1 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 (*Figure B.1*). The median NRMSEs for 60 days early were 17.27% (baseline) and 16.57% (optimized), and those for 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 B.4*), for all except potatoes (HU) (*Figure 3.5*, *Table B.4*) 60 days before harvest, and for all cases at the end of season (*Table B.4*; *Figure B.1*). 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 (*Figure 3.6*; *Figure B.3*).

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 3.3*;

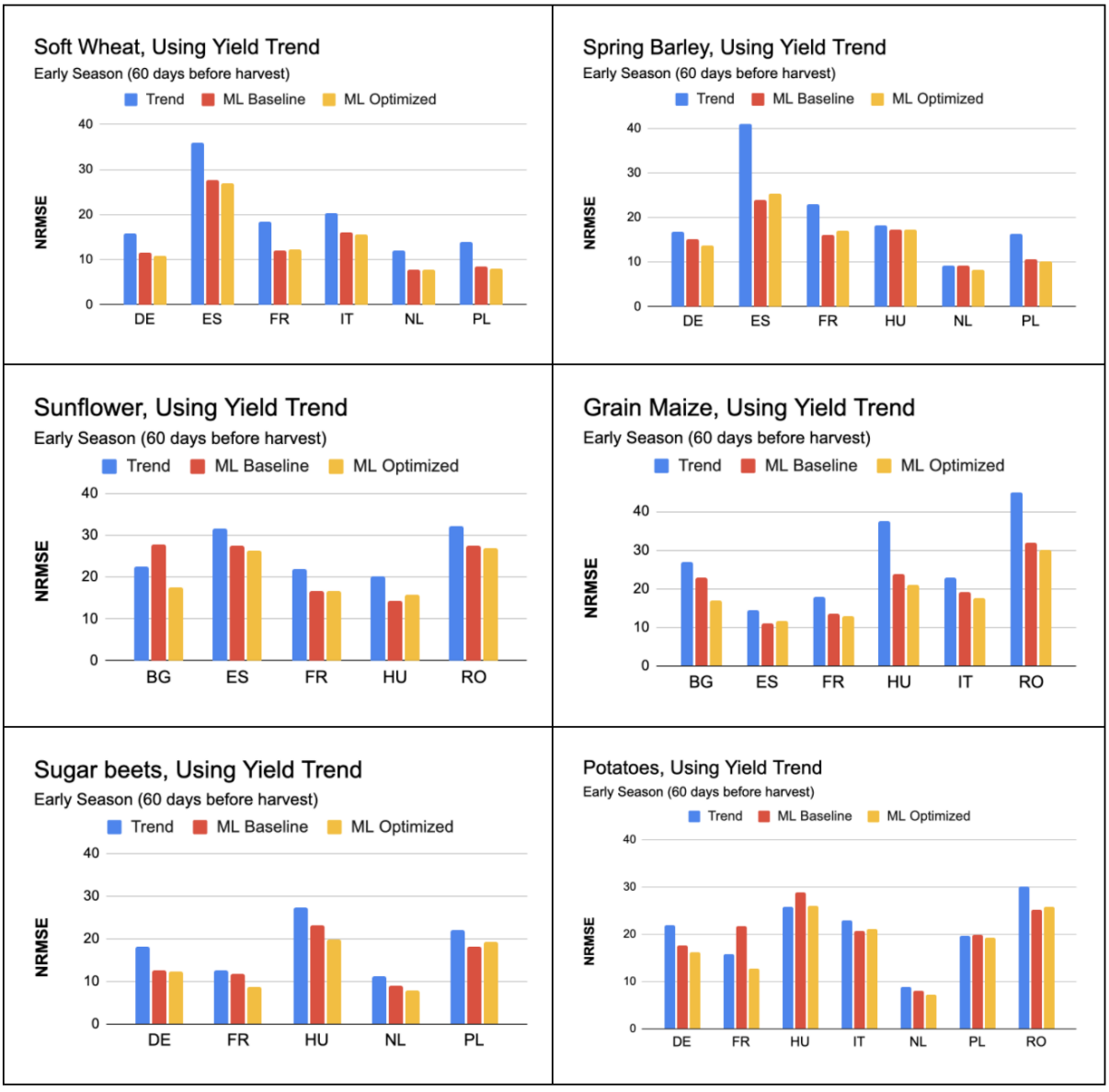


Figure 3.5: Normalized RMSE of regional forecasts 60 days before harvest. Regional forecasts from machine learning (baseline and optimized) were compared with the trend model. For machine learning models, we show results for the algorithm with the lowest NRMSE.

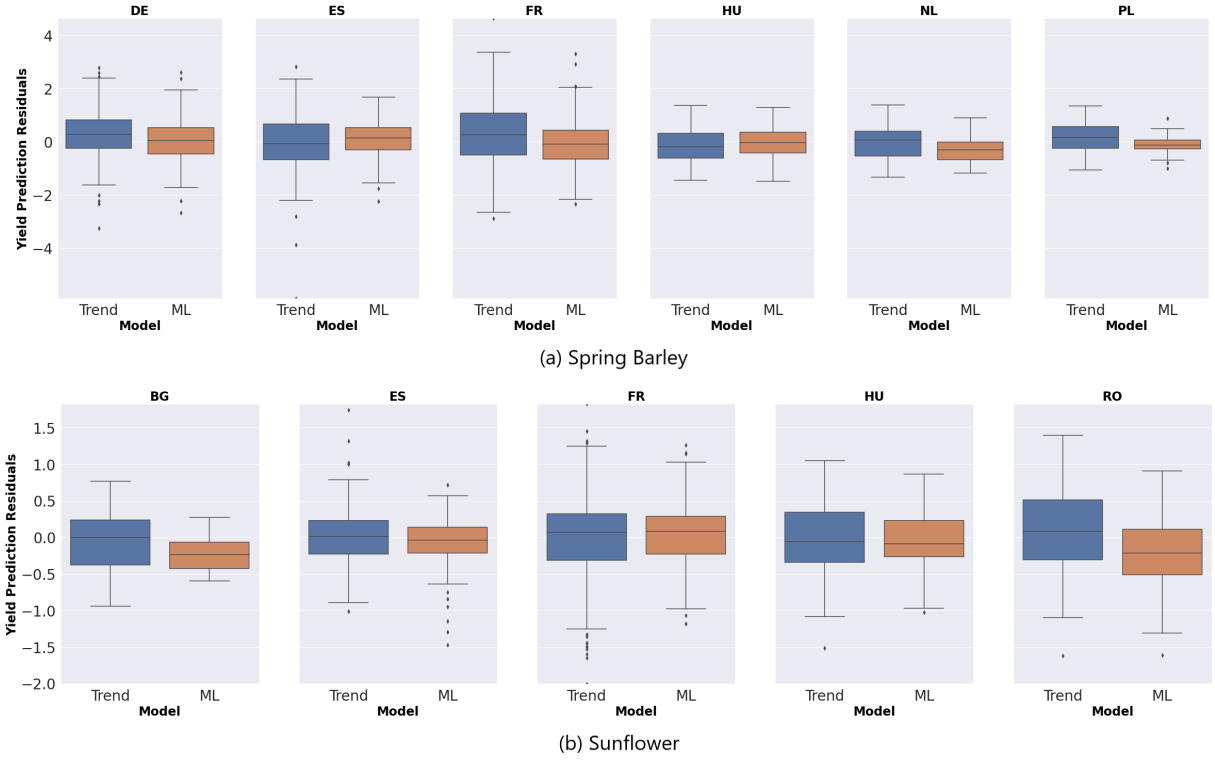


Figure 3.6: Boxplots of regional yield residuals 60 days before harvest. The trend model (blue) has a higher variance than machine learning (orange). For machine learning, we show results for the algorithm with the lowest NRMSE. *Figure B.3* shows boxplots for other crops.

Table 3.3: Coefficient of variation for regional prediction residuals 60 days before harvest. For cases where yield trend residuals had a low average and high variance, machine learning prediction residuals almost always had a lower coefficient of variation, indicating lower uncertainty.

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

Table B.5). The low CV of machine learning provided confidence on the quality of regional 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 3.3*). 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 B.6*, *Figure B.7*, *Figure B.8*). Similarly, machine learning and MCYFS forecasts at national level follow the reported yields quite closely (*Figure 3.8a*). However, absolute prediction residuals for many regions were high (25-50%, orange) and very high (>50%, red) in 2012, 2014, 2015 and 2016 (*Figure B.5*). Such disparity between regional and national level shows the limitation of national forecasts and the added value of regional forecasts.

3.3.2 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%) (*Figure 3.7a*; *Figure B.6a*). 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%) (*Figure 3.7b*, *Figure B.6b*). 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 (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%) (*Figure 3.7c*, *Figure B.6c*). 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 (*Figure 3.7c*).

3.3.3 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 B.6*). 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 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.

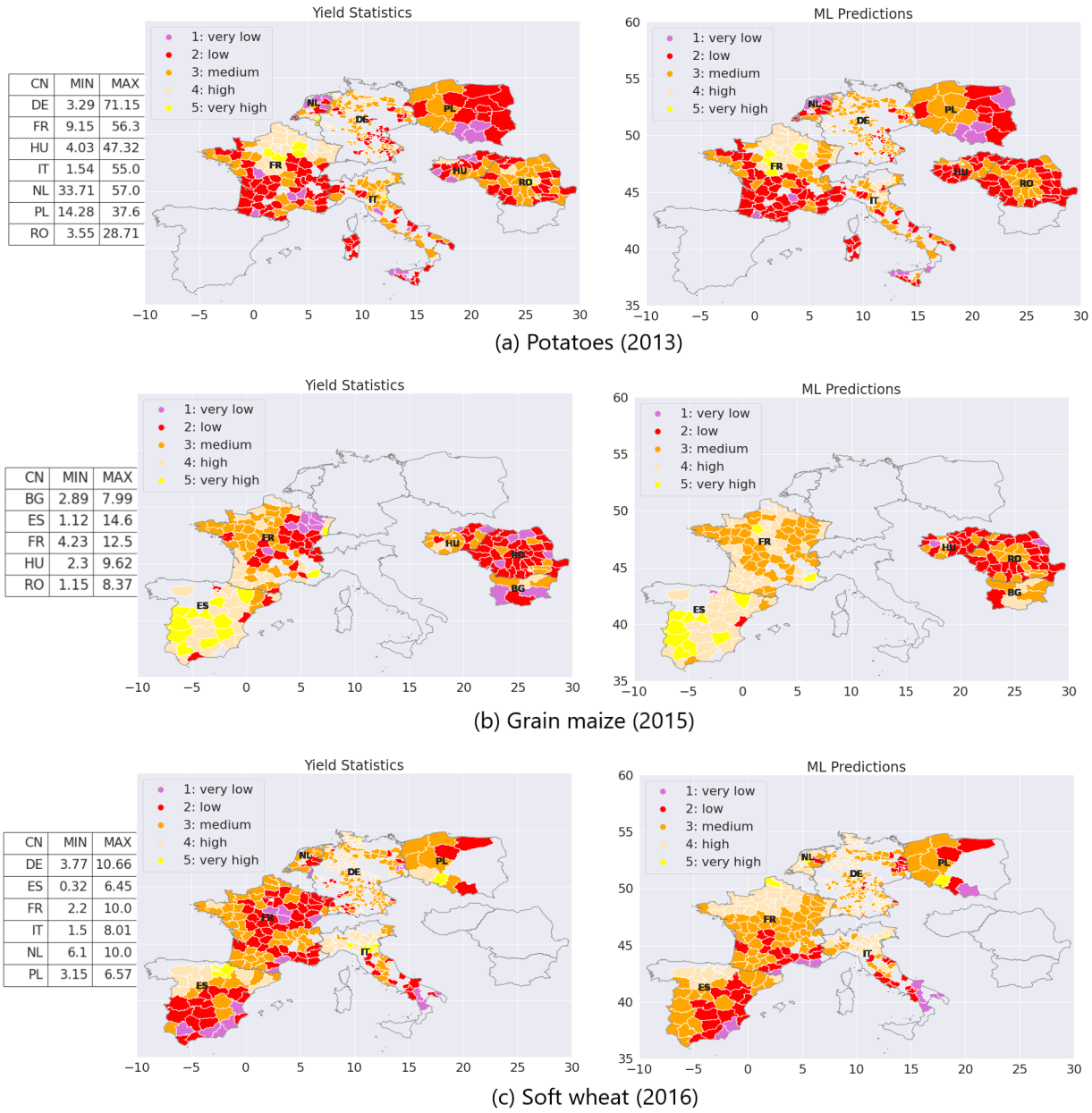


Figure 3.7: Regional forecasts 60 days before harvest vs reported yields. (a) 2013 - an average harvest for potatoes. (b) 2015 - an extreme harvest for grain maize. (c) 2016 - an extreme harvest for soft wheat (mainly in the north of FR). The machine learning models and yield classes are per country. Very low: up to 20% above the min yield; Low: 20-40%; Medium: 40-60%; High: 60-80%; Very high >80%.

At the end of season, normalized RMSE for machine learning were lower than MCYFS for 13 out of 35 cases (*Figure B.7*). 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 & Baruth, 2019).

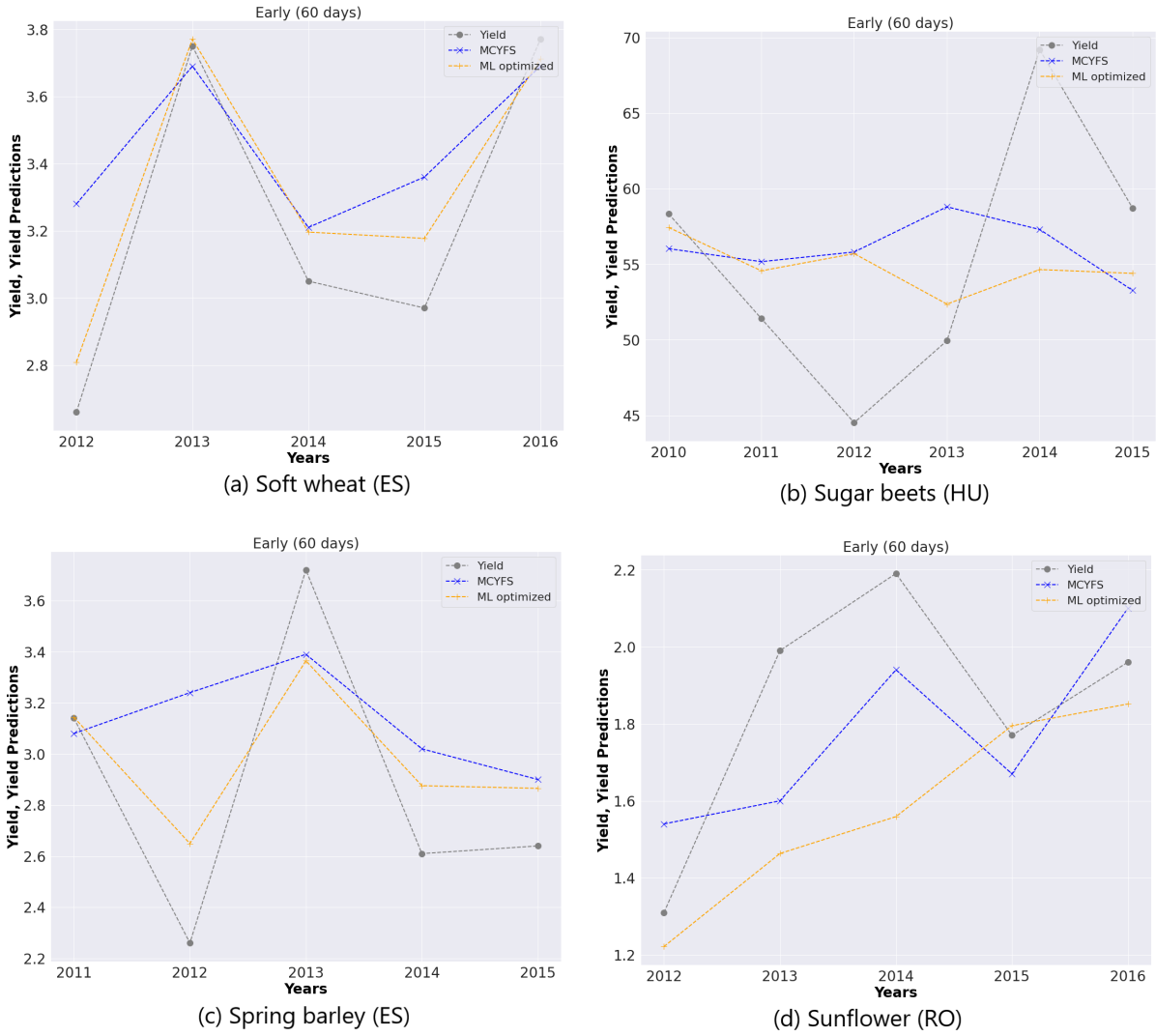


Figure 3.8: National forecasts 60 days before harvest compared with reported yields. For machine learning, we selected the algorithm with the lowest NRMSE. (a) Both machine learning and MCYFS capture the temporal variability. (b) Both do not capture temporal variability. (c) Machine learning performs better than MCYFS. (d) MCYFS outperforms machine learning.

Machine learning and MCYFS capture 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) (*Figure 3.8a,b*; *Figure B.9a,b*). Machine learning follows the reported yield better than MCYFS in certain cases, such as soft wheat (NL), spring barley (ES) and spring barley (PL). Similarly, MCYFS captures the variability better than ML in others, such as sunflower (BG), sunflower (RO), grain maize (FR) and potatoes (FR) (*Figure 3.8c,d*; *Figure B.9c,d*). Overall, we could see the added value of machine learning in some cases and its limitations in others.

3.4 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 & 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 B.6*) in which machine learning performs significantly better than MCYFS early in the season. Our results indicated 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 chapter, 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 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 (*Figure 3.6*; *Figure B.3*). We attempted to capture weather extremes using z-score features, but they were not always effective (for example, Grain Maize (2015); *Figure 3.7*b). 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 (*Figure B.4*). 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 chapter. 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 (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 chapter, 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.

3.5 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.

Chapter 4

Interpretability of deep learning models for crop yield forecasting

This chapter is based on:

Dilli Paudel, Allard de Wit, Hendrik Boogaard, Diego Marcos, Sjoukje Osinga, and Ioannis N Athanasiadis. Interpretability of deep learning models for crop yield forecasting. *Computers and Electronics in Agriculture*, 206:107663, 2023. doi:10.1016/j.compag.2023.107663.

Abstract

Machine learning models for crop yield forecasting often rely on expert-designed features or predictors. The effectiveness and interpretability of these handcrafted features depends on the expertise of the people designing them. Neural networks have the ability to learn features directly from input data and train the feature learning and prediction steps simultaneously. In this paper, we evaluate the performance and interpretability of neural network models for crop yield forecasting using data from the MARS Crop Yield Forecasting System of the European Commission's Joint Research Centre. The selected neural networks can handle sequential or time series data and include long short-term memory (LSTM) recurrent neural network and 1-dimensional convolutional neural network (1DCNN). Performance was compared with a linear trend model and a Gradient-Boosted Decision Trees (GBDT) model, trained using hand-designed features. Feature importance scores of input variables were computed using feature attribution methods and were analyzed by crop yield modeling and agronomy experts. Results showed that LSTM models perform statistically better than GBDT models for soft wheat in Germany and similar to GBDT models for all other case studies. In addition, LSTM models captured the effect of yield trend, static features (e.g. elevation, soil water holding capacity) and biomass features on crop yield well, but struggled to capture the impact of extreme temperature and moisture conditions. Our work shows the potential of deep learning to automatically learn features and produce reliable crop yield forecasts, and highlights the importance and challenges of involving human stakeholders in assessing model interpretability.

4.1 Introduction

Crop yield forecasts provide useful information to many stakeholders, including farmers, policymakers and commodity traders, for strategic decisions related to food security and market access (Basso & Liu, 2019; Chipanshi et al., 2015). Crop yield is influenced by complex interactions among crop-specific, environmental and management-related factors. Such complexity makes understanding yield forecasting models challenging but also critical. In recent years, machine learning methods have become popular in crop yield forecasting (Chlingaryan et al., 2018; Van Klompenburg et al., 2020). Reliability of these methods depends on how well their forecasts can be interpreted in human understandable terms.

Interpretability is defined as the degree to which humans can understand the causes of a decision (Doshi-Velez & Kim, 2017; Miller, 2019). Interpretability is related to trust and trust is often based on understanding how model predictions change when inputs are changed. Accuracy and efficiency are two other requirements for interpretability (Rüping, 2006). Accuracy means the forecasts must be close to observed values; efficiency means humans must understand model behavior within a limited amount of time. The common solution to interpretability is to build simple and inherently understandable models, such as linear models, process-based models, or decision trees (Molnar, 2022; Ribeiro et al., 2016). Linear regression – with crop model outputs, weather variables and remote sensing indicators – is commonly used in crop yield forecasting due to its simplicity and interpretability (van der Velde & Nisini, 2019; Statistics Canada, 2020; Lobell et al., 2015). Focusing only on inherently interpretable models would limit the type of relationships that can be modeled and hence result in less accuracy and usability (Ribeiro et al., 2016).

Machine learning and deep learning models can learn complex relationships, but they are often seen as black boxes because of a lack of understanding about how they make predictions (McGovern et al., 2019). Some methods, such as decision trees and their ensembles, are inherently interpretable. For other methods, including neural networks, feature attribution methods (Montavon et al., 2018; Ancona et al., 2018; Lundberg & Lee, 2017) provide an alternative to lack of inherent interpretability. They treat the original model as a black box and analyze predictions of the model to learn post-hoc explanations about relationships between predictors and crop yield. Interpretable machine learning and explainable artificial intelligence (explainable AI or XAI) are growing areas of research (Samek et al., 2019; McGovern et al., 2019; Xu et al., 2019). With the help of feature attribution methods, deep learning models have become easier to understand and interpret.

Deep learning also provides benefits of automatic feature learning. Standard machine learning methods involve a step to design features or predictors based on expert knowledge. Expertise-based feature design can produce meaningful features but has some shortcomings that limit its usefulness at large scale. First, except for certain features with well-defined formulas (e.g. vegetation indices), the feature design process is manual and time consuming (Bengio et al., 2013). Second, the effectiveness of handcrafted features depends on the expertise of the people designing them. Third, methods using handcrafted features often keep the feature design and prediction steps separate. This separation prevents updates to the feature design step when prediction models are trained with supervision labels (e.g. yield statistics). Neural

networks can extract features or representations directly from input data. Automatic feature learning not only removes the dependence on expertise, but also optimizes both feature extraction and prediction steps using supervision signals from training labels. Such combined learning improves the discriminative power of learned features (Wang & Yang, 2018).

Many studies have used deep learning for crop yield forecasting (Van Klompenburg et al., 2020; Gavahi et al., 2021; Oikonomidis et al., 2022)), and some of them have compared performance with standard machine learning methods and analyzed feature influence (Wolanin et al., 2020; Khaki et al., 2020; Shook et al., 2021; Nayak et al., 2022). However, they do not address challenges of explaining feature importance or model behavior to human stakeholders. Reasons explaining how forecasts were made are just as important as the forecasts themselves. We present an approach that involves human experts in the design of hand-crafted features as well as in the assessment of features and relationships learned by neural networks. Using this approach, we evaluate the accuracy and interpretability of deep learning models for crop yield forecasting. In particular, we seek to answer two questions: (i) Given the same input data, how well do deep learning models perform compared to standard machine learning models that use expert-designed features? (ii) According to experts, do deep learning models make predictions based on expected or plausible relationships? To answer the first question, forecasting performance was compared with a Gradient-Boosted Decision Trees model. To answer the second question, interpretability was assessed by a combination of quantitative and qualitative methods. Feature importance scores were obtained from post-hoc analysis of deep learning models and plotted to indicate the magnitude and direction (positive or negative) of impact on yield. The observed relationships and relative importance of features were analyzed by the same group of crop yield modeling and agronomy experts who previously provided input for feature design. Our approach provides a framework to compare performance between standard machine learning and deep learning methods, and includes human stakeholder feedback in interpretability assessment. Performance was compared for two crops – soft wheat and grain maize – and five countries: Germany, Spain, France, Hungary and Italy. Interpretability analysis was restricted to soft wheat and grain maize in France. Our work sheds light on the potential of neural networks to automatically learn meaningful features and produce reliable crop yield forecasts. Automating feature extraction reduces the dependence on manual feature design for large-scale crop yield forecasting.

The rest of the chapter is structured as follows. *Section 4.2* describes the data and methods, *Section 4.3* presents the results, *Section 4.4* discusses our findings and outlines areas for future work and *Section 4.5* summarizes our conclusions. *Appendix C* provides details and supporting evidence not included in the main text.

4.2 Methods

Our objective was to evaluate the accuracy and interpretability of deep learning methods for crop yield forecasting. Three types of models were built to assess the skill of neural networks to automatically learn features and produce accurate yield forecasts (*Figure 4.1*). First, linear trend models provided a baseline (the "null" model) for prediction skill. Second, Gradient-Boosted Decision Trees (GBDT) models represented standard machine learning

methods trained with expert-designed features. Other studies have compared performance with linear models, support vector machines (Boser et al., 1992; Cortes & Vapnik, 1995) and Random Forests (Breiman, 2001). In our case, the choice of GBDT was motivated by its performance in regional crop yield forecasting in Europe (Paudel et al., 2021, 2022a). Third, deep learning models were built using architectures that can extract features from seasonal time series data. For interpretability, feature attribution methods were used to analyze the forecasts of deep learning models and extract quantitative measures of feature importance, indicating the magnitude and direction (positive or negative) of influence on crop yield. The relative importance of features and their influence on yield were analyzed and validated qualitatively by human experts based on their knowledge and experience.

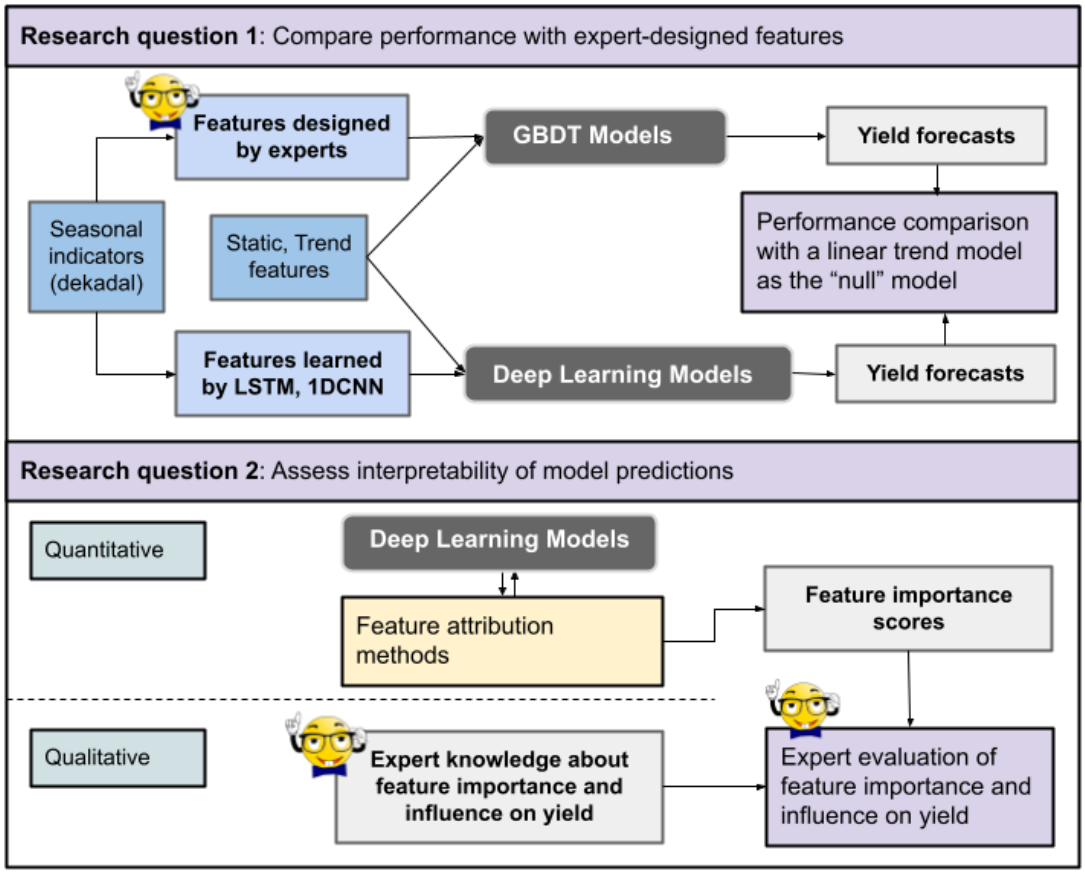


Figure 4.1: Framework to assess performance and interpretability of deep learning models. Performance was compared with a trend model and a GBDT model. Features for GBDT were designed by experts while deep learning models extracted features automatically. Long Short-Term Memory (LSTM) and 1-dimensional Convolutional Neural Network (1DCNN) were selected to learn features from time series of seasonal indicators. Static (e.g. elevation) and yield trend features were the same for both GBDT and deep learning. Feature importance scores learned by post-hoc analysis of deep learning model predictions represented quantitative measures of interpretability. They indicated the size of feature influence and the positive or negative impact on yield. Human experts with knowledge about factors affecting yield analyzed and provided feedback on the relative importance of features and the relationships with yield.

Table 4.1: Data sources summary. Case studies covered two crops and five countries: soft wheat (Germany, Spain, France, Italy) and grain maize (Spain, France, Hungary, Italy).

Data	Type of data	Indicators, Source
WOFOST crop model outputs	Seasonal time series (dekadal)	Water-limited dry weight biomass (WLIM_YB, $kg\ ha^{-1}$), Water-limited dry weight storage organs (WLIM_YS, $kg\ ha^{-1}$), Water-limited leaf area index (WLAI, $m^2\ m^{-2}$), development stage (DVS, 0-200), root-zone soil moisture as % of soil water holding capacity (RSM). **Source**: MCYFS. See Lecerf et al. (2019).
Meteo	Seasonal time series (dekadal)	Maximum daily air temperature(TMAX, °C), Minimum daily air temperature (TMIN, °C, average daily air temperature (TAVG, °C), sum of daily precipitation (PREC, mm), sum of daily evapotranspiration of short vegetation (ET0, mm) (Penman-Monteith, Allen et al. (1998)), climate water balance (CWB = PREC - ET0, mm). **Source**: MCYFS. See Lecerf et al. (2019).
Remote Sensing	Seasonal time series (dekadal)	Fraction of Absorbed Photosynthetically Active Radiation (Smoothed) (FAPAR). **Source**: MCYFS. See Copernicus GLS (2020).
GAES	Static	Agro-environmental zone identifiers. **Source**: Global agro-environmental stratification (Mücher et al., 2016).
Irrigated area	Static	Irrigated total area (IRRIG_AREA_ALL, ha), irrigated grain maize area (IRRIG_AREA2, ha) and irrigated cereals area as proxy for soft wheat (IRRIG_AREA90, ha). **Source**: EC-JRC (2022).
Elevation, slope	Static	Average elevation (AVG_ELEV, m), standard deviation of elevation (STD_ELEV, m), average slope (AVG_SLOPE, degrees), standard deviation of slope (STD_SLOPE, degrees) **Source**: USGS-EROS (2021).
Soil	Static	Soil water holding capacity (SM_WHC). **Source**: MCYFS. See Lecerf et al. (2019).
Field Size	Static	Average field size (AVG_FIELD_SIZE, ha), standard deviation of field size (STD_FIELD_SIZE, ha) **Source**: Lesiv et al. (2019).
Yield	Yearly	Yield at NUTS3 level (t/ha). **Source**: FR-Agreste (2020); DE-RegionalStatistiks (2020); Eurostat (2021a); EC-JRC (2022).

4.2.1 Data

Our data came from the MARS Crop Yield Forecasting System (MCYFS) of the European Commission's Joint Research Centre (EC-JRC, 2022; MARSWiki, 2021) and Eurostat (Eurostat, 2021a). Seasonal time series included outputs of the WOFOST crop model (van Diepen et al., 1989; de Wit et al., 2019), weather variables and remote sensing indicators aggregated to NUTS3 regions (*Table 4.1*). NUTS (Nomenclature of Territorial Units for Statistics) is a hierarchical system of dividing the territory of the European Union for statistics and policy

(Eurostat, 2016b). Yield values of five previous years were used to learn the yield trend, which captures the effect of technological improvements. Static data on soil water holding capacity, elevation, slope, field sizes and irrigated area was used to capture spatial differences among regions not covered by seasonal data (Paudel et al., 2022a). In addition, agro-environmental zones were added as categorical variables to account for other agro-climatic differences. Case studies covered two crops – soft wheat and grain maize – and five countries: Germany (DE), Spain (ES), France (FR), Hungary (HU) and Italy (IT) (*Figure C.1*). Models were trained with NUTS3 yield statistics as ground-truths. Remote sensing data and yield data determined the total data size, which ranged from 300 labeled instances for grain maize (HU) to 1950 for soft wheat (DE), and in most cases covered the years 1999 to 2018. The test set consisted of the most recent 30% of available years. Model hyperparameters were optimized using a five-fold validation scheme (*Figure 3.3*).

4.2.2 Trend and GBDT Models

The trend models fitted a line through yield values of five previous years. GBDT represents a standard machine learning algorithm that requires expertise-based features. GBDT is an ensemble of decision trees that relies on boosting (Friedman, 2001) for growing the trees. For GBDT, the crop calendar was inferred from WOFOST-simulated development stages and used to design features that capture the impact of various predictors during different crop calendar periods (Paudel et al. (2021); *Figure C.2*). We requested five experts to complete a survey about important predictors of crop yield in each period: pre-emergence (p0), emergence (p1), vegetative (p2), flowering (p3), yield formation (p4), maturity (p5). *Section 4.2.5* provides some details about the experts. For each period, they provided the seasonal indicators that influence crop growth and development and affect final crop yield. Results of the survey and follow-up discussions were used to design features for GBDT. The indicators selected for feature design are shown in *Table 4.2*. Except for the step to extract seasonal features (*Figure 4.2*), the input data and training and test splits (*Figure 3.3*) for both GBDT and deep learning models were identical.

4.2.3 Deep learning models

The same set of seasonal indicators considered important by experts were passed to deep learning models, but without feature design. Selected architectures included long short-term memory (LSTM) recurrent neural network and 1-dimensional convolutional neural network (1DCNN), which can automatically learn features from sequential data. LSTM processes sequential input one time step at a time and has a notion of memory to maintain or forget information from previous time steps. 1DCNN uses kernels or filters that slide across the input to create summaries of inputs covered by the size of the filter. *Section C.3* provides additional information on LSTM and 1DCNN. Features learned by LSTM or 1DCNN were combined with yield trend and static data and passed to the output layer (*Figure 4.2*). The model parameters or weights were optimized using the Adam optimizer (Kingma & Ba, 2014), with a batch size of 16. The hyperparameters learning rate and weight decay (aka L2-penalty) lambda were optimized using custom 5-fold validation (*Figure 3.3*; Paudel et al. (2022a)). Models trained with optimal hyperparameters were evaluated on the validation set with early stopping: training stopped after the validation error increased for two successive epochs. The

Table 4.2: Feature design table for GBDT. The feature design table from Paudel et al. (2022a) was updated based on expert survey and follow-up discussions. Experts identified the indicators that were important at different stages of crop growth and development. WOFOST indicators: Water-limited dry weight biomass (WLIM_YB), water-limited dry weight storage organs (WLIM_YS), water-limited leaf area index (WLAI), root-zone soil moisture (RSM). Weather variables: maximum, minimum, average daily air temperature (TMAX, TMIN, TAVG); sum of daily precipitation (PREC); climate water balance (CWB). Remote sensing indicator: Fraction of Absorbed Photosynthetically Active Radiation (FAPAR).

Period	**Maximum Values**	**Average Values, Average of cumulative values***	**z-scores based on long term avg, std**
Pre-emergence (p0)		CWB*	
Planting, Emergence (p1)		TAVG, CWB*	TMIN, PREC
Vegetative (p2)	WLIM_YB, WLAI	RSM, TAVG, CWB*, FAPAR	RSM
Flowering (p3)			RSM, PREC, TMAX
Yield Formation (p4)	WLIM_YB, WLIM_YS, WLAI	RSM, FAPAR	RSM
Maturity, Harvest (p5)		PREC	PREC

optimized hyperparameters and early stopping epochs were used to evaluate the models on the test set.

4.2.4 Feature attribution methods

We considered three post-hoc feature attribution methods: Occlusion (Zeiler & Fergus, 2014), Integrated Gradients (Sundararajan et al., 2017), GradientShap (Lundberg & Lee, 2017). Occlusion is similar to sensitivity analysis in that it replaces a portion of feature data with baselines (zeros or random values) and compares the differences in prediction errors. Integrated gradients computes feature importance by approximating the integral of gradients (or partial derivatives) of the model outputs to the inputs along a path from baselines to inputs. The baselines can be zeros or random values. Ancona et al. (2018) found that Occlusion better identifies a small number of important features, but Integrated Gradients is better at capturing global nonlinear effects and interactions among features. We selected GradientShap, based on SHAP (Shapley Additive Explanations, Lundberg & Lee (2017)), to include desirable properties of Shapley values from cooperative game theory. Shapley values (Shapley, 1953) capture feature importance for linear models in the presence of multicollinearity (Lipovetsky & Conklin, 2001). GradientShap uses expected gradients to approximate Shapley values. Expected gradients can be considered an approximation of Integrated Gradients with many baselines and one point in the path between the input and the baseline. The explanations produced by GradientShap and Integrated Gradients are additive. For GradientShap, summing the contributions of each feature approximates the output of the original model (Lundberg & Lee, 2017). For Integrated Gradients, contributions of features add up to the difference between output at the input and output at the baseline

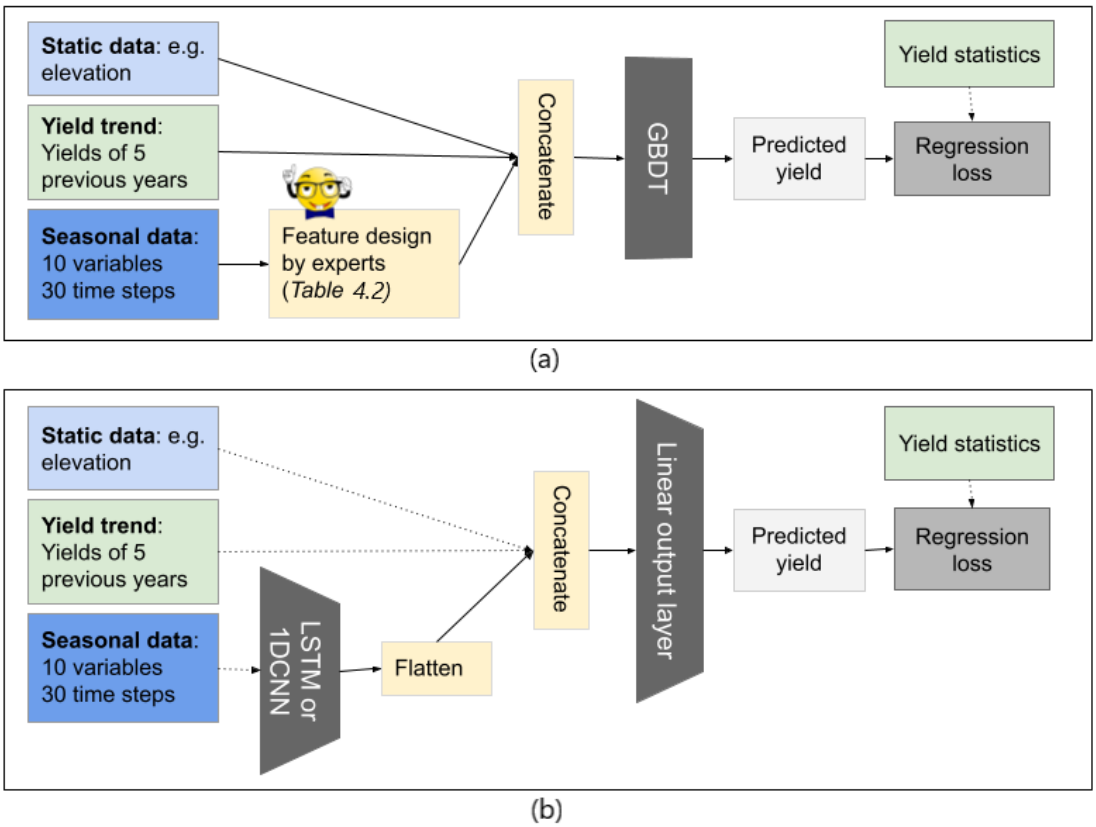


Figure 4.2: Deep learning framework (b) compared with the GBDT setup (a). In the case of GBDT, seasonal features were designed by experts. For deep learning, seasonal features were learned by training a Long Short-Term Memory (LSTM) or 1-dimensional Convolutional Neural Network (1DCNN) on seasonal indicator data. Yield trend features and static features were concatenated with seasonal features and passed to the output layer. The framework is kept as similar as possible with the GBDT setup.

(Sundararajan et al., 2017). GradientShap assumes that input features are independent. This assumption makes analysis of feature contributions easier, but ignores feature interactions. In this chapter, all methods were used as implemented in Captum (Kokhlikyan et al., 2020) - a model interpretability framework for PyTorch (Paszke et al., 2019).

4.2.5 Evaluation

We evaluated performance and interpretability of deep learning models in three steps. First, three types of models were built as described in *Section 4.2.2* and *4.2.3*. Forecasts from deep learning models were compared with the trend model and GBDT model. Second, feature attribution methods were used to extract importance scores of inputs passed to deep learning models. Third, human experts analyzed the interpretability of extracted relationships between features and yield as well as relative importance of features. In the case of deep learning, we use the term features loosely to mean the features learned from seasonal data as well as trend features and static data.

Evaluation of yield forecasts

Model performance was compared using normalized root mean squared error (NRMSE), defined to be the RMSE divided by the average yield of the test set. To produce early season forecasts, both GBDT and deep learning models used seasonal data up to 60 days before harvest. Both GBDT and deep learning predictions were collected from ten models to account for the effect of random seed or weight initializations. We used the average NRMSE of ten models to compare performance of trend, GBDT and deep learning methods (LSTM and 1DCNN). Similarly, prediction residuals used for boxplots and statistical tests were averaged across the ten models. Significance of model performance was evaluated using the Mann-Whitney U test (Mann & Whitney, 1947), which is a non-parametric version of Student's t-Test for independent samples. Variance and outliers were analyzed using boxplots of prediction residuals (predicted yield - reported yield). Spatial variability of yields and yield forecasts was qualitatively analyzed for the test years in France.

Evaluation of feature importance and interpretability

We selected five experts to provide input on important predictors of crop yield. Among them, four of them also participated in evaluation of feature importance and interpretability. Experts were selected based on familiarity with factors affecting soft wheat and grain maize yields, specifically in France. One of them is from Wageningen Plant Production Systems and has a background in agronomy and yield gap analysis. Two of them are senior researchers at Wageningen Environmental Research with experience in crop modeling. Two of them are from the European Commission's Joint Research Centre at Ispra where they work as MCYFS analysts to produce national-level crop yield forecasts. All experts know the input data well and have experience building crop yield forecasting models. However, they have not used deep learning methods and did not participate in the design of the deep learning architecture or selection of feature attribution methods.

The survey about important predictors of crop yield was conducted before running deep learning experiments. Therefore, the experts provided their prior knowledge about how each feature influenced yield. The scale used for the survey was: strong negative influence (-1), mild negative influence (-0.5), no influence (0), mild positive influence (0.5), strong positive influence (1). Experts scored how static features and seasonal indicators in different periods of the crop growing season affected the final yield. Among seasonal indicators, water-limited leaf area index (WLAI) and precipitation (PREC) were not included in the expert survey, mainly because they were correlated with other biomass and moisture indicators. They were later added based on the suggestions from experts during a follow up discussion.

To assess interpretability of deep learning models, we relied on quantitative importance scores from feature attribution methods and qualitative agreement scores from experts. For each feature, feature importance scores were plotted against feature values to show the positive or negative influence of the feature on yield. High importance for high feature values indicated a positive influence, while high importance for low feature values indicated a negative influence. We summed feature importance scores from a hundred runs – ten runs of a feature attribution method for ten models – to account for random initialization of weights and random baselines. Because the feature importance scores are additive, they can be seen as the contribution

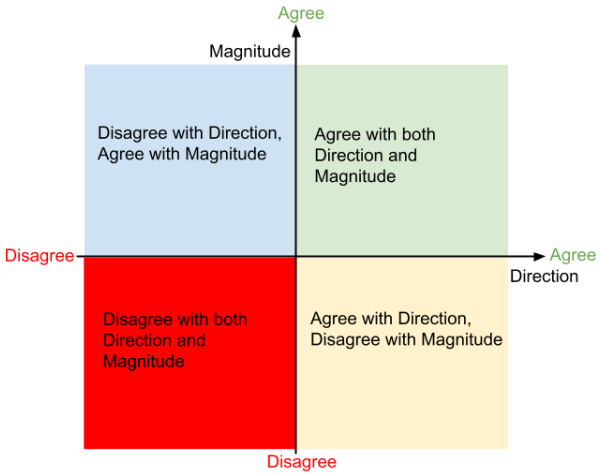


Figure 4.3: Agreement, disagreement quadrants for expert evaluation of interpretability. The quadrants represent agreement between experts and feature importance scores in two dimensions: the direction (positive or negative) of feature impact on yield (Direction), and the relative magnitude of feature influence (Magnitude).

(positive or negative) of each feature to the yield prediction.

Feature importance scores were analyzed for soft wheat and grain maize in France. France data is relatively large and of sufficiently high quality (Schauberger et al., 2018). Yields in France also show significant spatial and temporal variability. Yield variability is commonly divided into three components: average yield, yield trend and deviation from the trend (e.g. Dagnelie et al. (1983)). Inputs to deep learning models included static data to capture regional variation in average yields caused by topography, management and some agro-climatic differences. Trend features accounted for the multi-annual yield trend attributed to technological improvements. Seasonal features learned from dekadal time series were expected to capture yearly deviations from the trend. In line with expected effects on yield, importance scores were summarized for five classes of features: static data, yield trend features, seasonal biomass features, seasonal temperature features and seasonal moisture features. The importance of seasonal features was summarized by crop calendar periods (*Figure C.2*): pre-emergence (p0), emergence (p1), vegetative (p2), flowering (p3), yield formation (p4), maturity (p5).

Experts evaluated the interpretability of feature importance from deep learning based on their knowledge of factors affecting crop yield and the interactions among them. In particular, they provided scores representing whether they agreed or disagreed with the positive or negative impact of features on yield (Direction), and the relative magnitude of feature influence on yield (Magnitude) *Figure 4.3*). Yield referred to the end-of-season yield, and relative magnitude meant how the scores of one feature compared with those of another. As an extra option, they could indicate that they did not understand the relationships shown by feature importance plots. Interacting effects of multiple features were analyzed in a discussion session. When completing surveys to enter their scores, experts were requested to provide short notes explaining their scores, especially for disagreements.

4.3 Results

For performance comparison, we ran deep learning models with both LSTM and 1DCNN layers using data from 60 days before harvest. Most of the results reported in this paper correspond to the LSTM version due to its superior performance in the validation set (*Figure C.5*). Results for interpretability analysis are for LSTM and GradientShap. We selected GradientShap because the importance scores had lower variance across multiple runs (*Table C.3*). Scores from GradientShap were similar to those from Integrated Gradients based on Wasserstein distances. The distances between GradientShap and Integrated Gradients scores were 0.005 for soft wheat (FR) and 0.004 for grain maize (FR). The corresponding distances between GradientShap and Occlusion scores were 0.215 and 0.287.

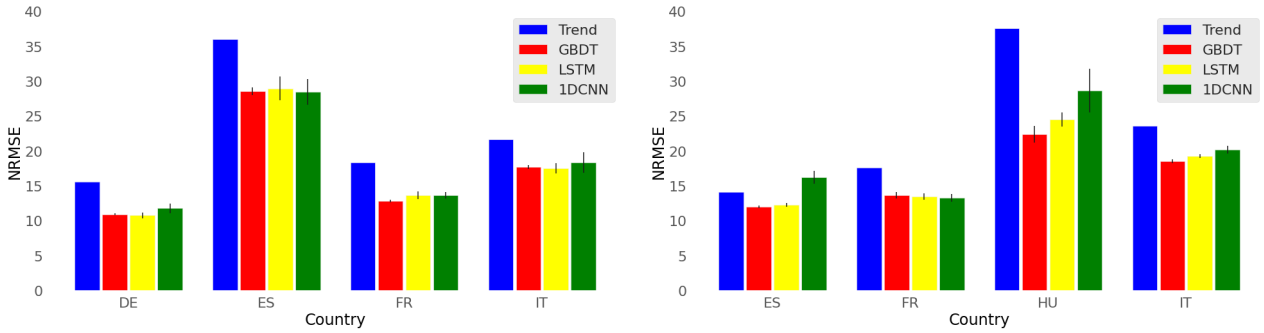
4.3.1 Performance comparison

Soft wheat LSTM models were better than GBDT models for DE (p-value 0.001) and statistically similar to GBDT models for other cases (p-values between 0.373 and 0.71). They were significantly better than trend models for all countries (*Table C.1*). Similarly, per-country average NRMSEs were similar for GBDT and LSTM (differences less than 1%) (*Figure 4.4a*), and they were lower than trend NRMSEs. Grain maize performance was less impressive. LSTM forecasts were still statistically similar to GBDT (p-values between 0.298 and 0.566) (*Table C.2*), and both GBDT and LSTM models had lower average NRMSEs than the trend model (*Figure 4.4b*). However, p-values from Mann-Whitney U Test (between 0.396 and 0.996) showed that GBDT and LSTM models were not significantly better than the trend model. For both crops, LSTM models showed more variability across the ten runs (i.e. a higher standard deviation) than GBDT. Evidently, weight initializations had a bigger impact on deep learning models than random seed had on GBDT.

Boxplots of prediction residuals and spatial maps for FR also showed similar patterns of performance differences between the two crops. In the case of soft wheat, prediction residuals from GBDT and LSTM models had lower spread and fewer outliers than those from trend models (*Figure 4.4c*). For soft wheat (DE), LSTM residuals were closer to zero than GBDT ones. For others, there were no significant differences between LSTM and GBDT. Spatial maps for soft wheat (FR) (*Figure C.6*) showed that both GBDT and LSTM predictions were close to the yield statistics for all test years except 2016, when there were significant yield losses in the north of FR (see Ben-Ari et al. (2018)). For grain maize, prediction residuals for GBDT and LSTM had slightly lower spread than trend models for ES, FR and HU (*Figure 4.4c*); for IT, the boxplots of all models were similar. GBDT and LSTM forecasts for grain maize (FR) (*Figure C.7*) were quite different from reported yields not only in 2015, when there were yield losses in Central Europe, but also in 2014, 2016 and 2017. Overall, grain maize forecasts were less accurate than soft wheat.

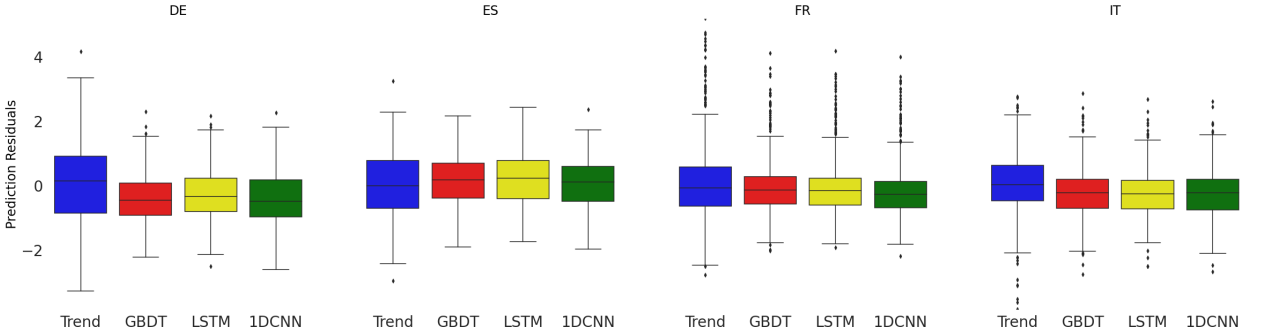
4.3.2 Interpretability of feature importance

Feature importance scores showed that trend features had the highest importance (*Figure C.9*). In general, static features ranked second after trend, and seasonal features had quite small importance values. Most experts agreed with this relative importance ranking. Trend

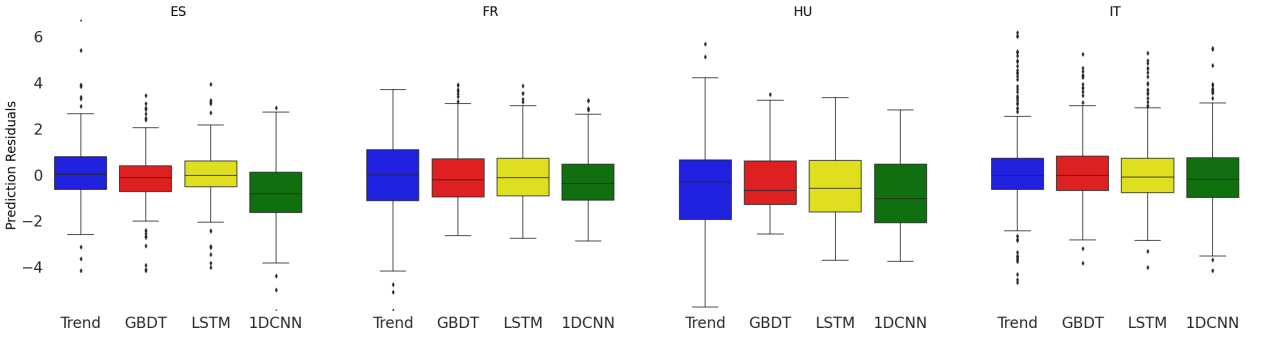


(a) NRMSEs, Soft wheat

(b) NRMSEs, Grain maize



(c) Box plots of prediction residuals, Soft wheat



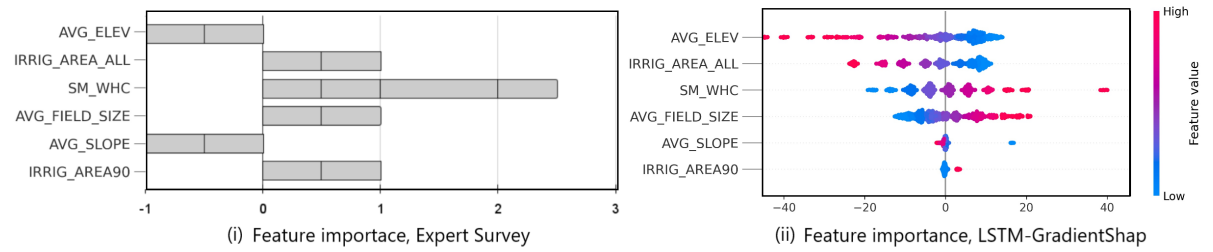
(d) Box plots of prediction residuals, Grain maize

Figure 4.4: Performance comparison 60 days before harvest. (a), (b): Average Normalized RMSE of deep learning models for different countries compared with a trend model and GBDT. (c), (d): Boxplots of prediction residuals (yield predictions - yield statistics). Prediction residuals and NRMSE were computed using predictions for all regions within a country in all test years. The values were then averaged across ten models. The error bars in (a) and (b) indicate the standard deviation of NRMSEs for the ten models.

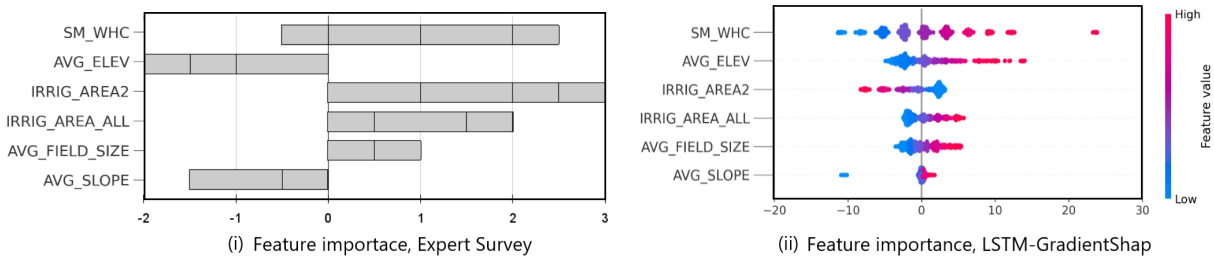
and static features reduce model bias and set the yield level, so they have high importance. Year-to-year deviation from the trend is usually relatively small, and hence seasonal features have low importance. This was particularly evident for grain maize; trend models produced statistically similar forecasts to GBDT and LSTM models, and the importance of most seasonal features were close to zero.

Static features

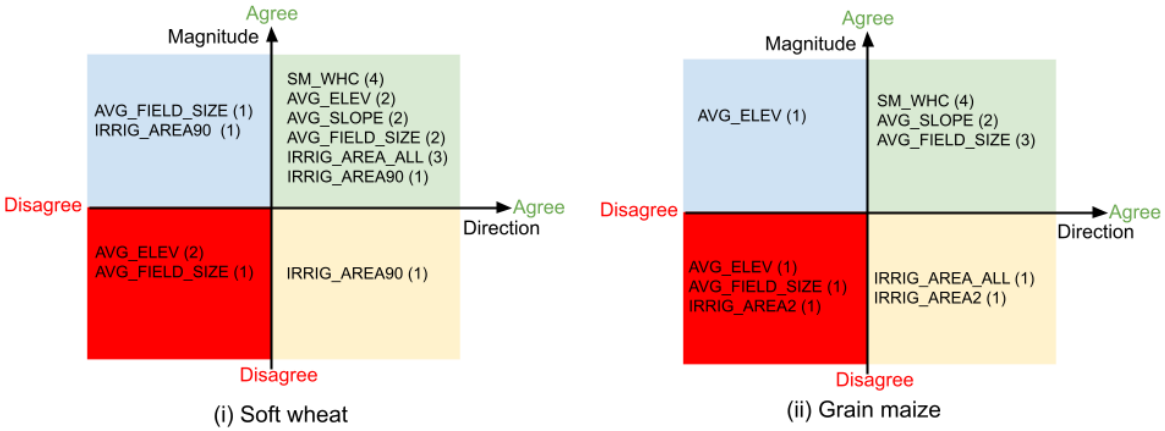
Agro-environmental zone (AEZ) features or AEZ identifiers captured the regional variation in average yields for both soft wheat and grain maize (*Figure C.8*). For soft wheat, feature



(a) Soft wheat: Feature importance from the expert survey (i) and deep learning models (ii).



(b) Grain maize: Feature importance from the expert survey (i) and deep learning models (ii).



(c) Interpretability scores from experts: (i) Soft wheat, (ii) Grain maize.

Figure 4.5: Importance and interpretability of static features. In (a) and (b), the scale used for the expert survey was: strong negative influence (−1), mild negative influence (−0.5), no influence (0), mild positive influence (0.5), strong positive influence (1). The divisions within each bar represent how different experts voted, e.g. experts assigned −0.5, 1, 1 and 0.5 for SM_WHC in (b). Importance scores from deep learning models show the magnitude and direction (positive or negative) of feature influence on yield. Feature values going low to high (blue to red) from left to right represent a positive influence and vice versa. In (c), the axes are direction (positive or negative) of impact on yield (Direction) and relative magnitude of feature influence (Magnitude), and the numbers in brackets represent the number of experts. For features with less than four experts, some experts did not vote or did not understand the influence. Static features: soil water holding capacity (SM_WHC), average elevation (AVG_ELEV), average slope (AVG_SLOPE), average field size (AVG_FIELD_SIZE), total irrigated area (IRRIG_AREA_ALL), cereals irrigated area as proxy for soft wheat (IRRIG_AREA90), and maize irrigated area (IRRIG_AREA2).

importance plots showed that AEZs in the north had a positive relation with yield, which is reflected by the high average yields. Similarly, AEZs in the south had a negative relation with yield. For grain maize, the pattern was generally reversed: AEZs in the south had

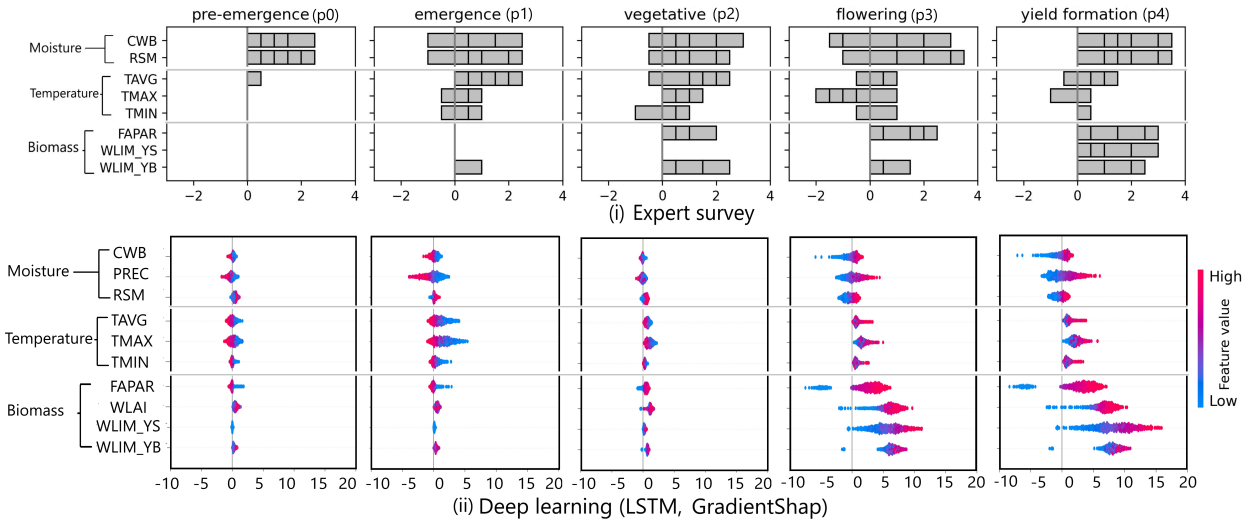
higher average yields and a positive relation indicated by importance plots.

Among other static data, experts expected soil water holding capacity (SM_WHC), average field size and irrigated areas to have a positive influence; elevation and slope to have a negative influence (*Figure 4.5a(i)*, *Figure 4.5b(i)*). For soft wheat, feature importance plots showed expected relations except for total irrigated area (*Figure 4.5a(ii)*). After discussing potential interactions among features, experts agreed that the negative relation is due to high irrigated areas in the south where average soft wheat yields are low. Some experts disagreed with the negative influence of elevation and positive influence of field size (*Figure 4.5c(i)*). The effect of elevation was later explained by the negative correlation with SM_WHC, which had a positive influence on yield. Similarly, the positive influence of field size was interpreted in conjunction with higher average yields in the north of France and lower average yields in the south. For grain maize, the negative relation with grain maize irrigated areas and positive relation with elevation were difficult to understand (*Figure 4.5b(ii)*, *Figure 4.5c(ii)*).

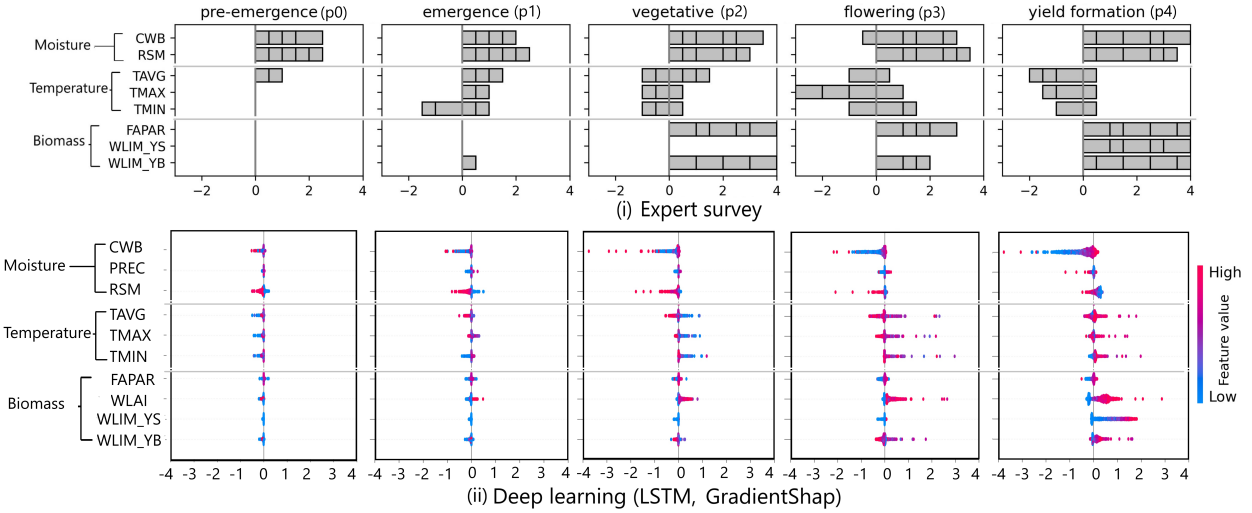
Seasonal features

The LSTM network learned seasonal features from a matrix of ten seasonal indicators and thirty dekadal values per indicator. For interpretability, importance scores for each indicator were summarized according to crop calendar periods. The importance scores showed clear relationships between features and soft wheat yields: the biomass features were important later in the season, while temperature and moisture features were influential in early season as well (*Figure 4.6a*). Furthermore, the influence of temperatures and two moisture features (precipitation (PREC) and climate water balance (CWB)) flipped from negative to positive around the vegetative period (p2). For grain maize, the influence of biomass features were still understandable, but the effects of temperature and moisture features were less clear (*Figure 4.6b*). In the rest of this section, we compare importance scores from deep learning models with expert scores and summarize the interpretability ratings from experts for each class of seasonal features.

Biomass features: Experts expected biomass features to have a positive influence on yield from vegetative period (p2) onwards (*Figure 4.6a(i)*, *Figure 4.6b(i)*). For soft wheat, all four biomass features (water-limited dry weight biomass (WLIM_YB), water-limited dry weight storage organs (WLIM_YS), water-limited leaf area index (WLAI), fraction of absorbed photosynthetically active radiation (FAPAR)) had positive relationships with yield in flowering (p3) and yield formation (p4) (*Figure 4.6a(ii)*, *Figure 4.7a(ii)*). Experts generally found these relationships interpretable (*Figure 4.6c(i)*). One expert who disagreed noted that biomass features are not good predictors of national level yields. For grain maize, biomass features had smaller importance scores (almost 10x compared to soft wheat) and the influence of WLIM_YB was not clear until yield formation (p4). WLAI had a consistent positive influence from vegetative period (p2) onwards, and WLIM_YS had the expected positive influence during yield formation (p4) (*Figure 4.6b(ii)*, *Figure 4.7b*). Most of the experts agreed with the importance of WLAI, but not other indicators (*Figure 4.6c(ii)*). Contrary to expert expectations, FAPAR did not show up as an important feature. FAPAR is less important when WLAI is above a certain threshold (see Gitelson et al. (2014)), which is likely because grain maize has high WLAI.

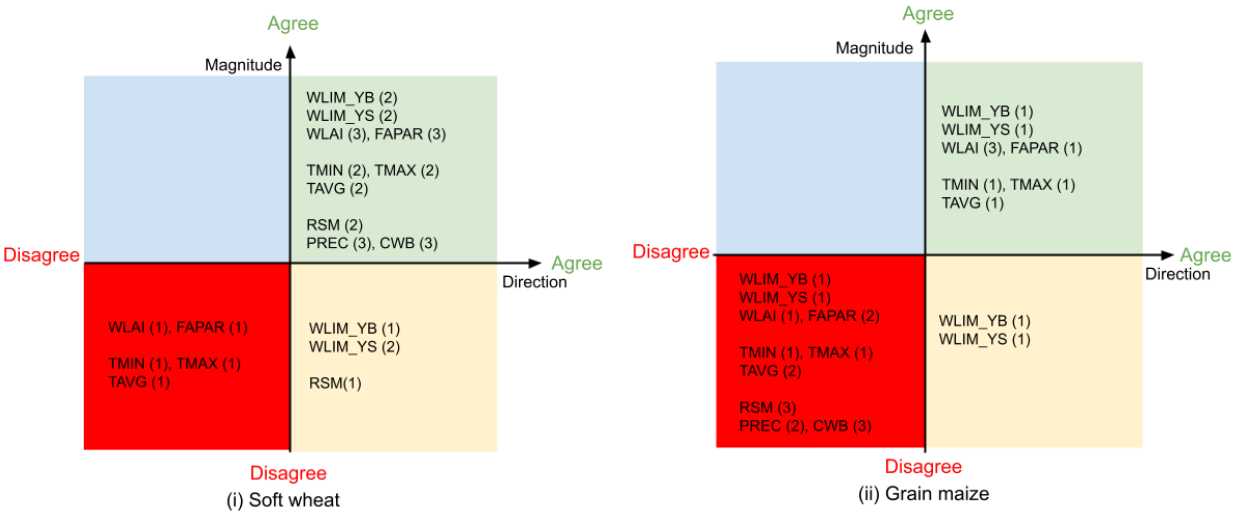


(a) Soft wheat: Feature importance from the expert survey (i) and deep learning models (ii).



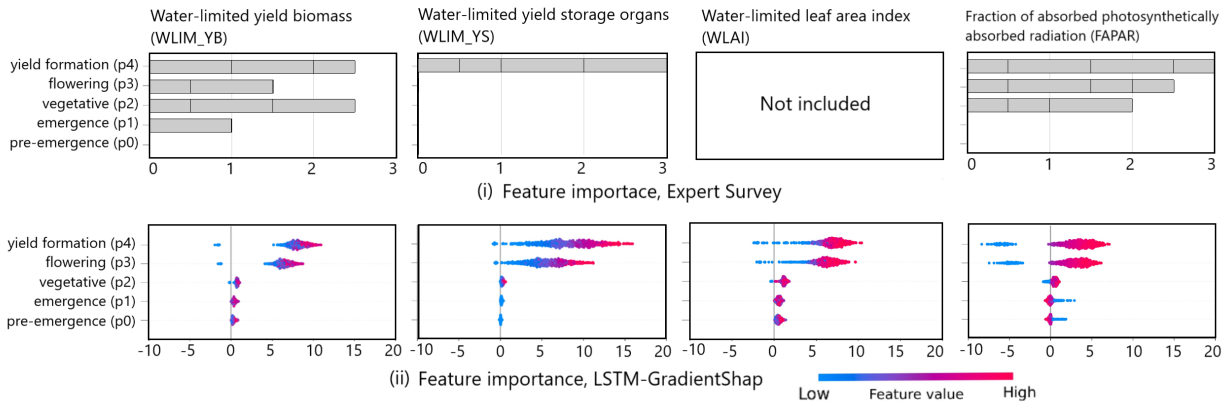
(b) Grain maize: Feature importance from the expert survey (i) and deep learning models (ii).

Figure 4.6: Importance and interpretability of seasonal features 60 days before harvest (continued in the next page).

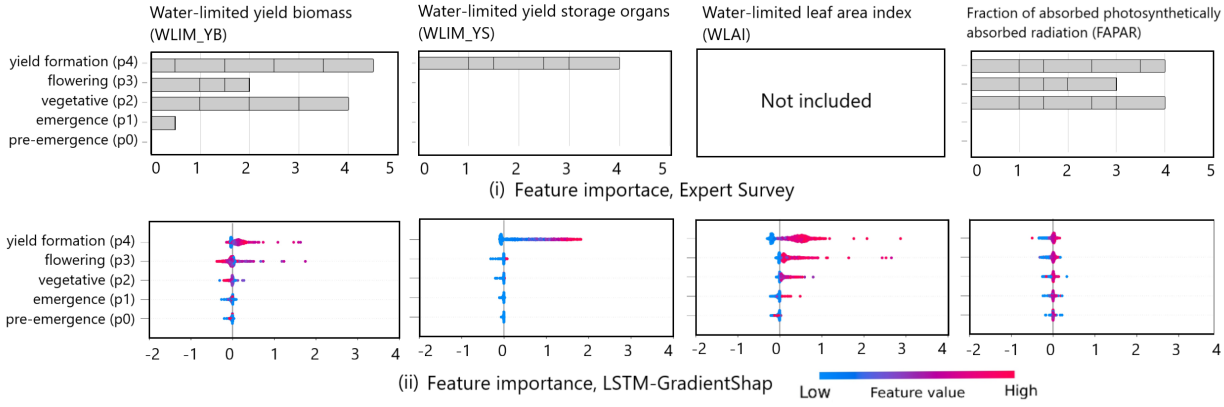


(c) Interpretability scores from experts: (i) Soft wheat, (ii) Grain maize.

Figure 4.6: Importance and interpretability of seasonal features 60 days before harvest (continued). In (a) and (b), the scale used for the expert survey was: strong negative influence (−1), mild negative influence (−0.5), no influence (0), mild positive influence (0.5), strong positive influence (1). The divisions within each bar represent how different experts voted. Importance scores from deep learning show the magnitude and direction (positive or negative) of feature influence on yield. Feature values going low to high (blue to red) from left to right represent a positive relation and vice versa. In (c), the axes are direction (positive or negative) of impact on yield (Direction) and relative magnitude of feature influence (Magnitude), and the numbers in brackets represent the number of experts. For features with less than four experts, some experts did not vote or did not understand the influence. Biomass features: water-limited dry weight biomass (WLIM_YB), water-limited dry weight storage organs (WLIM_YS), water-limited leaf area index (WLAI), Fraction of Absorbed Photosynthetically Active Radiation (FAPAR). Temperature features: minimum, maximum, average daily air temperature (TMAX, TMIN, TAVG). Moisture features: root-zone soil moisture (RSM), sum of daily precipitation (PREC), climate water balance (CWB).



(a) Soft wheat

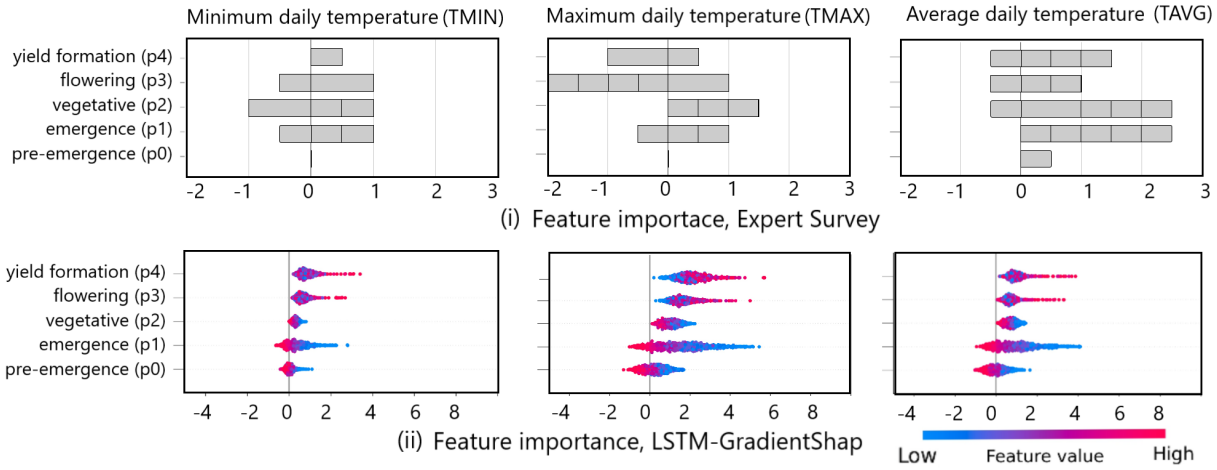


(b) Grain maize

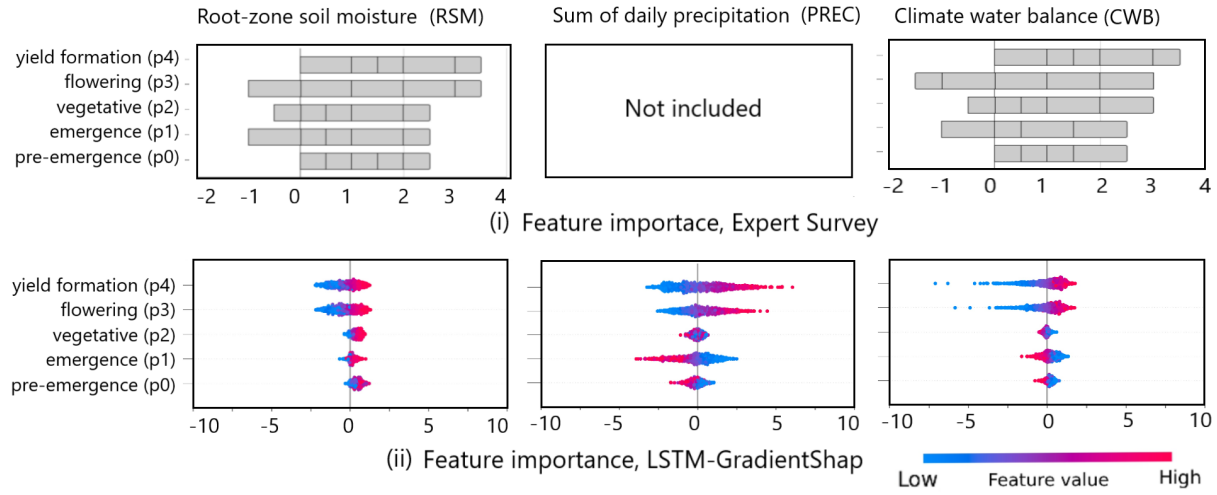
Figure 4.7: Importance of biomass features 60 days before harvest. The scale used for the expert survey was: strong negative influence (-1), mild negative influence (-0.5), no influence (0), mild positive influence (0.5), strong positive influence (1). The divisions within each bar represent how different experts voted. Importance scores from deep learning show the magnitude and direction (positive or negative) of feature influence on yield. Feature values going low to high (blue to red) from left to right represent a positive relation and vice versa.

Temperature Features: For soft wheat, temperatures (minimum (TMIN), maximum (TMAX), average daily temperature (TAVG)) had a negative relation with yield until the vegetative period (p2) and positive relation around flowering (p3) and later (*Figure 4.6a(ii)*, *Figure 4.8a(ii)*). This means lower temperatures were preferred early in the season; higher temperatures contributed to higher yields later in the season. For grain maize, the effect was mixed and very small throughout the season (*Figure 4.6b(ii)*, *Figure C.10a(ii)*). Experts expected very low temperatures around emergence (p1) and vegetative (p2) periods and very high temperatures around flowering (p3) and yield formation (p4) to have a negative influence on yields of both crops (*Figure 4.6a(i)*, *Figure 4.6b(i)*). Some of them agreed with feature importance plots because temperatures in France could be optimal for the crops (*Figure 4.6c*). In any case, effects of extreme temperatures were not captured by feature importance from deep learning (*Figure 4.6a*, *Figure 4.6b*, *Figure 4.8a*, *Figure C.10a*). High temperatures had high importance scores for soft wheat around flowering (p3) and yield formation (p4). For grain maize, temperatures did have some negative influence around flowering, but the

importance scores were very close to zero (*Figure 4.6b(ii)*, *Figure C.10a(ii)*).



(a) Temperature features



(b) Moisture features

Figure 4.8: Importance of temperature and moisture features for soft wheat 60 days before harvest. The scale used for the expert survey was: strong negative influence (-1), mild negative influence (-0.5), no influence (0), mild positive influence (0.5), strong positive influence (1). The divisions within each bar represent how different experts voted. Importance scores from deep learning show the magnitude and direction (positive or negative) of feature influence on yield. Feature values going low to high (blue to red) from left to right represent a positive relation and vice versa. Corresponding plots for grain maize can be found in *Figure C.10*

Moisture features: Experts expected the impact of root-zone soil moisture (RSM) and climate water balance (CWB) to be similar and positive throughout the season (*Figure 4.6a(i)*, *Figure 4.6b(i)*). For soft wheat, importance scores from deep learning showed different relationships. RSM had a small but consistent positive influence throughout the season. Precipitation (PREC) and CWB had a negative effect in early season and a positive effect from flowering (p3) onwards (*Figure 4.6a(ii)*, *Figure 4.8b(ii)*). Despite differences with their prior expectations, experts found these relationships interpretable (*Figure 4.6c(i)*). The only disagreement was with the lower relative importance of RSM; they expected RSM to be as important as CWB later in the season (*Figure 4.8b*). For grain maize, experts

found the negligible effect of PREC and the negative influence of RSM on yield difficult to interpret (*Figure 4.6c(ii)*). The effect of CWB was mostly positive, but many of the scores were too close to zero (*Figure 4.6b(ii)*, *Figure C.10b(ii)*). Hence the experts did not find the relationships understandable (*Figure 4.6c(ii)*).

4.4 Discussion

Previous studies have found that deep learning models can be used for crop yield forecasting (You et al., 2017; Khaki & Wang, 2019; Nevavuori et al., 2019; Wolanin et al., 2020). Deep learning can significantly improve performance when data size is large (at least around 10000). Some studies in the US have used more complex architectures, combining CNN with LSTMs or 3 dimensional CNNs with convolutional LSTMs (You et al., 2017; Khaki et al., 2020; Gavahi et al., 2021). Their results were better than standard machine learning methods, such as Ridge Regression (Hoerl & Kennard, 1970), LASSO Regression (Tibshirani, 1996), Decision Trees and Random Forests. In our case, data sizes were smaller, ranging from around 300 to 1950 labeled instances, and hence we chose simpler architectures. Even then, the performance of deep learning models was better than GBDT for soft wheat (DE), which had the largest data size, and statistically similar to GBDT for other case studies. GBDT and deep learning models outperformed the linear trend models for soft wheat but not grain maize. Similarly, spatial variability maps for France showed that soft wheat forecasts for both deep learning models and GBDT were closer to reported statistics than grain maize forecasts. The main takeaway is that deep learning can automatically extract features that perform similarly to expert-designed features. In other words, manual feature design does not make model performance significantly better. Therefore, deep learning provides a solution to the limitations of manual feature design at large scale.

Our results indicated that crop yield forecasts from deep learning models can be explained using post-hoc feature attribution methods, especially when forecasting accuracy is high. Feature attributions were found to be interpretable for trend features (*Figure C.9*), most static features (*Figure 4.5*) and biomass features (*Figure 4.7*). Similarly, attributions of moisture features correctly captured the relationships in different periods of soft wheat season (*Figure 4.8b*). Exceptions included the effect of extreme temperatures on both crops (*Figure 4.8a*, *Figure C.10a*) and irrigation and moisture on grain maize (*Figure 4.5*, *Figure C.10b*). Our framework and models did not account for the relative rarity of extreme temperature events. As a result, both low temperatures in early season and high temperatures later in the season had positive rather than negative importance. The influence of moisture was affected by whether the crop was irrigated, and the influence of irrigated areas was sometimes unclear because the data was static. Despite some limitations, deep learning models captured some relationships not expected by experts. For example, experts did not consider the effects of high temperatures in early season; feature importance from deep learning showed a negative effect of high temperatures (*Figure 4.8a*). Similarly, experts expected a consistent positive effect of RSM and CWB throughout the season, but feature importance from deep learning showed that CWB has a negative effect on soft wheat yields in early season and a positive effect from flowering (p3) onwards (*Figure 4.8b*). Experts later agreed with these relationships. In some cases, experts did not agree with the relationships or relative importance. Some of the

disagreements stemmed from different perspectives on the relationships between features and yield. For example, some experts disagreed with the negative relation between elevation and soft wheat yields (*Figure 4.5a*), but they interpreted it based on interactions with soil water holding capacity. Similarly, some experts found small importance values of FAPAR for grain maize hard to understand, while others explained it based on high WLAI (*Figure 4.6c(ii)*). Overall, feature influence matched expert knowledge and experience more for soft wheat than for grain maize. For grain maize, the linear trend model performed as good as GBDT and LSTM models. Therefore, comparisons of GBDT and LSTM forecasts with reported yields (Figure A.7, A.8) and low importance scores of seasonal features tell a similar story. The interannual variability related to seasonal features may be low, hence the relationships between these features and yield were less clear. In addition, as noted by one of the experts, modeling irrigated and non-irrigated systems together could have been another confounding factor. This limitation comes from regional yield statistics, which in most countries do not have separate values for irrigated and non-irrigated systems. Separate models for the two systems could produce more interpretable relationships.

Many factors contribute to model interpretability and agreement or disagreement with experts. First, experts expected certain relationships based on prior knowledge and experience. The deep learning framework for regional crop yield forecasting was different from the setups they have previously used. Similarly, relationships actually present in data can be complex and difficult to verify. Second, the choice of neural network architectures and models was by no means the best. Neural networks have a large number of parameters and need large amounts of training data to generalize well (Molnar, 2022). We selected simpler architectures to balance model capacity and complexity. Third, feature attribution methods used also have their limitations. The additive explanations produced by GradientShap were easy to understand, but they only approximated model behavior. Interpretable neural networks are not yet common. Finally, presentation of feature importance or explanations to experts is a social process (Miller, 2019) and affects how they understand model behavior. We presented positive or negative contributions of individual features to yield forecasts. This approach ignores interactions among features although certain combined interactions were considered in the interactive session with experts.

Analyzing interpretability is difficult and hence we relied on a qualitative analysis of interpretability by experts. Robust hypothesis testing would be required to establish whether explanations of deep learning models match existing knowledge or provide new insights (McGovern et al., 2019). Tests could ask human subjects to select feature attributions or explanations that correspond to given input and output, or suggest an output given inputs and feature attributions (Doshi-Velez & Kim, 2017). Such tests require many participants and significant time commitment from the experts (Narayanan et al., 2018). Therefore, we chose a simpler approach of expert surveys followed by an interactive evaluation. Interpretability scores showed that experts did not always agree when selecting important predictors (*Figures 4.5, 4.6, 4.7, 4.8, C.10*) or scoring interpretability of feature importance (*Figures 4.5, 4.6*). Despite such differences, it is important to engage humans in assessment of interpretability. Models can be interpretable in many ways and experts may view the same explanations differently. The fact that deep learning models did not find certain features important does not mean they are physically unimportant (McGovern et al., 2019). Overall, we developed

and tested a method to assess model interpretability with feedback from human experts that went beyond feature importance plots, which are often not interpretable on their own.

Deep learning methods are nowadays common in agricultural applications including crop yield forecasting (Van Klompenburg et al., 2020; Kamilaris & Prenafeta-Boldú, 2018). The need for model interpretability will continue to grow with the widespread use of deep learning. In this paper, we selected human experts familiar with the input data and the agronomic principles driving crop growth and development to evaluate interpretability. This familiarity allowed us to simplify the process of explaining feature attributions to them. Experts were able to understand feature importance plots and provide judgments about their interpretability. Hence we were able to focus more on understanding how well deep learning captures expected agronomic relationships, and less on social and cognitive factors affecting interpretability (Narayanan et al., 2018). Nevertheless, social and cognitive factors are important and interpretability analysis should not be limited to visual feature importance plots. Other stakeholders that use crop yield forecasts for decision making may need a different method for explaining yield forecasting models to them. More accurate assessment of interpretability is possible with iterative improvement of the evaluation process. Deep learning models may produce accurate crop yield forecasts, but they are useful only when they can be trusted and used in real-world applications. Interpretability will play an important role in bridging the gap between model building and decision making.

4.5 Conclusions

We evaluated the performance and interpretability of deep learning models for soft wheat and grain maize at regional level in Europe. Performance was found to be statistically similar to or better (in one case) than a standard machine learning algorithm that relies on expert-designed features. Therefore, deep learning provides benefits of automatic feature learning that can address limitations of manual feature design at large scale. Similarly, feature attribution methods provide post-hoc explanations for model predictions that are generally interpretable, especially when forecasting accuracy is high. Such explanations indicate how each feature contributes to the yield prediction, but not their interacting effects. Feature importance scores from deep learning models correctly captured the influence of most static features, yield trend and biomass features on crop yield. In some cases, they also identified relationships not expected by experts (e.g. the negative effect of climate water balance on soft wheat in early season). On the other hand, deep learning models struggled to capture the impact of extreme temperatures on both crops and the effect of irrigation and moisture on grain maize. We found that human assessment of interpretability is challenging, but nonetheless important. Limitations exist in data, model building, extraction of feature attributions and presentation of explanations to human stakeholders. Some of these limitations can be addressed by continued engagement of human stakeholders and iterative improvement of the evaluation process. Interpretability of models is crucial for building trust in them and using them to guide decision making.

Chapter 5

A weakly supervised framework for high-resolution crop yield forecasts

This chapter is based on:

Dilli R. Paudel, Diego Marcos, Allard de Wit, Hendrik Boogaard, and Ioannis N. Athanasiadis. A weakly supervised framework for high resolution crop yield forecasts. *Environmental Research Letters* (Under Review), 2023b.

Abstract

Predictor inputs and label data for crop yield forecasting are not always available at the same spatial resolution. We propose a deep learning framework that uses high resolution inputs and low resolution labels to produce crop yield forecasts for both spatial levels. The forecasting model is calibrated by weak supervision from low resolution crop area and yield statistics. We evaluated the framework by disaggregating regional yields in Europe from parent statistical regions to sub-regions for five countries (Germany, Spain, France, Hungary, Italy) and two crops (soft wheat and potatoes). Similarly, county-level corn yields and crop areas were disaggregated to 10-km grid level in the United States (US). Performance of weakly supervised models was compared with naive disaggregation models, which assigned the low resolution forecast for a region or county to all high resolution subregions or grids, and strongly supervised models trained with high resolution yield labels. The weakly supervised (WS) models in Europe were statistically similar to strongly supervised models and better than the naive trend model. In the US, the WS model forecasts for 10-km grids were significantly better than both the naive models and strongly supervised models. Based on Kendall's rank correlation coefficient, the WS model forecasts captured significant amounts of high resolution yield variability. Looking at an extreme harvest and the following season's harvest, we found that the WS model not using the low resolution trend captured high resolution yield differences better. Combining information from two versions of the WS model – the WS Trend model (using low resolution yield trend) and the WS No Trend model (not using yield trend) – provided good estimates of yields as well as spatial differences among sub-regions or grids. Higher resolution crop yield forecasts are useful to policymakers and other stakeholders for local analysis and monitoring. Weakly supervised deep learning methods provide a way to produce such forecasts even in the absence of high resolution yield data.

5.1 Introduction

Crop yield forecasts for a country or region are more reliable when they can be explained using yield variability at high spatial resolutions. High resolution crop yield forecasts improve the effectiveness of policy interventions targeted to food security, agricultural production and resource sustainability (You et al., 2014). They are also useful to farmers, commodity traders and other local stakeholders involved in processing, transporting and selling farm products.

Predictor inputs and label data for crop yield forecasting are often not available at the same spatial resolution. Label data (e.g. yield statistics) are published for administrative regions, such as counties, states or provinces. Weather inputs are available at grid-level (EC-JRC, 2022; Thornton et al., 2020) and soil and remote sensing data at sub-kilometer resolutions (ESDAC, 2021; Poggio et al., 2021; Copernicus ESA, 2022). Common statistical and machine learning methods require both inputs and labels at the same spatial level. When each data point has a corresponding label at the same spatial level, the model gets strong supervision from the label. This means strongly supervised models can be built only at the administrative levels where yield statistics are published. Therefore, predictor inputs are aggregated to the same level as yield data. In the absence of strong supervision from high resolution labels, standard statistical and machine learning methods cannot produce high resolution forecasts. High resolution yield data may be unavailable for various reasons. For example, yield statistics are rarely published at grid level, and farm or field level yield data are not openly available. When label data is not available at high resolution, learning is still possible using high resolution inputs and low resolution labels. Such learning is called weakly supervised learning (Zhou, 2018). Deep learning models can be weakly supervised by using high resolution inputs to produce high resolution yield forecasts, which can be aggregated to low resolution and compared with the labels there. Weakly supervised models limit the spatial aggregation required for input data and produce high resolution yield forecasts even when there is a shortage or absence of high resolution yield data.

Many studies have used deep learning for crop yield forecasting (Fan et al., 2021; Shahhosseini et al., 2021; Wolanin et al., 2020; Khaki et al., 2020), but they do not disaggregate yields to high resolutions. Folberth et al. (2019) attempted disaggregation using Gradient Boosting (Friedman, 2001) and Random Forests (Breiman, 2001), which are standard machine learning methods. The models were trained on low resolution inputs and labels and later applied them to higher resolution inputs. This approach assumes that low resolution and high resolution data come from the same distribution, which is unlikely. Other methods of disaggregating crop yields exist, for example, area-to-point kriging (Brus et al., 2018; Steinbuch et al., 2020) and spatial allocation based on cross-entropy method (You et al., 2014) or remote sensing indicators (Shirsath et al., 2020; Kang & Özdoğan, 2019). We draw inspiration from Jacobs et al. (2018), who trained a convolutional neural network and an aggregation layer to predict pixel-level population density from high resolution satellite images and low resolution density statistics. To our knowledge, weakly supervised methods have been used for computer vision tasks, such as detecting or counting objects (Wang et al., 2022; Chandra et al., 2020; Ghosal et al., 2019), but not to disaggregate crop yields to high resolution.

We propose a weakly supervised deep learning framework that produces crop yield forecasts for both high and low resolutions using high resolution inputs and low resolution labels. The framework is weakly supervised because models to produce high resolution yield forecasts are trained with low resolution labels. Since crop area statistics may be unavailable at high resolution, the framework also estimates the crop area weights required for aggregation. Our objective was to build and evaluate crop yield forecasting models that can produce high resolution yield forecasts even when high resolution yields and crop areas are unavailable. This objective was divided into three sub-objectives. First, we assessed the ability of weakly supervised models to disaggregate crop yields from low to high resolution. Second, we evaluated the quality of low resolution yield forecasts produced using high resolution inputs. The framework was validated in two different settings (Europe and the United States), both in terms of agro-environmental factors and spatial resolution. In Europe, labels from NUTS2 regions (low resolution) and inputs from constituent NUTS3 regions (high resolution) were used to produce yield forecasts for both NUTS2 and NUTS3 regions. NUTS (Nomenclature of Territorial Units for Statistics) is a hierarchical system of dividing the territory of the European Union for statistics and policy (Eurostat, 2016b). In the United States (US), labels from counties (low resolution) and inputs from constituent 10-km grids (high resolution) were used to produce yield forecasts for both grids and counties. For weak supervision to be useful to stakeholders, high resolution forecasts need to capture some yield variability among NUTS3 regions within a NUTS2 region or grids within a county. Therefore, as the third sub-objective, we analyzed how well weak supervision captures yield variability at high resolution for an extreme harvest and the following season's harvest. Our analysis included two crops (soft wheat and potatoes) and five countries (Germany, Spain, France, Hungary, Italy) in Europe and corn in the US.

The rest of the chapter is structured as follows: Section 5.2 describes data and methods; Section 5.3 presents the results; and Section 5.4 discusses our findings and outlines directions for future work. *Appendix D* provides additional details and supporting evidence not included in Section 5.2, 5.3 and 5.4.

5.2 Methods

Our objective was to build and evaluate weakly supervised models that use high resolution inputs and low resolution labels to produce crop yield forecasts for both spatial resolutions. To address the first two sub-objectives, performance of the weakly supervised models was compared with two types of models: strongly supervised models and naive disaggregation models. Strongly supervised models were built at high and low resolutions with inputs and labels from the same spatial level. Forecasts from the strongly supervised low resolution models were naively disaggregated to high resolution by assigning the forecast for parent region or county to all constituent sub-regions or grids. The naive disaggregation models served as the "null" models with no prediction skill, while the strongly supervised models provided the bar to beat. At high resolution, the performance of WS models was compared with the naive disaggregation models and high resolution strongly supervised models. At low resolution, the WS models were compared with low resolution strongly supervised models. For the third sub-objective, we analyzed the spatial yield variability among NUTS3 regions

within a NUTS2 region and among 10-km grids within a county for selected test years.

5.2.1 Data

European data came from the MARS Crop Yield Forecasting System of the European Commission's Joint Research Centre (MARSWiki, 2021) and the Eurostat (Eurostat, 2021a). The data covered two crops (soft wheat and potatoes) and five countries: Germany (DE), Spain (ES), France (FR), Hungary (HU) and Italy (IT). Data from all countries was combined to build one prediction model per crop. The decision to combine data from multiple countries was motivated by the small number of NUTS2 labels available for weak supervision. Seasonal data included outputs of the WOFOST crop model (van Diepen et al., 1989; Supit et al., 1994; de Wit et al., 2019), weather variables and remote sensing indicators aggregated to NUTS3 and NUTS2 (*Table 5.1*). The yield trend was captured using yield values of five previous years. Static differences among regions were captured by soil water holding capacity and agro-environmental features, such as elevation, slope, field sizes, irrigated area (Paudel et al., 2022a). In addition, agro-environmental zones and countries were added as categorical variables to account for other agro-climatic and administrative differences. Reported yield and crop area statistics served as labels. In most cases, we had data from 1999 to 2018. The most recent 30% of the years were allocated to the test set. From the remaining 70% training years, 5 most recent years were used for a sliding-window 5-fold validation (*Figure 3.3*) to optimize hyperparameters.

For the US, county crop yields and crop areas were exported from the National Agricultural Statistics Service (NASS) of the US Department of Agriculture (USDA) (USDA-NASS, 2022). 10-km grid inputs came from the Climate Data Store of the Copernicus Climate Change Service (Copernicus CDS, 2022) and Copernicus Global Land Service (Copernicus GLS, 2020). Seasonal inputs consisted of crop productivity indicators produced by a simple crop model (de Wit et al., 2022) and weather and remote sensing indicators (*Table D.1*). The only static input was soil water holding capacity. Grid-level yields published by Deines et al. (2021) were considered ground-truths for grid-level validation. These yields are not yield statistics, but produced with the Scalable Crop Yield Mapper (SCYM) methodology of Lobell et al. (2015). We used them for validation because official yield statistics are not available for 10-km grids. Overall, we had data from 2000 to 2018. Training and test splits followed a 70-30% scheme, similar to European data above. Since the data size was significantly large (about ten times compared to the European data), hyperparameter optimization used a single validation set (five most recent years from the training set), instead of a 5-fold sliding validation.

5.2.2 The weakly supervised framework

The weakly supervised framework modified the deep learning framework from *Chapter 4* (Paudel et al., 2023a) to include low resolution trend features and an aggregation layer (*Figure 5.1*). We considered Long Short-Term Memory (LSTM) and 1-dimensional convolutional neural networks (1DCNN) architectures to extract features from seasonal data. Both LSTMs and 1DCNNs have been used in literature with sequential or time series data (You et al., 2017; Khaki et al., 2020). LSTM was selected due to its superior validation set performance (*Figure D.1*). Seasonal data at high resolution (NUTS3 in Europe and 10 km grids in the US),

Table 5.1: Data sources for Europe. In Europe, data sources covered two crops and five countries: soft wheat (DE, ES, FR, IT) and potatoes (DE, FR, HU, IT). The US data covered corn. Data sources for the US are shown in *Table D.1.*

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 index ($m^2\ m^{-2}$), Development stage ($0-200$), root-zone soil moisture as % of soil water holding capacity, sum of water limited transpiration (cm). **Source**: MCYFS. See Lecerf et al. (2019).
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), climate water balance = (PREC - ET0) (mm), sum of daily global incoming shortwave radiation ($KJ\ m^{-2}\ d^{-1}$). **Source**: MCYFS. See Lecerf et al. (2019).
Remote Sensing	Fraction of Absorbed Photosynthetically Active Radiation (Smoothed) (FAPAR). **Source**: MCYFS. See Copernicus GLS (2020).
GAES	Agro-environmental zone identifiers. **Source**: Global agro-environmental stratification (Mücher et al., 2016).
Crop Areas	Crop production areas (ha). **Source**: Eurostat (Eurostat, 2021a) and MCYFS (EC-JRC, 2022).
Irrigated area	Irrigated total area and irrigated crop-specific area (ha). **Source**: EC-JRC (2022).
Elevation, slope	Average and standard deviation elevation (m) and slope ($degrees$) **Source**: USGS-EROS (2021).
Soil	Soil water holding capacity. **Source**: MCYFS. See Lecerf et al. (2019).
Field Size	Average and standard deviation (ha). **Source**: Lesiv et al. (2019).
Yield	Yield at NUTS3 level ($t\ ha^{-}1$). NUTS2 level yields were produced by aggregating NUTS3 yields. **Source**: FR-Agreste (2020); DE-RegionalStatistiks (2020); Eurostat (2021a); EC-JRC (2022).

including crop productivity indicators, weather and remote sensing indicators, was processed by LSTM. Features from the LSTM layers, together with static agro-environmental data and yield trend features (based on NUTS2 yields in Europe and county yields in the US), were passed to the output layer (*Figure 5.1*). The output layer produced high resolution (NUTS3 or 10-km grid) yield forecasts and crop area fractions. We believe remote sensing indicators can help predict crop area fractions (crop production area/total land area), but not the absolute crop areas. The aggregation layer multiplied predicted crop area fractions with land areas to produce NUTS3 or grid-level crop areas, and used them to calculate crop area weights for aggregation. High resolution yield forecasts were then aggregated to low resolution. The framework was supervised with NUTS2 or county-level yields and crop areas. Data from all NUTS3 regions within a NUTS2 region formed a batch to enable aggregation of NUTS3 forecasts. Similarly, data from all 10-km grids within a county formed a batch to aggregate grid forecasts.

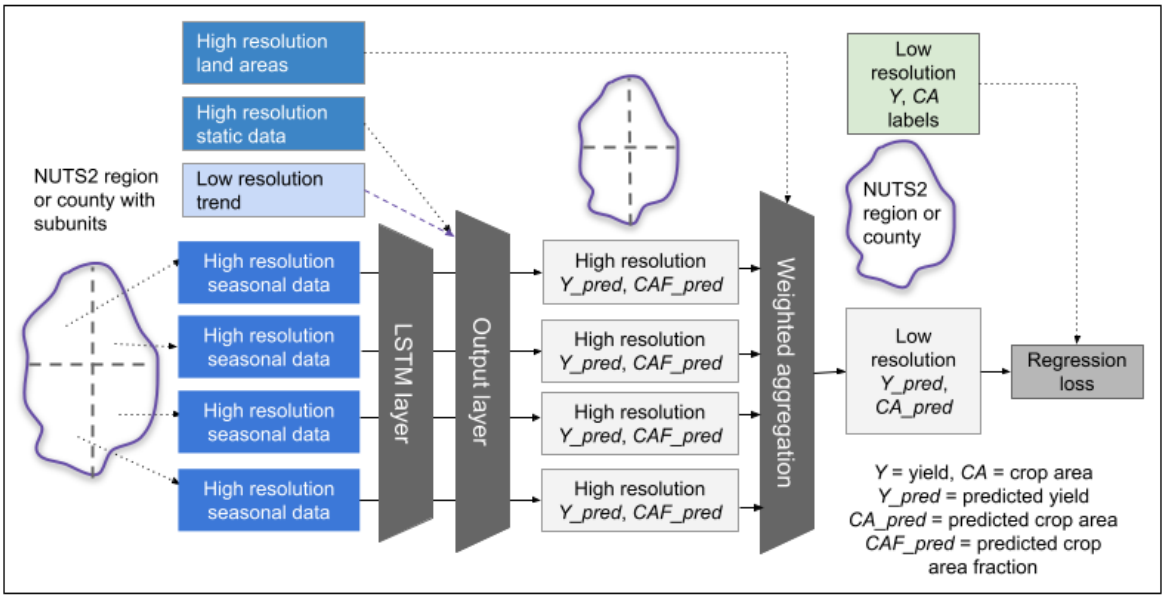


Figure 5.1: Weakly supervised framework to produce high resolution crop yield forecasts. The framework used seasonal and static data from high resolution and yield trend from low resolution to produce high resolution forecasts, which were then aggregated to low resolution. The framework was weakly supervised by comparing the aggregated forecasts with low resolution yields and crop areas. In Europe, low resolution refers to NUTS2 regions, and high resolution refers to NUTS3 regions. In the US, counties and 10-km grids represent low and high resolutions respectively.

5.2.3 Evaluation

Forecasts from the weakly supervised (WS) model were evaluated at both spatial resolutions: NUTS3 and NUTS2 in Europe; and 10-km grids and counties in the US. For performance comparison, three types of strongly supervised models were built at both high and low spatial resolutions: linear trend, Gradient-Boosted Decision Trees (GBDT) and Long Short-Term Memory (LSTM) networks (*Figure 5.2*). The trend models fitted a line through yield values of five previous years. GBDT represented a standard machine learning algorithm that was trained using expertise-based features. Features for GBDT were designed the same way as in (Paudel et al., 2022a). LSTM represented a strongly supervised deep learning method. The LSTM framework was adapted from Paudel et al. (2023a), with the exception that twelve (instead of ten) seasonal indicators were selected to match those used for GBDT feature design. Low resolution (LR) and high resolution (HR) models used inputs and labels at the corresponding spatial resolution. For example, features and yields for the LR GBDT model came from NUTS2 regions in Europe and counties in the US. Similarly, inputs and yields for the HR LSTM model came from NUTS3 regions in Europe and 10-km grids in the US.

GBDT, LSTM and the WS model predictions were collected from ten models to account for the effect of random seeds or random weight initializations. We used the average normalized root mean squared errors (NRMSE), normalized by average yield of the test set, of ten models to compare performance. Variance and outliers were analyzed using boxplots of prediction residuals (predicted yield - reported yield). Significance of model performance was evaluated using the Mann-Whitney U test (Mann & Whitney, 1947), which is a non-parametric version

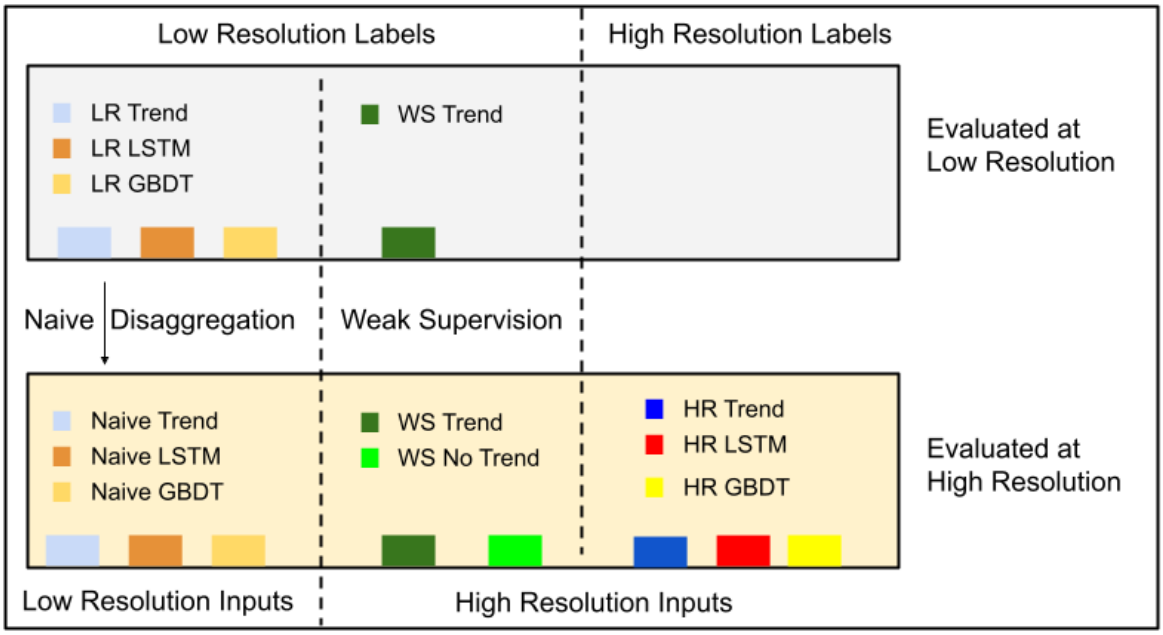


Figure 5.2: Evaluation framework. The low resolution (LR) models and high resolution (HR) models were trained with inputs and labels from the same (low or high) resolution. The weakly supervised (WS) models were trained with high resolution inputs and low resolution labels. The naive disaggregation models assigned the forecast of parent region or county to all constituent sub-regions or grids. High resolution forecasts from weakly supervised (WS) models were compared with naive disaggregation models and the HR models. Low resolution forecasts from the WS models were compared with the LR models.

of Student's t-Test for independent samples. Prediction residuals used for boxplots and statistical tests were averaged across the ten models.

Evaluation of low resolution yield forecasts

At low resolution (NUTS2 in Europe, counties in the US), WS model forecasts were compared with the strongly supervised low resolution models: LR trend, LR GBDT and LR LSTM models. All forecasts were made 60 days before harvest.

Evaluation of high resolution yield forecasts

The high resolution forecasts of the WS models were compared with two sets of models. The first set included the naive disaggregation models: naive Trend, naive GBDT and naive LSTM models. The naive disaggregation models assigned the forecast of LR models for a parent NUTS2 region or county to the constituent NUTS3 regions or grids. This means, for a particular test year, the forecasts for all NUTS3 regions within a NUTS2 region were the same. Comparison with the naive disaggregation models served as a test of prediction skill. The second set included the strongly supervised models at high resolution: HR Trend, HR GBDT and HR LSTM. Unlike the naive disaggregation models and the WS models, the HR models were trained with high resolution labels. Therefore, the HR models provided the high bars for performance. The HR GBDT, HR LSTM and the WS models used the same high resolution inputs, except yield trend features. The HR models had access to the high

resolution yield trend; the WS models only had the low resolution trend. In Europe, NUTS3 yield statistics were used to train the HR models. In the US, the HR models were trained with modeled grid-level yields from Deines et al. (2021). All forecasts were made 60 days before harvest.

Spatial variability of high resolution forecasts

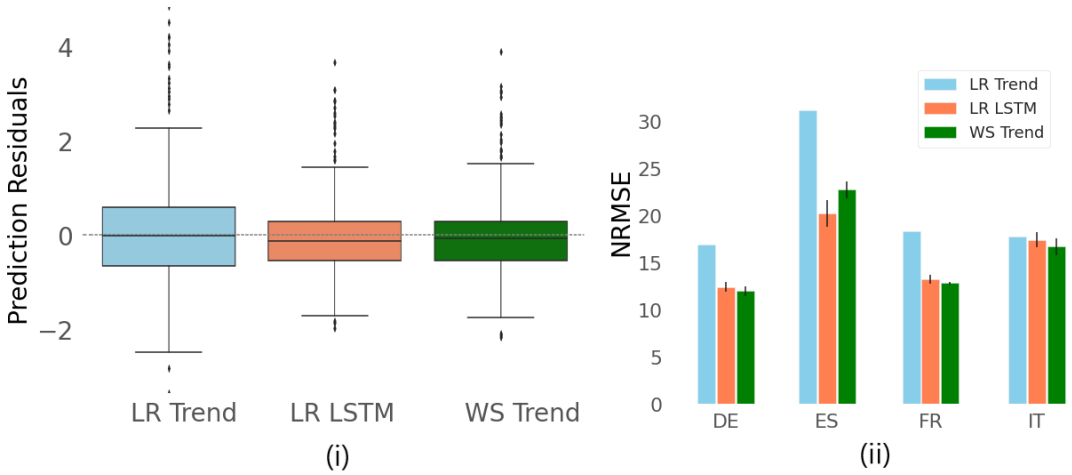
In addition to performance comparisons, we also analyzed the spatial variability of NUTS3 and grid-level forecasts from disaggregation models (naive trend and WS models). A significant part of yield variability is explained using the yield trend attributed to factors such as technological improvements (see Lecerf et al. (2019)). In the version of WS model shown in (*Figure 5.1*), we expected the low resolution trend features to make the model more accurate, but suppress spatial variability at high resolution. Therefore, we ran another version of the WS model without NUTS2 or county trend to learn yield differences among constituent NUTS3 regions or 10-km grids. The two versions are called WS Trend model and WS No Trend model (*Figure 5.2*). Kendall's rank correlation coefficient (or Kendall's tau, Kendall (1938)) was used to quantify the skill to capture spatial yield variability among NUTS3 regions for soft wheat in Europe and among 10-km grids for corn in the US. For example, NUTS3 yield forecasts within the same NUTS2 region were ranked and compared with the ranking of NUTS3 yields to compute Kendall's tau. Kendall's tau of the WS models were compared with those of HR models. A high correlation (and significance based on p-value) would show that forecasts captured the relative differences in yields among NUTS3 regions. To illustrate the spatial yield variability captured by different models, maps of yield forecasts vs yields were plotted for an extreme harvest and the following season's harvest. In Europe, NUTS3 regions were selected based on maximum acreage for soft wheat (FR) and years (2016 and 2017) based on significant yield losses reported in the north of FR in 2016 (see Ben-Ari et al. (2018)). In the US, spatial variability of 10-km grid yields was analyzed for 2012, when there was a severe drought (Rippey, 2015), and 2013.

5.3 Results

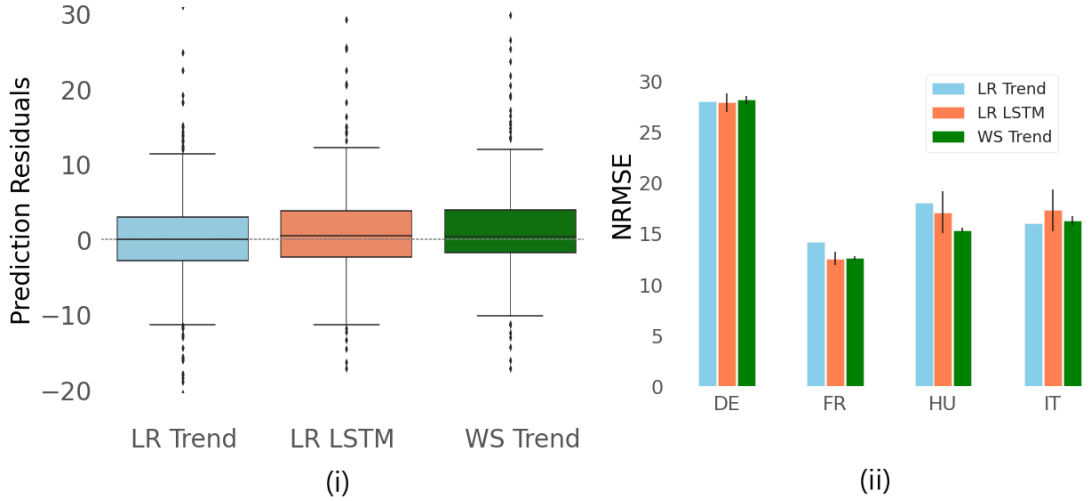
In this section, performance comparison results are shown for the WS Trend model and spatial variability analysis includes the WS No Trend model as well. Among the naive disaggregation models and strongly supervised (LR and HR) models, the GBDT models are not shown because their forecasts were statistically similar to LSTM forecasts.

5.3.1 Evaluation of low resolution yield forecasts

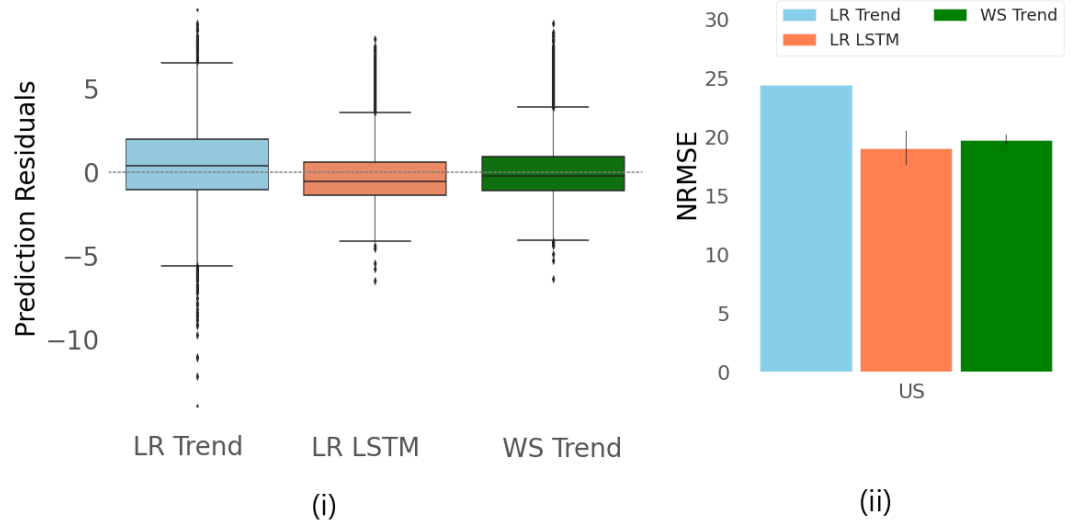
For soft wheat at NUTS2, the WS Trend model forecasts were statistically similar to those of all LR models (p-values between 0.1424 and 0.8534) and the LR models were also statistically similar to each other (*Table D.2*). Box plots showed that both WS Trend and LR LSTM underestimated yields compared to LR Trend, but they had smaller residuals and fewer outliers. Per-country NRMSEs were also lower for LR LSTM and WS Trend (*Figure 5.3a*). For potatoes, the LR Trend model was statistically better than LR GBDT and LR LSTM models (p-values 0.0037 and 0.0128) and the WS Trend model (p-value 0.002) (*Table D.3*),



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



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



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

Figure 5.3: Evaluation of low resolution forecasts 60 days before harvest. Prediction residuals used for the boxplots and per-country NRMSEs were averaged across ten models. The whiskers in the bar plots indicate the standard deviation of NRMSEs for the ten models.

but these differences were less evident in the box plots and per-country NRMSEs (*Figure 5.3b*). Overall, the WS Trend model was not significantly better than the LR models despite using high resolution inputs. In the US, high resolution inputs did make the WS Trend model significantly better; county-level corn forecasts were better than those from all LR models (p-values near 0) (*Table D.4*). The LR LSTM had a similar NRMSE, but underestimated the yields more compared to the WS Trend model (*Figure 5.3c*).

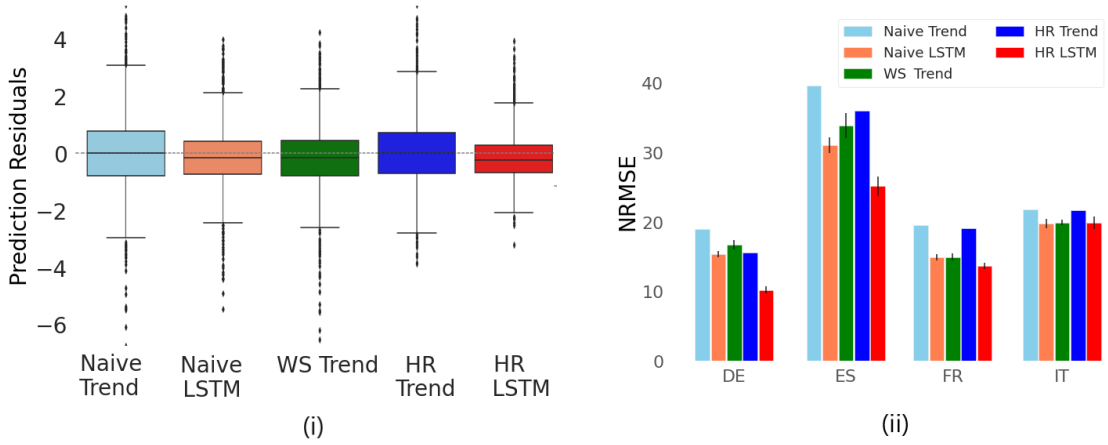
5.3.2 Evaluation of high resolution yield forecasts

For NUTS3 soft wheat forecasts in Europe, the WS Trend model was significantly better than the Naive Trend and HR Trend models (p-values near zero) and similar to Naive GBDT and Naive LSTM as well as HR GBDT and HR LSTM models (p-values 0.1202 and 0.731) (*Table D.5*). Interestingly, HR LSTM and HR GBDT were statistically similar to Naive GBDT and Naive LSTM (p-values between 0.0741 and 0.2797), and HR Trend was also similar to Naive Trend (p-value 0.369). Box plots of prediction residuals agreed with the statistical test results: HR Trend looked similar to Naive Trend and WS Trend, Naive LSTM and HR LSTM looked similar to each other. WS Trend had lower NRMSEs than Naive Trend and HR Trend but higher than Naive LSTM for DE and ES (*Figure 5.4a*). For potatoes, the WS Trend model was still better than the Naive Trend model (p-value 0.0228), but statistically similar to other naive models and the HR models, including the HR Trend model (p-value 0.0756). Once again, HR LSTM and HR GBDT were statistically similar to both Naive GBDT and Naive LSTM (p-values between 0.2854 and 0.9417), and HR Trend was similar to Naive Trend (p-value 0.3523) (*Table D.6*). Box plots and NRMSEs showed that WS Trend model was similar to Naive LSTM, but worse than HR Trend and HR LSTM (*Figure 5.4b*). Overall, performance results did not show weak supervision to be better than naive disaggregation to NUTS3. For corn in the US, weak supervision did produce better high resolution forecasts than the naive disaggregation models as well as the HR models (all p-values near zero) (*Table D.7*). Box plots showed that WS Trend had prediction residuals smaller than Naive Trend and HR Trend and closer to zero than Naive LSTM and also HR LSTM. NRMSEs were the lowest for WS Trend and HR LSTM (*Figure 5.4c*).

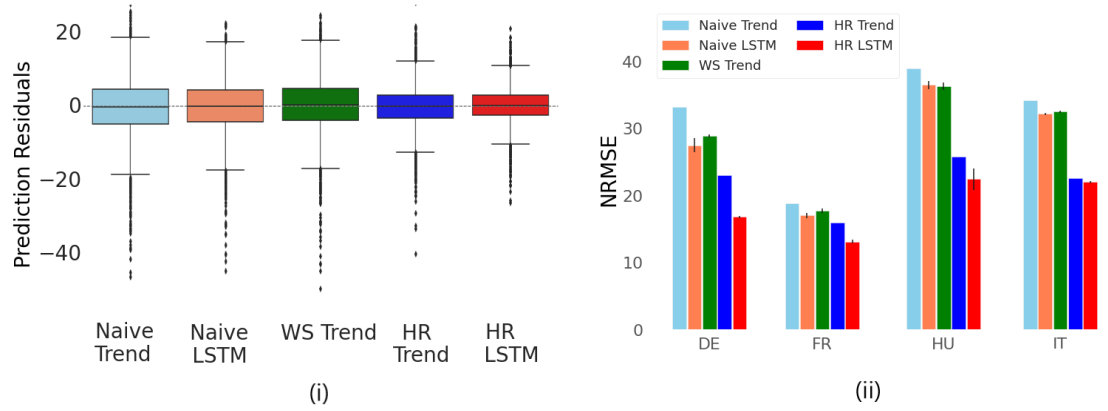
5.3.3 Spatial variability of high resolution forecasts

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

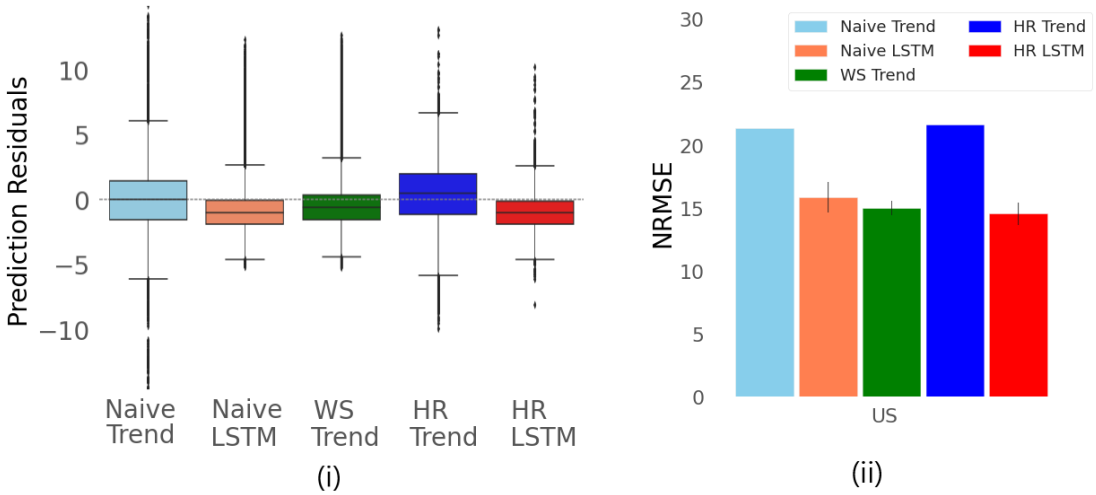
For soft wheat in France, the Naive Trend model predicted higher yields in 2016, with an



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



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



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

Figure 5.4: Evaluation of high resolution forecasts 60 days before harvest. Prediction residuals used for the boxplots and per-country NRMSEs were averaged across ten models. The whiskers in the bar plots indicate the standard deviation of NRMSEs for the ten models.

average prediction residual of 0.717. The maps showed that Naive LSTM, WS Trend and HR LSTM models were influenced by the yield trend and overestimated yields (average prediction residuals: 0.36, 0.28 and 0.39 respectively). Their forecasts looked quite similar (*Figure 5.5a*). The WS No Trend model captured the yield losses better with an average residual of 0.064. In 2017, the Naive Trend model was heavily influenced by the 2016 yields (especially in the middle: FRI3, FRB0, FRC1), while all the other models predicted values closer to the reported yields (*Figure 5.5b*). As expected, the naive models provided no information about NUTS3 level yield variability. The WS Trend model forecasts looked similar to Naive LSTM forecasts and did not show visible differences among NUTS3 sub-regions within NUTS2 regions. The WS No Trend model captured such differences better, and the forecasts looked similar to HR LSTM forecasts. Because it did not use yield trend, the WS No Trend model underestimated the yields: the average prediction residual was -0.516 compared to -0.135 for the WS Trend model and -0.061 for HR LSTM model.

For corn in the US, we evaluated spatial variability for 2012 and 2013 because of the well-known drought of 2012 (Rippey, 2015), and we only compared Naive Trend, WS Trend and WS No Trend models. For 2012, the WS No Trend model captured the yield losses better, with the WS Trend model overestimating the yields (*Figure D.2*). The mean prediction residual was 4.0 for Naive Trend, 3.99 for WS Trend and 2.51 for WS No Trend. The WS No Trend model also captured differences among grids better than the WS Trend model. In 2013, the Naive Trend was heavily influenced by the 2012 yield losses. The WS Trend model also forecasted lower yields for some grids in Missouri because of the influence of trend (*Figure D.3*). The WS No Trend model matched the yields from Deines et al. (2021) better in those grids. Overall, the low resolution trend was more useful to WS Trend models in Europe (NUTS2 to NUTS3 regions) than in the US (county to 10-km grids). The Pearson's r for NUTS2 trend and NUTS3 yields was 0.81 for soft wheat and 0.61 for potatoes. The corresponding value for county trend and 10-km grid yields was 0.33. This makes sense because NUTS3 regions are much larger than 10-km grids and yield trend is more pronounced at larger spatial levels, where variability due to other factors tends to average out.

5.4 Discussion

Standard machine learning methods rely on supervision labels to forecast crop yields and this reliance can be a limitation when yield data is unavailable. Weakly supervised methods using deep learning address this limitation by learning from high resolution inputs and low resolution labels. In Europe, weakly supervised models were statistically better than the Naive Trend models but similar to Naive GBDT and LSTM models. Because of this similarity with naive models, we could not say that weakly supervised models were better. Even then, forecasts from weak supervision were useful for two reasons. First, the WS Trend model was also statistically similar to HR GBDT and HR LSTM models, which were themselves similar to the Naive GBDT and Naive LSTM models. Therefore, WS Trend model forecasts were as good as those from the strongly supervised HR models. Second, the WS models, especially the WS No Trend model, captured some NUTS3-level yield variability (*Figure 5.5*). Combining information from the two WS models could provide more accurate estimates of yields as well as spatial differences among NUTS3 regions.

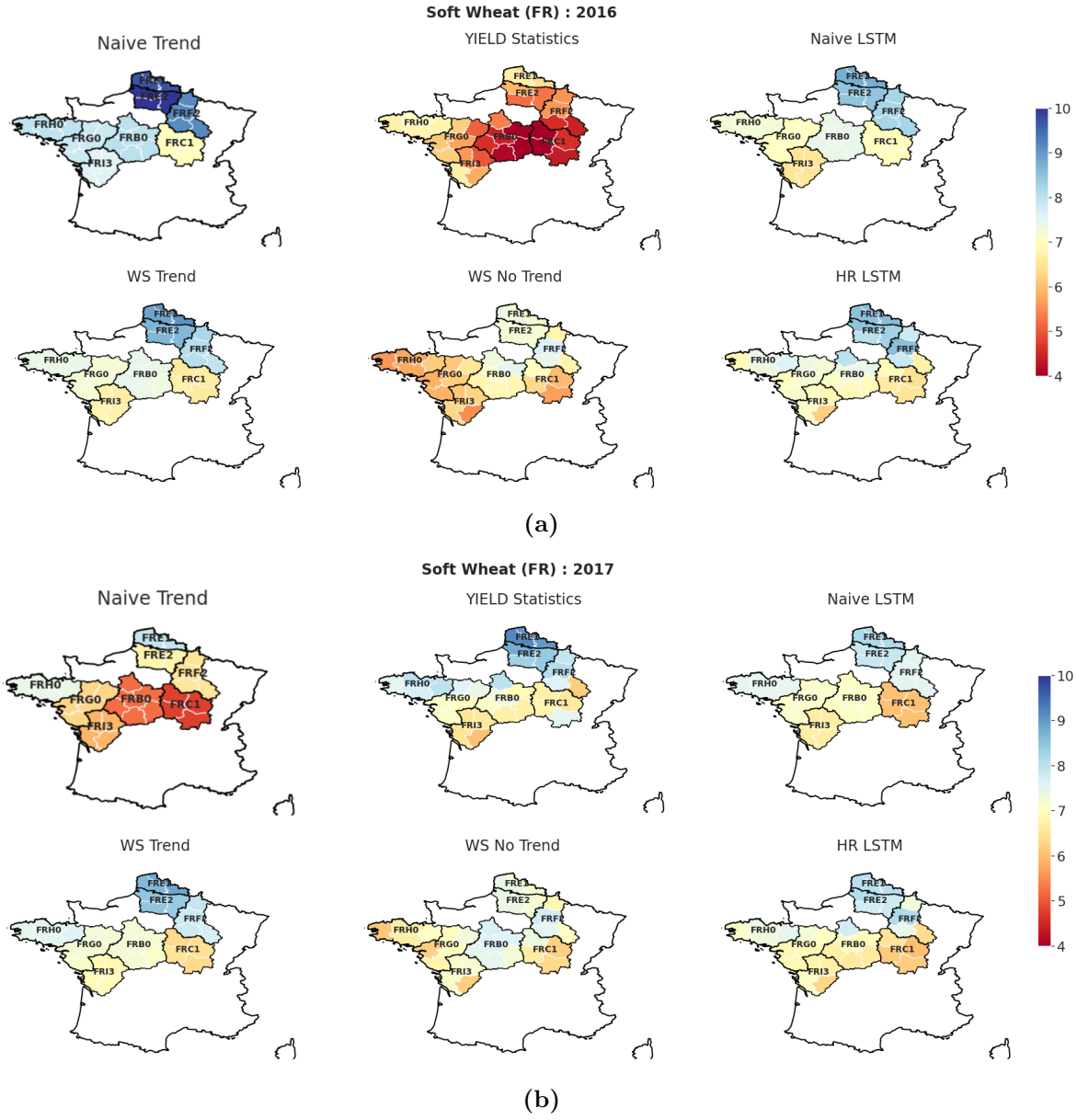


Figure 5.5: Spatial variability of soft wheat yields and forecasts. (a): FR 2016. (b): FR 2017.

Our weakly supervised framework was adapted to produce corn yield forecasts for 10-km grids in the US with minimal changes. Changes involved preprocessing data from Copernicus Data Store and NASS, instead of MCYFS, and handling a larger data size (approximately ten times bigger). The WS Trend models were significantly better than strongly supervised models at both county level and 10-km grid level. Both WS models (Trend and No Trend) also captured some grid-level variability within counties (*Figure D.2, D.3*). Furthermore, the WS No Trend model captured some of the yield losses due to drought in 2012. County-level NRMSEs were quite similar to those reported by other studies that used complex architectures. For example, Khaki et al. (2020) reported an NRMSE of 9% for corn for test years 2016, 2017, 2018. The WS model had a corresponding NRMSE of 10.54%. Future work could experiment with architectures that combine CNN and RNNs (e.g. convolutional LSTM), another crop model (e.g. WOFOST or APSIM (Holzworth et al., 2016)), and additional information on

farm management to improve the performance of weakly supervised models.

We have shown that weak supervision is capable of disaggregating crop yields from low to high resolution. Similarly, the WS models produce reliable forecasts for low resolution as well. Our analysis of spatial yield variability at high resolution relied on Kendall's tau values and qualitative comparisons for selected years. Results showed that the WS No Trend models were able to capture some spatial differences among NUTS3 regions or 10-km grids. In Europe, we noted that information from the WS Trend model and WS No Trend model could be combined to make them useful. This paper did not delve into how to combine yield forecast information from the two models. In general, the WS Trend model forecasts indicate where the yield level should be. The WS No Trend model provides information about deviations from that yield level. In the US, the WS No Trend model performed better than the WS Trend model in both 2012 and 2013. Future work could address stakeholder needs by developing a consistent method of selecting one WS model or combining information from the two WS models.

Performance of weakly supervised models was quite different between Europe and the US. In the US, WS Trend models were better than even the strongly supervised HR models. In Europe, they were similar to Naive GBDT and Naive LSTM models as well as HR GBDT and HR LSTM models. To understand this similarity among HR models, WS Trend model and Naive models, we looked at the effect of yield trend. In Europe, the Naive Trend models, based on the low resolution trend, were also statistically similar to the HR Trend models for both soft wheat and potatoes. Therefore, similarity between low resolution and high resolution trends, coupled with high correlation between low resolution trend and high resolution yields (Pearson's r: 0.81 for soft wheat and 0.61 for potatoes), had a significant influence in making the models similar to each other. In the US, Naive Trend and HR Trend were statistically different. At the same time, the low resolution trend was correlated less with high resolution yields (Pearson's r: 0.33). The weaker influence of low resolution trend may have helped the WS models to learn better from high resolution inputs. Consequently, the WS Trend model was statistically better than other models at both resolutions. Apart from the influence of yield trend, factors that could have affected weak supervision include differences in spatial resolution, the number of sub-regions or grids within a region or county, the quality of crop area weights produced, quality of low resolution labels and yield variability at high resolution. Research into how each of these factors affect performance of weakly supervised models will bring clarity to when and where weak supervision will work.

In this chapter, we have scratched the surface of high resolution crop yield forecasting with weak supervision. We see three areas that need further research to gauge the benefits and limitations of weakly supervised methods. First, more work is needed to understand the scale differences that can be handled by weak supervision. For example, weak supervision worked well between counties to 10-km grids in the US and less so between NUTS2 and NUTS3 regions in Europe. For very large differences in resolution (e.g. NUTS3 regions or counties to 1km grids), supervision signals from low resolution labels may be insufficient to capture high resolution differences. Third, predictor inputs must be suitable to capture yield variability at selected resolutions. Crop simulation outputs and weather variables may correlate well with yields at NUTS3 or 10-km grids, but become less relevant at farm or parcel

level. High resolution remote sensing data, for example from Sentinel satellites (Copernicus ESA, 2022), and ground measurements may provide better predictors for farm level yields. Fourth, we experimented with standard neural network architectures. Future work could investigate other architectures that are more suitable for weak supervision. As mentioned above, architectures that combine strengths of CNNs and RNNs to learn both spatial and temporal features are also worth exploring. Data size and quality will always play a role due to the data-driven nature of neural networks.

Crop yield predictors will become available at increasingly high resolution. Yield data may be missing due to many reasons, including privacy concerns. When there is an imbalance between spatial resolutions of inputs and yields, weak supervised methods provide a solution. Our approach will continue to work when high resolution yield data becomes available for some regions but not others. Deep learning may also provide a way to better optimize the crop area weights. High resolution crop areas, when available, will remove the need to estimate them and further improve the quality of yield forecasts. High resolution crop yield forecasts provide useful information to policymakers and other stakeholders for local analysis and monitoring. We have shown that weakly supervised methods can produce such forecasts in the absence of high resolution labels.

5.5 Conclusions

We designed a weakly supervised framework to train deep learning models that can learn from high resolution inputs and low resolution labels to produce crop yield forecasts for both resolutions. Evidence from NUTS2 to NUTS3 disaggregation in Europe and county to 10-km grids disaggregation in the US showed that weak supervision produces quite reliable yield forecasts in different settings, both in terms of agro-climatic factors and spatial resolutions. Forecasts from weakly supervised models not using the low resolution yield trend captured a significant amount of high resolution yield variability and produced more accurate forecasts for extreme harvests. The framework can be improved with additional data sources, including high resolution crop areas, a better understanding of factors affecting the performance of weak supervision and neural network architectures that can capture both spatial and temporal differences. Overall, high resolution crop yield forecasts are useful to farmers, policymakers and other stakeholders as they provide more detailed information about local yield variability. Weakly supervised methods provide a way to produce such forecasts when high resolution yield data is unavailable.

Chapter 6

Synthesis

6.1 Review of research objectives

The main objective of this thesis was to investigate the benefits and challenges of using machine learning for large-scale crop yield forecasting. This objective was divided into four sub-objectives:

1. **Design a generic explainable, modular and reusable machine learning workflow to forecast crop yields for multiple crops and countries.** The workflow, called a machine learning baseline (*Chapter 2*), used features designed by experts based on agronomic principles of crop growth and development. Workflow components were kept modular to enable updates or improvements over time. Configuration options, such as crop and country, allowed us to reuse the workflow for many case studies.

2. **Evaluate the benefits and limitations of machine learning to produce regional crop yield forecasts in Europe.** In *Chapter 3*, the machine learning baseline from *Chapter 2* was improved in many ways and evaluated on six crops and nine major crop-growing countries of Europe at both regional level and national level. Similarly, regional forecasts were analyzed to demonstrate how well they captured the spatial yield variability for an average harvest and two extreme harvests.

3. **Assess the performance and interpretability of deep learning models for crop yield forecasting.** In *Chapter 4*, expert-designed features were replaced by features automatically learned by neural networks, which were then evaluated for forecasting performance and interpretability. For interpretability, deep learning model forecasts were analyzed to extract feature importance scores. Human experts then assessed the relative importance of features and their positive or negative influence on yield.

4. **Design a framework to produce high resolution crop yield forecasts when high resolution yield data are unavailable.** The weakly supervised deep learning framework, described in *Chapter 5*, produces high resolution forecasts using high resolution inputs and low resolution labels. Weakly supervised models were evaluated on their ability to disaggregate crop yields from low to high resolution and to produce low resolution yield forecasts using high resolution inputs. The spatial variability captured by high resolution yield forecasts was analyzed for an extreme harvest and the following season's harvest.

The rest of the chapter is structured as follows: *Section 6.2* recapitulates the main findings; *Section 6.3* reflects on our findings, contributions and limitations; *Section 6.4* discusses future outlook in terms of challenges, opportunities and directions for future research; and *Section 6.5* summarizes our conclusions.

6.2 Main findings

Forecasts from machine learning (*Chapter 2* and *3*) and deep learning models (*Chapter 4* and *5*) shed light on the benefits and challenges of using machine learning and deep learning for large-scale crop yield forecasting. We found that machine learning can combine domain knowledge with data-driven learning and provide several benefits:

- Our data sizes (200 to 2000 labeled instances) were relatively small (see also Meroni et al. (2021)) compared to studies in the US (Shahhosseini et al., 2021; Khaki et al., 2020), which have used up to 10000 labeled instances. Even then, forecasts from machine learning methods were comparable in performance with forecasts from an operational large-scale yield forecasting system.
- Regional or sub-national differences average out at national level and national forecasts do not reveal about those differences. Our results showed that machine learning can complement the expert-driven approaches, similar to MCYFS, to operationalize regional crop yield forecasting at large scale.
- Deep learning methods automatically learned features that performed similar to expert-designed ones.
- Deep learning model forecasts were based on plausible relationships between features and crop yield. Feature influence matched expert knowledge particularly well for cases with high forecasting accuracy (and low errors). Therefore, accuracy and interpretability of models were closely related.
- Deep learning combined with weak supervision produced reliable high resolution forecasts in the absence of high resolution labels for two different settings (Europe and the United States), both in terms of agro-environmental factors and spatial resolution.
- Machine learning and deep learning provided increased automation required to scale crop yield forecasting to many case studies and higher resolution.

Our results also identified some challenges to further realize the benefits of machine learning. Some of these challenges are discussed in greater detail in *Section 6.4*.

- **Data size and quality**: Due to small data sizes, machine learning models sometimes overfitted the training data and did not generalize well to the test data. In some cases (e.g. Romania), a larger data size did not produce better results, raising data quality concerns.
- **Forecasting setup**: Forecasting setup was selected to enable performance comparison with MCYFS and to forecast yields for potentially heterogeneous regions at sub-national level. Although the setup included features to identify extreme conditions, yield forecasts were mostly influenced by the yield trend, and less by seasonal features. As a result, yield extremes were not captured well, and forecasting errors did not always improve as the season progressed.
- **Customized tools and benchmarks**: Standard machine learning and deep learning libraries are not designed with agricultural data in mind. Customizations are required to capture spatial and temporal relationships in agricultural data. Our machine learning workflows served as custom tools and produced baseline or benchmark results using MCYFS data. Due to the absence of public benchmarks, it was challenging to compare results with other studies using machine learning to forecast yields.
- **Interpretability**: Human experts were able to understand feature importance scores due to their familiarity with the data and the principles of crop growth and development.

Even then, they did not always agree with each other when selecting important predictors or scoring interpretability. Considerable effort is required to make yield forecasting models understandable to stakeholders.

6.3 Reflection

In *Section 1.4.3*, we motivated the use of machine learning for large-scale crop yield forecasting with two main observations. First, operational large scale systems (e.g. MCYFS) do not use machine learning despite its benefits, such as modeling complex relationships and increased automation. Such systems preferred simple and interpretable statistical models built using crop model outputs, remote sensing indicators and survey results. Second, previous studies on crop yield forecasting with machine learning had not sufficiently addressed challenges of large scale application. They usually focused on specific crops and locations (e.g. Pantazi et al. (2016); Cai et al. (2019)). Therefore, we set out to build machine learning and deep learning solutions that would work at large scale.

6.3.1 Scientific contributions

This thesis serves as a roadmap for large-scale crop yield forecasting by identifying key requirements (the "what") and designing and implementing workflows to address them (the "how"). Some of these requirements were noted in *Section 1.4.2*. In summary, automated workflows were needed to provide a consistent and reproducible method of forecasting yields across multiple spatial levels (e.g. grids, regions, countries). The workflows had to account for data sparsity and produce crop yield forecasts backed by explanations understandable to human stakeholders. The four sub-objectives addressed these requirements and, in the process, made scientific contributions related to defining the forecasting setup, designing workflows for large scale, handling missing data and evaluating interpretability of model forecasts.

The machine learning workflows from *Chapter 2* and *3* defined the forecasting setup in three ways. First, they combined agronomic principles of crop modeling with data-driven learning, an approach called knowledge-guided machine learning (Willard et al., 2022; Liu et al., 2022). Features were designed using expert knowledge from data that included crop model outputs, and feature selection was used to identify the most predictive features. Li et al. (2021) have followed a similar approach for crop yield forecasting in China. Second, the workflows defined a scheme to split data based on time, with validation and test years coming after the training years. This custom split includes a 5-fold sliding validation to limit overfitting and to prevent information leakage from the test set. When switching to deep learning, the setup was kept similar to enable comparison with standard machine learning algorithms. Some studies have used random splits (Cai et al., 2019; Lischeid et al., 2022), which are useful to predict yields of missing years but not for forecasting using historical data. Because random splits do not impose a time constraint on training and test sets, information flow is less constrained and performance of models can be inflated. Third, machine learning forecasts were produced at the highest resolution possible and aggregated based on crop area weights to produce forecasts for lower resolutions including national level. This approach provides a consistent

method for forecasting crop yields at multiple spatial levels.

In addition to the explainable feature design and time-based data splits, the machine learning workflows included customizations to address requirements of large scale systems and to test alternative setups. We ran experiments with the baseline and the improved workflow to shed light on the added value of these customizations, such as predicting residuals (yields - yield trend) and grouping regions according to agro-environmental zones (AEZ). Modeling residuals is motivated by the subdivision of yield variability into mean yield, multi-annual yield trend and residual variation (see Dagnelie et al. (1983)), with seasonal features expected to capture the residual (year-to-year) variation from the trend. Modeling yield residuals did not significantly improve the accuracy metrics. Predicting residuals may make sense for per-region and per-year models. Machine learning is less suitable in such scenarios because the length of yield time series is small. Per-AEZ models also did not significantly improve performance, mainly because a few AEZs had very small data sizes. These results show the limitations of both agricultural data as well as machine learning models. In any case, our explorations and outcomes are relevant to other researchers working with MCYFS data; they show that certain setups do not provide significant performance gains. Similarly, results from the baseline and improved workflow serve as benchmarks for regional and national level yield forecasts in Europe.

The workflows emphasized automation, reproducibility, modularity and reusability to make them useful for large scale applications. Increased automation helped us iterate quickly and improve the workflows, and made model outputs reproducible, barring some stochasticity related to random initializations. Human experts can use saved outputs for post-hoc analysis of forecasting performance and interpretability. Machine learning and deep learning workflows were kept modular to support updates and improvements over time. Indicators selected to design features can be updated based on new knowledge or context. Similarly, researchers can experiment with new machine learning algorithms or neural network architectures. The emphasis on reusability shifted focus from point solutions to specific case studies to generic and scalable workflows for many crops and locations. With MCYFS data, the workflows can be extended to other crops and countries in Europe. Applications to other continents are possible when equivalent crop productivity indicators (e.g. dry-weight yield biomass, leaf area, development stage) are available. In *Chapter 5*, the improved machine learning workflow was adapted to build a corn yield forecasting model in the US, proving its usefulness outside of Europe.

In *Chapter 5*, we also designed a weakly supervised deep learning framework to tackle the shortage of label data (yields and crop areas) at high resolution. Strongly supervised models can be built only at the administrative levels where yield statistics are published (NUTS2 or NUTS3 in Europe and counties in the US). We showed that weak supervision can produce yield forecasts at higher resolutions than these. Weakly supervised methods have been previously used to detect or count objects (Wang et al., 2022; Chandra et al., 2020; Ghosal et al., 2019) and to estimate density (Jacobs et al., 2018), but not – as far as we know – to forecast crop yields. The weakly supervised framework is useful to researchers working on similar problems, where labels may be missing for various reasons. Weakly supervised models can work as intermediate solutions in places where high resolution labels can be collected

over time. Strongly supervised models can be built when the number of labeled instances becomes large.

Reliability of machine learning and deep learning models depends on how well their forecasts can be understood by human stakeholders. The machine learning workflows took an organic approach to interpretability. Feature design incorporated domain knowledge from agronomy, crop modeling and yield forecasting experts. Therefore, when provided with feature importance scores, interpreting them was relatively straightforward. With deep learning models (*Chapter 4*), understanding what neural networks had learned and how yield forecasts were made was challenging. Many studies look at feature importance or activation maps for interpretability (Wolanin et al., 2020; Khaki et al., 2020; Mateo-Sanchis et al., 2021), and some look at interactions strengths or effect size as well (Nayak et al., 2022; Inglis et al., 2022). However, they do not validate whether their analysis is understandable to human stakeholders. We engaged human experts and highlighted the role of human stakeholders in assessing interpretability. Quantitative comparisons between feature importance scores from deep learning models and expert-provided scores, using metrics such as Wasserstein distance, were still difficult to interpret. Therefore, feature importance scores were plotted to show the positive or negative influence of features on yield forecasts. Experts provided feedback about whether the relative importance of features and their impact on yield matched their knowledge and experience. Overall, we developed an approach to assess model interpretability with feedback from human experts that went beyond feature importance plots, which are often not understandable on their own.

6.3.2 Societal relevance

Timely, consistent and publicly available crop yield forecasts provide unbiased information to stakeholders (Jiang et al., 2020), including farmers, commodity traders, logistics companies and governments. In the US, the US Department of Agriculture (USDA) publishes crop supply and demand estimates for the country and the world (USDA-NASS, 2012). In Europe, JRC forecasts are shared by the Directorate General for Agriculture and Rural Development (DG-AGRI) of the European Commission with policymakers (in the European Parliament and Member States), stakeholders monitoring crop production and trade and the Agricultural Market Information System (AMIS) (van der Velde et al., 2018). Crop yield forecasts influence commodity markets and the decisions of businesses and governments (USDA-NASS, 2012). They also contribute to global market transparency (van der Velde et al., 2018).

The USDA and JRC forecasts are published at the country or state level. Their usefulness can be improved when they are supported by consistent and reliable high resolution forecasts. In *Chapter 3*, we underscored the need for timely publication of regional crop yield forecasts in Europe, and showed that machine learning workflows provide a consistent and reproducible method to forecast both sub-national and national yields. In *Chapter 5*, weak supervision was used to produce yield forecasts up to 10-km grid level. Our results have demonstrated the added value of machine learning in terms of forecasting performance, automation, reproducibility and consistency across crops, countries and spatial levels. Similarly, our machine learning workflows provide guidance on how to integrate domain knowledge and operational requirements at large scale. Therefore, operational large-scale systems, such as

MCYFS of JRC and National Agricultural Statistics Service (NASS) of USDA, can leverage our workflows and include machine learning in their toolbox to benefit from growing amounts of data and to address modeling limitations of existing methods. In the meantime, JRC has started to harmonize and publish regional crop yield statistics (Cerrani & López Lozano, 2022; EC-JRC, 2022), and MCYFS analysts are testing regional yield forecasting for winter crops in Ukraine.

Our workflow designs, implementations and sample data are openly accessible (see the *Appendices*) to support the equal access to crop yield forecasting models for all stakeholders. As of January 2023, sample data published in Zenodo has close to 2500 downloads. Our work has also been shared with JRC and used in workshops involving stakeholders from two EU Horizon 2020 projects (CYBELE and Dragon). Furthermore, a narrowed-down version of our workflow has been developed as a tutorial for the MSc students of the machine learning course at Wageningen University, since 2021.

6.3.3 Constraints and limitations

Due to their data-driven nature, performance of machine learning and deep learning models heavily depends on data size and data quality. Except when predicting corn yields in the US (*Chapter 5*), all the work in this thesis used MCYFS and Eurostat data. The choice of MCYFS and Eurostat data was motivated by four reasons. First, the MCYFS database is well maintained; there was no need to collect or preprocess data. Second, working closely with the JRC gave us an opportunity to benefit from their knowledge and experience in crop yield forecasting. Third, regional machine learning forecasts could be aggregated to national level and compared with MCYFS forecasts. Fourth, the machine learning workflows could incorporate requirements of an operational large scale system. On the flipside, MCYFS and Eurostat data have size and quality concerns, which put some constraints on our research:

- In some cases, data size was quite small due to a short time series of yield statistics and the alignment of different data sources. For deep learning, data size was not big enough to train complex architectures that learn both spatial and temporal relationships.
- Regional yield statistics are not always reliable because the collection and curation protocols for these statistics vary across countries (López-Lozano et al., 2015). Machine learning models are trained with these yield statistics as ground-truths. Poor quality yield data resulted in high forecasting errors in some cases (e.g. Romania) even when data size was relatively large.
- Yield statistics are not always reported separately for irrigated and rainfed areas, and the available irrigation masks are mostly static. Therefore, irrigated and rainfed systems were modeled together. Modeling them together led to unclear effects of irrigation and moisture on crop yield (*Chapter 4*).
- Crop areas, used as weights to aggregate inputs and yields, have uncertainties due to the quality of crop masks as well as statistics collection and interpolation methods (Cerrani & López Lozano, 2017). These uncertainties affected the amount of yield variability captured by forecasting models.

- The input data did not capture all factors contributing to yield variability and the influence of predictors may have been spatially and temporally inconsistent (see Lecerf et al. (2019)). Our input data did not include farm management information, such as cultivars and fertilization, and the irrigation data was static. Therefore, machine learning and deep learning models could not capture the effects of these factors on yield variability.

Our setup choices were motivated by existing domain knowledge, requirements of large scale systems, such as MCYFS, and the need to capture spatial and temporal yield variability simultaneously for diverse regions. These choices led to some limitations that affected model performance and interpretability:

- Comparisons between the baseline and optimized workflows showed that our setup has a limited ability to capture the spatial differences among heterogeneous regions. The baseline only used one static feature: soil water holding capacity. In the optimized workflow, other static features, such as elevation, slope and field size, were added. These static features did not significantly improve performance. As a result, forecasting errors remained high for both the baseline and the optimized setup (e.g. Spring Barley in Spain and France).
- Results from both machine learning and deep learning workflows showed that yield forecasts were mostly influenced by the yield trend and less by seasonal features. Similarly, forecasting errors did not improve significantly after a certain point in the season (see also Meroni et al. (2021)). As a result, forecasts agreed with yield statistics for average harvests but less so for extreme harvests. Apart from the seasonal features for extreme conditions, we did not explore ways to model yield extremes.
- In *Chapter 4*, our approach to interpretability assessment relied on experts who were familiar with the data and yield forecasting methods. Our method does not work for other stakeholders, such as farmers and commodity traders, who may not have experts available on demand. We also did not address social and cognitive considerations that affect reproducibility, such as the composition of stakeholders, their prior knowledge and the amount of time required to understand interpretability metrics. Feature importance plots may be hard to interpret, and ways to visualize feature interactions may be required (Molnar, 2022). When sharing visualizations or asking for feedback, stakeholders may interpret some plots or questions in different ways. A framework is needed to systematically assess interpretability by accounting for social and cognitive factors (Miller, 2019).

Overall, we combined strengths of crop models and remote sensing with machine learning, but did not address the limitations of crop models and remote sensing data used. Crop model outputs come from a version of WOFOST that does not account for certain field conditions, such as nutrient limitations and pests (de Wit et al., 2019). Our machine learning models used some crop model inputs (weather and soil), crop model outputs and remote sensing indicators and tried to capture the effects of factors not included in crop models. This approach worked to a certain extent, as indicated by the performance comparison with MCYFS, but also had limitations. Remote sensing data included a generic and crop-agnostic

indicator (i.e. FAPAR) that captures the relationship between biomass production and the photosynthetically active portion of solar radiation absorbed by the canopy (Monteith, 1972; Daughtry et al., 1992). We did not explore data assimilation in crop models (de Wit & van Diepen, 2007), which involves reinitializing or recalibrating crop models or overriding certain state variables based on remote sensing (Venancio et al., 2019; Doraiswamy et al., 2003).

6.4 The road ahead

We can imagine a future where stakeholders get yield forecasts and supporting explanations tailored to their needs. In-situ data collected by researchers, private companies or governments and simulated or synthetic data are shared to complement each other. Machine learning and deep learning models could then be trained to forecast crop yields. Models pre-trained on data-rich regions, crop model simulations or synthetic data could provide a starting point for forecasting yields in data-sparse areas. Yield forecasts could be produced in the highest resolution possible (e.g. farms or small grids) and aggregated to low resolutions, ensuring a consistent method of producing forecasts across scales, from local to regional, national and global levels. Customized tools and benchmarks would be used to compare setups and model forecasts and improve them over time. Model building, evaluation and finetuning would be automated, except to incorporate stakeholder input. In addition to forecasting yields, machine learning would contribute to pest and disease detection (Behmann et al., 2015) and crop protection (Ip et al., 2018). Overall, machine learning would become an enabler for a sustainable and more productive agriculture (Benos et al., 2021).

To get to the future just described, efforts to forecast crop yields at large scale have to deal with diverse conditions and requirements related to crops, regions of the world, spatial scales and stakeholders. Some requirements will vary by spatial scale and others will have to be consistent across scales. Our work provides a roadmap, by identifying key requirements and ways to address them (*Section 6.3.1*), but also has some limitations (*Section 6.3.3*). We emphasized the need for a consistent yield forecasting method across spatial levels. Our approach covers national and regional scale in Europe and counties and 10-km grids in the US. Deines et al. (2021) have proposed an approach that goes from pixels and fields to county level in the US. Some studies have applied machine learning to field-level crop yield forecasting (Cao et al., 2021; Feng et al., 2020). At farm or field level, large scale applications are challenging because data mostly comes from research fields (Kang & Özdoğan, 2019) or private companies (Kayad et al., 2019; Zhao et al., 2020; Deines et al., 2021) and coverage is limited to recent years (Deines et al., 2021). Large-scale collection and sharing of such data has to deal with financial and privacy considerations. Even then, some studies have used machine learning for field-level crop yield forecasting (Cao et al., 2021; Feng et al., 2020). Farm or field-level yield forecasts would complete the coverage from farm to global level and increase the participation of farmers in the data value chain (Elavarasan et al., 2018). Efforts are needed to bridge the gap between regional and farm-level yield forecasting so that crop yield forecasts at low resolution are supported by those from high resolution.

Gaps and challenges in large-scale crop yield forecasting have been identified by recent re-

views of agricultural systems modeling (Jones et al., 2017), local to regional yield forecasting approaches (Schauberger et al., 2020) and global agricultural monitoring systems (Fritz et al., 2019). Based on their findings and our experience, challenges mainly fall into four categories: data availability and quality, forecasting setup, tools and benchmarks, interpretability and stakeholder requirements. On the flipside, growing interest in machine learning, the convergence of knowledge-based methods, remote sensing and machine learning, and emergence of explainable AI tools provide opportunities to address many of these challenges. The rest of this section delves into the four categories of challenges facing large-scale crop yield forecasting. For each category, we discuss the main challenges, the trends and opportunities to address those challenges and specific recommendations for future work.

6.4.1 Data availability and quality

Challenges: Data may be unavailable in many places and, when available, data size and quality may not meet modeling requirements. Data shortage is especially severe in developing countries (Chitsiko et al., 2022; Jones et al., 2017; Schauberger et al., 2020) and for underrepresented crops (Schauberger et al., 2020). Gaps mainly exist in cropland and crop type maps (or crop masks), crop calendars and yield statistics (Fritz et al., 2019). Crop masks, and crop area weights derived from them, play an important role in aggregation of input data and yield forecasts. Most global and regional crop masks are static (Schauberger et al., 2020). Crop calendars, which determine the key phenological phases and the impact of stress factors, are useful for feature design and to interpret the influence of features on yield. Available crop calendars are too coarse in both space and time for locally adapted yield forecasting efforts (Schauberger et al., 2020). The availability of sub-national yield statistics is another challenge because machine learning models are trained with these statistics as labels. Large scale collection of high quality ground data requires a statistically sound sampling method and quality control in a timely manner (Defourny et al., 2019). The size and quality of data restrict the capabilities of models to account for factors that influence crop yields and the ability of researchers to evaluate models across a wide variety of conditions, which in turn limit human understanding and reliability of model predictions (Jones et al., 2017).

Opportunities: Three trends provide promising ways to address data availability challenges. First, the amount of data collected and published is growing (Lokers et al., 2016), and so is the interest in machine learning (Van Klompenburg et al., 2020; Benos et al., 2021). Initiatives to share data have gained momentum (e.g. the Global Open Data for Agriculture and Nutrition (GODAN), http://www.godan.info/, the CGIAR Big Data platform (https://bigdata.cgiar.org/)), and codes of conduct have been developed for sharing agricultural data (van der Burg et al., 2021). Building trust relationships to share farm level data (within limits of codes of conduct) would provide opportunities to train and validate machine learning models at the highest resolution possible. Second, high resolution remote sensing data provides vegetation indicators as predictors of crop yield (e.g. Johnson et al. (2016)) and can be combined with ground data to improve the quality of crop masks and crop calendars. For example, the Sen2Agri system of the European Space Agency (Defourny et al., 2019) provides a toolbox to produce crop type maps, and the EUROCROPS project is building a harmonized dataset of crop type information (Schneider et al., 2021) to validate such maps. Similarly, the WorldCereal project (Franch et al., 2022) has produced global

wheat and maize crop calendar maps at 0.5° resolution. In the future, coverage could be extended to other crops and accuracy could be improved by leveraging crowdsourcing and self-reporting by farmers (Fritz et al., 2019). Third, synthetic or modeled data provide alternative sources of input data. Examples include synthetic reanalysis weather datasets, such as ERA5 (Hersbach et al., 2020), and AgERA5 (Boogaard et al., 2022), and soil data, such as SoilGrids250 (Hengl et al., 2017). These weather and soil data can be used to run crop models at regional or grid level with additional information about crop cultivar, crop calendar and management practices. For example, de Wit et al. (2022) have combined earth observation data, renalyzed weather data and a simple crop model to produce crop productivity indicators and evapotranspiration indicators at 0.1° grids for dominant cropping patterns around the globe. Their work can be extended by adding a water balance model, especially taking into account the effect of irrigation. When crop models cannot be calibrated extensively, validation with some ground observations may indicate whether they improve model performance (see Liu et al. (2022)). Accurate calibration may not be necessary for machine learning or deep learning because crop model outputs are usually used as features in conjunction with other data or for pre-training neural networks (see Han et al. (2023)).

Recommendations and future work: Organizations responsible for collecting and sharing agricultural data, such as the Food and Agriculture Organization, Eurostat and national statistics offices, could review and standardize collection procedures across countries to produce consistent yield and crop area statistics across spatial levels. Review and standardization of protocols will help in the long run, but do not provide immediate solutions. In the meantime, data sharing and synthetic or modeled data can fill the gaps. Farm-level data, when available, could be shared by aggregating to small grids (see Deines et al. (2021)) where synthetic or modeled data have been published. Sharing aggregated rather than original farm yield data would also preserve privacy. When yield data is large, machine learning or deep learning models can be trained with the data; when yield data is sparse, certain forecasting setups can help to learn with low resolution labels or a small number of labels (see *Section 6.4.2*). Grid-level yield forecasts can be aggregated to regional (county, province, state) and national level based on crop area weights and validated there using published yield statistics. Crop production areas can come from sources such as the national cropland data layer in the US (Boryan et al., 2011) and farmers' declarations to the European subsidy control. On the input side, some studies have tested the value of synthetic weather data for crop model simulations (Qian et al., 2011; van Wart et al., 2015). Future work could evaluate whether synthetic and modeled data meet the requirements for crop modeling and yield forecasting (see Grassini et al. (2015)). Using synthetic or reanalyzed data and available in situ data, future research could also target data-sparse areas and underrepresented crops. Research in underrepresented crops is relevant to climate change adaptations and nutrient diversity (Schauberger et al., 2020).

Where data and statistics are available, their quality must be monitored in a timely manner. In Europe, JRC has developed a method to harmonize crop area statistics (Cerrani & López Lozano, 2017, 2022). The same method is being used to harmonize regional production (and yield) statistics. Data quality checks also involve detection and imputation of outliers. In *Chapter 3*, we explored ways to detect outliers and cleaned duplicate yield values. Methods based on clustering, measures of similarity and data distribution can be used to determine

possible outliers (Boukerche et al., 2020; Smiti, 2020). Once a value is determined to be an outlier, removing the value must be followed by data imputation. More research is needed to evaluate the impact of different outliers detection, validation and imputation strategies on yield forecasting.

6.4.2 Forecasting setup

Challenges: The main challenge in forecasting setup is the expertise needed to make setup choices, which must consider ways to incorporate domain knowledge and to address operational requirements as well as assumptions made by machine learning algorithms. These choices can be of at least four types. First, learning algorithms and architectures must be selected to handle small data sizes and sparse or missing labels. Second, spatial structure of data must be captured or data for one model must be selected based on agro-environmental similarities. Machine learning methods may struggle to find generalizable relationships when data comes from heterogeneous regions. Third, features must be designed or extracted to capture their influence on yield. Features need to account for the effect of the yield trend and extreme conditions that can lead to yield losses or surpluses. Fourth, combining strengths of common methods, such as crop models, remote sensing, statistical analysis and machine learning, may provide added benefits.

Opportunities: Weak supervision (*Chapter 5*) makes it possible to learn from low resolution labels when high resolution labels are missing. When input data is available and labels are missing for certain years, semi-supervised or self-supervised approaches can be used. Clustering-based approaches have been used to select data for a model from similar spatial units (e.g. Johnson et al. (2016)). Clustering may not be necessary when architectures can capture the spatial and temporal structure of data. Combining CNNs, which handle spatial information, and RNNs, which handle sequential information, has shown promising results in the US (You et al., 2017; Gavahi et al., 2021). The convergence of crop modeling, remote sensing and machine learning (Jeong et al., 2022; Lawes et al., 2022) provides opportunities to integrate detailed crop information from remote sensing into crop models (through data assimilation) and domain knowledge in the form of crop calendar and productivity indicators into machine learning. Domain knowledge can also come from experts, who can contribute to feature design and selection or design of architectures that capture spatial and temporal relationships. The trend of knowledge-guided machine learning (Willard et al., 2022; Liu et al., 2022) emphasizes designing architectures according to relationships among input data, adding biophysical constraints to the loss functions and using crop model outputs for knowledge-guided initialization of neural networks (see Han et al. (2023)).

Recommendations and future work: Our work focused almost entirely on supervised learning. Semi-supervised or self-supervised learning can be used to increase data size or learn with limited data. In MCYFS, input data, except FAPAR, is available from 1979 onwards. Semi-supervised learning can exclude FAPAR (and yield trend) and train machine learning models to hindcast for missing years. The predicted labels can then serve as pseudo labels for yield forecasting. In self-supervised learning, autoencoder networks can be trained to summarize seasonal time series data into features and to decode the original time series. Such self-supervision can be used to pre-train a large neural network (see Pylianidis & Athanasiadis

(2022)). Similar pre-training is possible with crop model simulated yields.

We investigated three ways of grouping the data spatially. In *Chapter 3*, per country and per AEZ models were evaluated. In *Chapter 5*, data from multiple countries were combined. Future work could investigate ways to select groups of regions that have consistent relationships between predictors and yield or design architectures to learn the spatial differences. How the data is split into training and test sets also has a significant bearing on performance. For our experiments, temporal splits were appropriate. In certain settings, spatial splits may be necessary. The impact of the spatial extent of the data on prediction performance needs to be further explored Filippi et al. (2019).

We deferred case-study specific setup choices to experts with knowledge of local conditions. It would be interesting to see how the models perform when optimizations are added based on knowledge of experts, e.g. at MCYFS. In *Chapter 2* and *3*, features were designed based on expert-defined tables of WOFOST outputs, weather variables and remote sensing indicators. The tables could be updated based on local knowledge or availability of additional data sources. Future research could also target data sources that account for additional factors that influence yield, such as nutrient limitations, pests, diseases and farm management decisions.

Finally, we did not focus on extreme yield losses or surpluses. Machine learning and deep learning algorithms, when trained with mean squared error or similar errors, fail to account for the extreme cases. Therefore, our models did not capture the effects of extreme weather very well. Similarly, most machine learning forecasting setups (except see Crane-Droesch (2018)), including ours, do not use weather forecasts for future timesteps and assume that no major unexpected yield-changing events occur after the prediction date (Schauberger et al., 2020). Some previous studies have used weather forecasts to improve crop model simulations (Cantelaube & Terres, 2005; Quiring & Legates, 2008) and statistical regression models (Ansarifar et al., 2021). Extreme events like floods or pest outbreaks can seriously change expectable harvest amounts (Schauberger et al., 2020). Future research could investigate how to adjust the setup to forecast yields for extreme harvests, especially in conjunction with forecasted weather.

6.4.3 Customized tools and benchmarks

Challenges: Machine learning and deep learning tools are usually designed for common use cases, such as computer vision and natural language processing. Customized tools are needed to address the requirements of agricultural applications (Osinga et al., 2022), including crop yield forecasting. Customizations have to address the data size considerations and setup choices mentioned earlier. In the absence of customized tools, individual researchers have to implement common setups and practices in yield forecasting from scratch, and such implementations will have concerns related to errors and reproducibility. Similarly, benchmarks are needed to compare model performance at large scale and to provide guidance on ways to improve model performance (van der Velde & Nisini, 2019). Our machine learning workflows can serve as benchmarks using MCYFS data, but there is a clear shortage of public datasets and performance benchmarks covering different spatial scales.

Opportunities: Many research publications (e.g. You et al. (2017); Khaki et al. (2020)), including ours, have shared software implementations. These implementations can be a starting point for building customized tools. Benchmarks, including datasets, pre-trained models and performance metrics, would benefit large-scale crop yield forecasting in at least five ways. First, certain predictor-yield relationships may only hold for a certain area or period of time (Schauberger et al., 2020). Benchmarks would help validate the generalizability of designs, models and findings from different case studies. Second, machine learning algorithms and neural network architectures could be selected based on performance on public benchmarks. Third, benchmarks would help determine the added value of additional predictors, such as farm economic data. Fourth, the value of synthetic or reanalysis-based data could be determined based on performance comparison with observed data. Data from large-scale crop yield forecasting systems, such as MCYFS and NASS, and globally available weather, soil and crop productivity indicators present opportunities to produce public benchmarks.

Recommendations and future work: Researchers with expertise in agronomy, software development and machine learning need to join hands to build mature, scalable and user-friendly tools customized for crop yield forecasting. For example, the biomedical community has a toolkit or workbench called the Galaxy-ML (Gu et al., 2021; The Galaxy Community, 2022). Our workflows implement customizations related to data-splits, crop calendar, feature design, yield trend estimation and predicting yield residuals. Ideas from our work and other publications (e.g. Li et al. (2021); Meroni et al. (2021)) could be used to build customized tools. Such tools could have options to specify the crop calendar (e.g. sowing, flowering and harvest dates), important predictors for different periods of crop calendar and farm management information to support feature design. Other customizations could address choices related to data preprocessing (data quality analysis, outliers detection and imputation) and forecasting setup (grouping data based on agro-climatic similarities and implementing architectures that learn the spatial and temporal structure of data).

Similarly, concerted effort is required to produce benchmarks at different spatial levels. For example, MCYFS and Eurostat data could be used to produce regional and national benchmarks in Europe. NASS data could be used to produce county, state and country level benchmarks in the US. Available data on weather, soil, farm management, yield and crop production area could be published as data benchmarks. Modeled crop productivity and water balance indicators, similar to WOFOST outputs or those from de Wit et al. (2022), could be included in data benchmarks. In terms of yield forecasting approaches, crop model outputs provide process-based benchmarks. Standard steps followed by MCYFS analysts can be automated to develop expertise-based benchmarks. Our workflows and methods from similar publications can be used to produce machine learning and deep learning benchmarks.

6.4.4 Interpretability and stakeholder requirements

Challenges: Applications of machine learning to crop yield forecasting will have to wrestle with the skepticism about what the models are learning and how they forecast yields. This skepticism is justified by the lack of a framework to systematically assess interpretability. Researchers coming from a crop modeling or statistical modeling background may find

machine learning to be a powerful tool, but lacking interpretability (see Lischeid et al. (2022)). The questions around interpretability are bigger than the skepticism from certain researchers. Interpretability has many aspects that pose their own difficulties. There is no consensus about what interpretability is or how to measure it (Molnar, 2022). Similarly, explaining measures of interpretability to stakeholders is non-trivial (Miller, 2019). Tools are required to visualize complex interactions between predictors and yield in human understandable ways and to share them with stakeholders. Interpretability is challenging because of different expectations or requirements of different stakeholders and the various ways in which prediction models may fit the data equally well (Fisher et al., 2019). Interpretability is also crucial for including machine learning models in policy recommendation or decision support systems, but the path from prediction or explanation to decision is not straightforward. Stakeholders, such as farmers, commodity traders and policymakers, have different requirements related to spatial resolutions, lead times and the decision support they need, and engaging them is critical to understand these requirements. Efforts to unify yield forecasting needs of all stakeholders will only work to a limited extent (Schauberger et al., 2020). Large-scale crop yield forecasting needs to narrow the gaps from model developers to decision support providers and the end users (Challinor, 2009).

Opportunities: The need for interpretability is growing with the widespread use of machine learning and deep learning. To address this need, explorations with machine learning are beginning to merge with knowledge or expertise-based methods. Therefore, the initial skepticism towards data-driven learning might actually help machine learning going forward. The cross-fertilization of ideas from machine learning with knowledge-based methods, such as crop models, will help address the limitations of both approaches. Promising trends include the knowledge-guided machine learning and the effort to design interpretable neural network architectures (e.g. Guo et al. (2019)). The area of interpretable machine learning or explainable AI is growing in other directions as well. Some studies have attempted to provide clarity on interpretability and suggested methods to evaluate it (Molnar, 2022; Hong et al., 2020; Doshi-Velez & Kim, 2017; Narayanan et al., 2018). Methods to extract feature attributions already exist (Montavon et al., 2018; Ancona et al., 2018; Lundberg & Lee, 2017). Research on visualization feature influence is also growing (Gritsenko et al., 2016; Lundberg & Lee, 2017; Olah et al., 2017; Ribeiro et al., 2016). These efforts provide some guidance on how to communicate interpretability metrics or visuals to stakeholders and to include their perspectives in the evaluation of interpretability. Some studies have also looked at methods of estimating uncertainty of forecasts (Wang et al., 2020; Leng & Hall, 2019) and ways of characterizing and addressing stakeholder requirements (Suresh et al., 2021; Challinor, 2009).

Recommendations and future work: In *Chapter 4*, we developed an approach to interpretability assessment using expert feedback. Our approach highlighted the need to involve human stakeholders, but had several limitations discussed in *Section 6.3.3*. Future work could focus on the following five areas. First, yield forecasts need to be supported by estimates of prediction uncertainty to give an idea about accuracy of forecasts. Uncertainty needs to be separated into input uncertainty and model uncertainty (choice of best model or best model parameters). Input uncertainty may come from low quality of input data or unobserved or unregistered data (e.g. actual cropping patterns, spread of pests and

diseases and farm management decisions) and the use of forecasted data in crop models (e.g. weather) (Schauberger et al., 2020). Second, the neural network architectures could be designed based on knowledge-guided approaches to make them more interpretable. Third, the research on feature attribution methods is growing. Future work could address their limitations or explore other ways of measuring interpretability, e.g. counterfactuals (see Schmitt et al. (2022)). Fourth, more work is needed to make interpretability measures friendlier to stakeholders who do not have knowledge of yield forecasting methods. In particular, metrics and visualizations must reveal how various features interact to influence yield. Finally, an interpretability assessment framework is needed to accommodate multiple equally interpretable models (Fisher et al., 2019) and different but equivalent perspectives on interpretability. Human evaluation of interpretability is not very common. Such a framework must also address challenges of social and cognitive aspects of interpretability, which deal with going from predictions to understanding and then to actual decisions. Research on crop yield forecasting has mostly focused on performance of forecasting models and, to a lesser extent, on interpretability of models. In the future, more effort could go into understanding stakeholder needs and addressing them.

6.5 Conclusions

Digital technologies, including artificial intelligence (AI) and machine learning, are expected to transform agriculture and contribute to increased food productivity and sustainability (Barrett & Rose, 2022; Connolly, 2016; Shepherd et al., 2020). The expected transformation will require significant institutional, social and technological changes to share the benefits equitably (Shepherd et al., 2020; Lajoie-O'Malley et al., 2020). Therefore, progress will be gradual and nonlinear (Rose et al., 2022). Even then, machine learning and deep learning methods are providing benefits to agricultural applications, including crop yield forecasting (Benos et al., 2021; Chlingaryan et al., 2018; Kamilaris & Prenafeta-Boldú, 2018). Machine learning and deep learning models improve forecasting performance when data size is large. Based on our own work, machine learning produced reliable results, comparable with an operational yield forecasting system, even with relatively small data sizes. Deep learning enabled automatic extraction of features that performed similar to expert-designed ones and produced forecasts based on plausible relationships between features and crop yield. Deep learning combined with weak supervision also produced high resolution yield forecasts in the absence of high resolution labels. Machine learning and deep learning also provided increased automation required to scale yield forecasting to many case studies and high resolution. Overall, our work showed that machine learning can incorporate domain knowledge and complement expert-driven approaches to provide a more automated, consistent and reproducible approach to crop yield forecasting across multiple spatial levels.

When looking at the bigger picture, gaps and challenges remain in large-scale crop yield forecasting. These challenges are related to data availability and quality, the expertise required to select the appropriate forecasting setup, a shortage of customized tools and benchmarks, concerns about interpretability and insufficient attention to stakeholder requirements. Current trends in crop yield forecasting are generating the right ingredients to address these challenges. Initiatives are needed to create and perfect recipes that connect various data management

and sharing, modeling and decision support efforts together to produce benchmarks and other public goods (Jones et al., 2017), including compatible datasets, customized tools, forecasting models and frameworks to assess performance and interpretability. Significant progress has been made in areas where researchers can work on their own; more effort is required in areas that involve collaborating with experts from another domain or interacting with stakeholders.

Appendices

Appendix A

A.1 Machine Learning Baseline Detailed Workflow

We designed a workflow with two parts. The first part includes steps that were specific to data sources. The second part includes steps that were independent of data sources. The two parts and the steps included were kept modular to allow incremental changes to the workflow.

A.1.1 Data Preprocessing

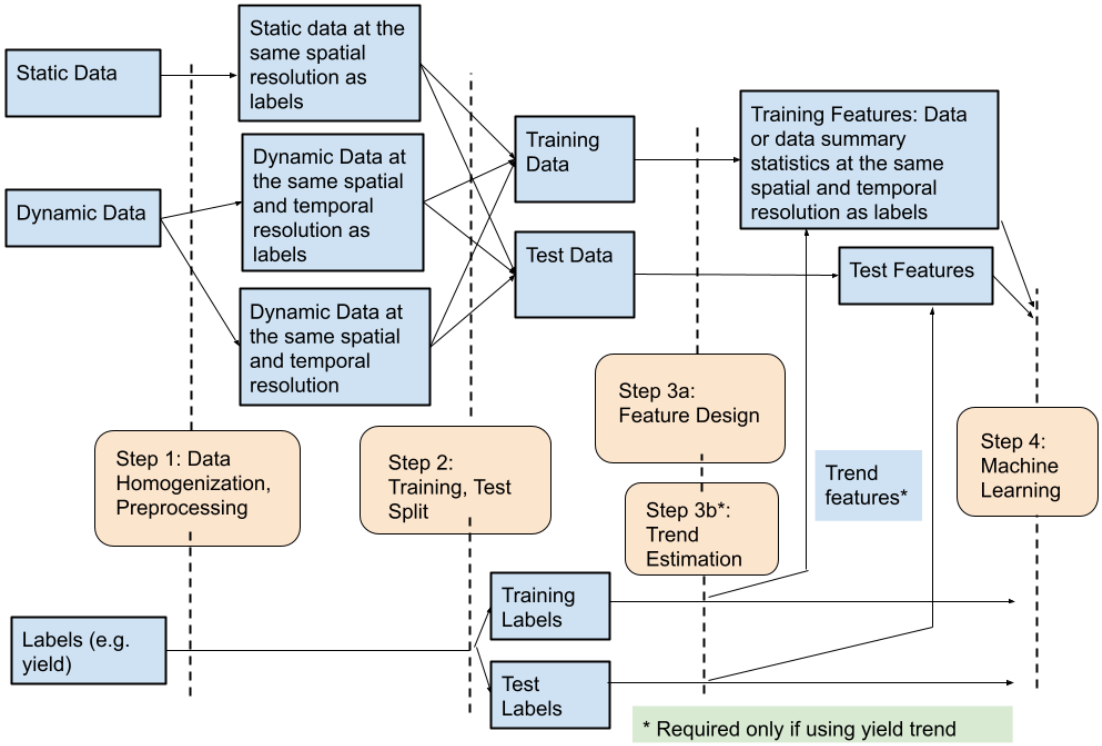


Figure A.1: Preprocessing and feature design. Preprocessing and feature design were used to convert data into features and labels.

During data preprocessing, we homogenized data in terms of file format, filenames and required columns, and aggregated them to the same spatial and temporal resolutions. Input data can come in two forms: (i) static data (values do not change over time), and (ii) dynamic data (values change over time). The labels or targets were yield statistics. Some data could be directly used as features. Others had to be aligned to the same spatial and temporal resolutions and aggregated to create features. In the baseline, the spatial resolution was different NUTS levels (Eurostat, 2016b). Temporal resolution was daily, dekadal, etc. Some crops, such as winter wheat, have growing seasons that cross the calendar year boundary. In order to make the workflow support such crops, we transformed the data to align with the campaign year instead of the calendar year. A campaign year started after the end of the previous growing season and stretched up to the end of the current growing season. Data imputation (i.e. handling missing data) and outliers detection are two other preprocessing steps that could improve the accuracy of crop yield predictions. These steps require a thorough analysis of data sources and were not included in the baseline. The baseline did,

however, filter out data with zero yield values because zero values cause problems, including division by zero.

To keep the steps modular, data could be preprocessed independently using other software, where appropriate (e.g. QGIS for shapefiles, QGIS Development Team (2020)) . The requirement for preprocessed data was that the data had the expected columns (e.g. NUTS region, calendar year) and was at the same spatial and temporal resolution.

Training and test split and feature design are covered in *Section 2.3*.

Yield Trend Estimation

We estimated a linear yield trend using a fixed 5-year trend window. We made the window length configurable to allow other options (e.g. 10 years). For every region and year, we used yield values of 5 previous years as yield trend features for machine learning. In the case of discontinuous data, the method selected the five previous years that were available.

A.1.2 Machine learning

The data sources-independent steps applied machine learning to features and labels (*Figure A.2*). For modularity, features from the first part of the workflow could be saved to a CSV file and loaded later for machine learning.

Feature Scaling

Feature scaling or normalization brings different feature values to similar ranges. The motivation is to prevent the sizes of feature values from affecting the learning process. Min-max normalization scales the features to [0, 1]:

$$x' = \frac{x - min(x)}{max(x) - min(x)}$$

where x is the old value and x' is the new value. Standardization or z-score normalization transforms feature values to have mean 0 and standard deviation 1:

$$x' = \frac{x - mean(x)}{\sigma}$$

where $mean(x)$ and σ are the mean and standard deviation of the feature values before standardization. In the baseline, we use z-score normalization to normalize feature values.

Feature scaling or normalization must be applied after training-test split to avoid information leakage. This is because calculating statistics like min, max, mean and standard deviation must not use the test data.

Feature selection

The motivation for feature selection is to remove features that might be correlated or those that capture the noise in training data, leading to overfitting. In the baseline we used three methods for feature selection: (i) Random Forest (RF), (ii) Recursive Feature Elimination (RFE), and (iii) Mutual Information Regression (MI). RF (Breiman, 2001) uses an ensemble

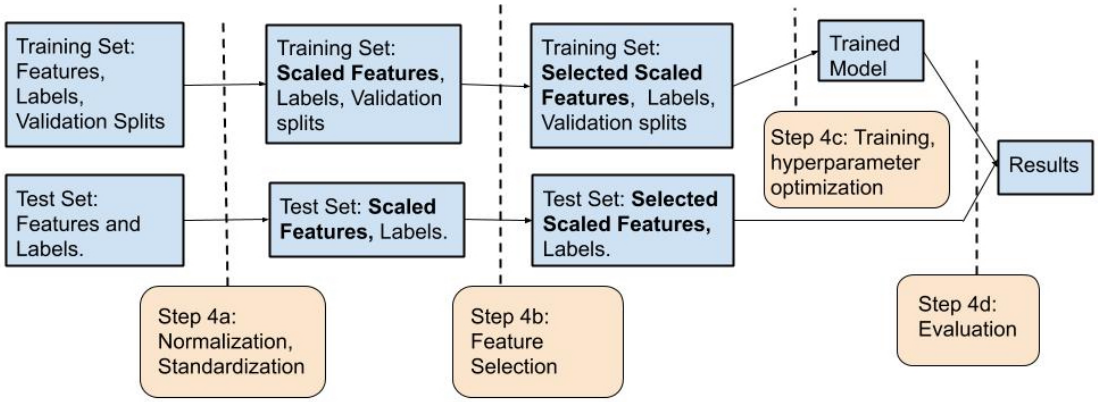


Figure A.2: Machine learning steps. Machine learning was applied using a pipeline consisting of feature scaling, feature selection and training (including hyperparameter optimization). A machine learning model was learned or fitted using only the training set. The fitted model was evaluated on the test set.

of weak decision tree models built by picking a random subset of candidate features to build the trees. The final prediction is an average of the predictions of individual trees. RFE (e.g. Granitto et al. (2006)), recursively eliminates unimportant features by evaluating a machine learning algorithm which provides feature weights or feature importances. In the baseline, we used RFE with Lasso regression (Tibshirani, 1996). Lasso applies the L1-norm penalty on feature weights to drive weights of unimportant features to zero. MI (Shannon, 1948) is a univariate feature selection method, similar to pearson's r (see Benesty et al. (2009)), that calculates the information content of individual features.

During feature selection, we optimized the number of features as a hyperparameter. We included feature selection in a pipeline consisting of feature scaling, feature selection, and training and evaluation. The training and evaluation stage used a standard machine learning algorithm. We passed the pipeline to a grid search method using 5-fold cross-validation or a time-based 5-fold sliding validation to optimize the hyperparameters. For each combination of hyperparameters, a 5-fold cross-validation or sliding validation was run to determine the mean score. Hyperparameter values with the best mean scores were selected as optimal. In the baseline, we focused on optimizing the number of features and used reasonable default values of hyperparameters for training. However, it is possible to do a more thorough hyperparameter optimization if required.

We created a pipeline consisting of feature scaling, feature selection and training stages to avoid information leakage during feature selection. 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.

Training and Evaluation

During training, we took the selected features and labels and re-optimized an algorithm's hyperparameters using grid search and 5-fold cross-validation or sliding validation. In this

step, we used a pipeline consisting of feature scaling and training. We selected the model with optimal values of hyperparameters for final evaluation. In the baseline, we optimized a small set of hyperparameters. It is possible to do a more thorough hyperparameter optimization and analysis if required. We evaluated the performance of 4 algorithms: (i) Ridge Regression (Ridge), (ii) K-nearest neighbors (KNN), (iii) Support Vectors Regression (SVR), (iv) Gradient Boosted Decision Trees (GBDT). Ridge (Hoerl & Kennard, 1970) is a linear model with L2-norm as the regularization penalty. KNN (Cover & Hart, 1967; Aha et al., 1991) is an instance-based method relying on similarity of data points in the feature space. In the baseline, KNN is weighted by distance. SVR, an extension of the support vector machines (Boser et al., 1992; Cortes & Vapnik, 1995) algorithm, is based on the principle of structural risk minimization (Guyon et al., 1992; Burges, 1998). For linear data, SVR finds a hyperplane with the largest margin of separation. For nonlinear data, SVR uses kernel functions to map the data to a high dimensional space and finds a separating hyperplane there. In the baseline, we applied SVR with the RBF (radial-basis function). GBDT is an ensemble method which uses gradient boosting (Friedman, 2001) to grow the trees and often provides more accurate results compared to random forests (see Hastie et al. (2009)).

We evaluated all algorithms based on the mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and the coefficient of determination or R^2. For MAE and RMSE, we calculated their normalized counterparts to compare them in percentage terms. The normalized errors were calculated by dividing the mean error with the mean true yield.

Mean Absolute Error (MAE):

$$MAE(Ytrue, Yobs) = \frac{1}{n} * \sum_{i=1}^{n} Ytrue_i - Yobs_i \tag{A.1}$$

Normalized MAE was calculated as MAE divided by the mean of the true yield values.

Mean Absolute Percentage Error (MAPE)

$$MAPE(Ytrue, Yobs) = \frac{1}{n} * \sum_{i=1}^{n} \frac{Ytrue_i - Yobs_i}{Ytrue_i} * 100 \tag{A.2}$$

Root Mean Squared Error (RMSE):

$$RMSE(Ytrue, Yobs) = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (Ytrue_i - Yobs_i)^2} \tag{A.3}$$

Normalized RMSE was calculated as RMSE divided by the mean of the true yield values.

Coefficient of determination or R2 score (R2):

$$R2(Ytrue, Yobs) = 1 - \frac{\sum_{i=1}^{n} (Ytrue_i - Yobs_i)^2}{\sum_{i=1}^{n} (Ytrue_i - mean(Ytrue))^2} \tag{A.4}$$

A.2 Input Data

Input data included WOFOST simulation outputs, weather observations, remote sensing, soil, region centroid coordinates and distance to coast, crop area fractions and yield statistics.

A.2.1 WOFOST Indicators

- WLIM_YB (water-limited dry weight biomass ($kg\ ha^{-1}$))
- WLIM_YS (water-limited dry weight storage organs ($kg\ ha^{-1}$))
- WLAI (water-limited leaf area divided by surface area ($m^2\ m^{-2}$))
- DVS (development stage (0-200))
- RSM (root-zone soil moisture as a percentage of soil water holding capacity)
- TWC (sum of water limited transpiration (cm))

A.2.2 Meteo Indicators

- TMAX (maximum daily air temperature (°C))
- TMIN (minimum daily air temperature (°C))
- TAVG (average daily air temperature (°C))
- PREC (sum of daily precipitation (mm))
- ET0 (sum of daily evapotranspiration of short vegetation (Penman-Monteith, Allen et al. (1998)) (mm))
- RAD (sum of daily global incoming shortwave radiation ($KJ\ m^{-2}\ d^{-1}$))
- CWB (climate water balance, calculated as PREC - ET0)

A.2.3 Remote Sensing Indicators

- FAPAR (Fraction of Absorbed Photosynthetically Active Radiation (Smoothed))

A.2.4 Soil Moisture Indicators

- SM_WP (wilting point)
- SM_FC (field Capacity)
- SM_WHC (water holding capacity, calculated as SM_FC - SM_WP)

A.2.5 Region Centroid Information

- CENTROID_X (longitude)
- CENTROID_Y (latitude)
- DIST_COAST (distance to coast)

A.2.6 Crop Areas

- CROP_AREA (absolute crop area (ha))

A.2.7 Crop Area Fractions

- FRACTION (fraction of crop area of parent NUTS region)

A.2.8 GAES

By GAES data, we mean the combination of elevation, slope, field size and irrigated area. For GAES zonation, see Mücher et al. (2016).

- AEZ_ID (agri-environmental zone ID)
- AVG_ELEV (average elevation)
- STD_ELEV (standard deviation of elevation)
- AVG_SLOPE (average slope)
- STD_SLOPE (standard deviation of slope)
- AVG_FIELD_SIZE (average estimated field size)
- STD_FIELD_SIZE (standard deviation of estimated field size)
- IRRG_AREA_ALL (total irrigated crop area)
- IRRG_AREA_(CROP_ID) (e.g. IRRIG_AREA6 for sugar beets, crop-specific irrigated area). NOTE that there was no crop-specific irrigated area for soft wheat and spring barley. For both these crops, irrigated cereals area was used as a proxy.

A.3 Software and data 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.

Table A.1: Data Summary. Most of the data sources came from the MCYFS database or Eurostat.

Data (indicators)	**Source (Years)**
WOFOST outputs aggregated to NUTS regions (WLIM_YB, WLIM_YS, WLAI, DVS, RSM, TWC)	MCYFS (see Lecerf et al. (2019)). (1979-2018)
Daily gridded weather observations aggregated to NUTS regions (only arable land area). (TAVG, TMIN, TMAX, PREC, ET0, CWB)	JRC (see EC-JRC (2022)). (1979-2018)
Remote sensing indicator (FAPAR) aggregated to NUTS regions (only arable land area)	MCYFS (see Copernicus GLS (2020)). (1999-2018)
Soil data aggregated to NUTS regions (only arable land area) (SM_WHC)	MCYFS (see Lecerf et al. (2019))
Centroids of NUTS region (CENTROID_X, CENTROID_Y, DIST_COAST)	Eurostat (Eurostat, 2020b) (Static)
Reported regional (NUTS2 or NUTS3) yield statistics	NL-CBS (2020), FR-Agreste (2020), DE-RegionalStatistiks (2020). (DE: 1999-2017, FR: 1989-2018, NL: 1994-2018)
Eurostat national (NUTS0) yield statistics	Eurostat (2021a); EC-JRC (2022). (ES, FR, IT, NL: 1971-2018; DE: 1975-2018; BG, HU, PL, RO: 1987-2018)
Crop Areas at NUTS2 or NUTS3	DE (Eurostat, 2021a). Others (EC-JRC, 2022). (BG: 1991-2017, DE: 1999-2018, ES: 1998-2016, FR: 1989-2018, HU: 1996-2016, IT: 1995-2018, NL: 1975-2017, PL: 1995-2017, RO: 1998-2017)
Eurostat national (NUTS0) yield statistics	Eurostat (2021a); EC-JRC (2022). (FR, NL: 1971-2018; DE: 1975-2018)
Crop Area Fractions (of parent NUTS region)	MCYFS (EC-JRC, 2022)). (1979-2018)
Past MCYFS Yield Forecasts (date, Yield forecast)	MCYFS (see van der Velde & Nisini (2019)). (BG, HU, PL, RO: 1999-2018; DE, ES, FR, IT, NL: 1993-2018)

A.4 Results: Supplemental Figures

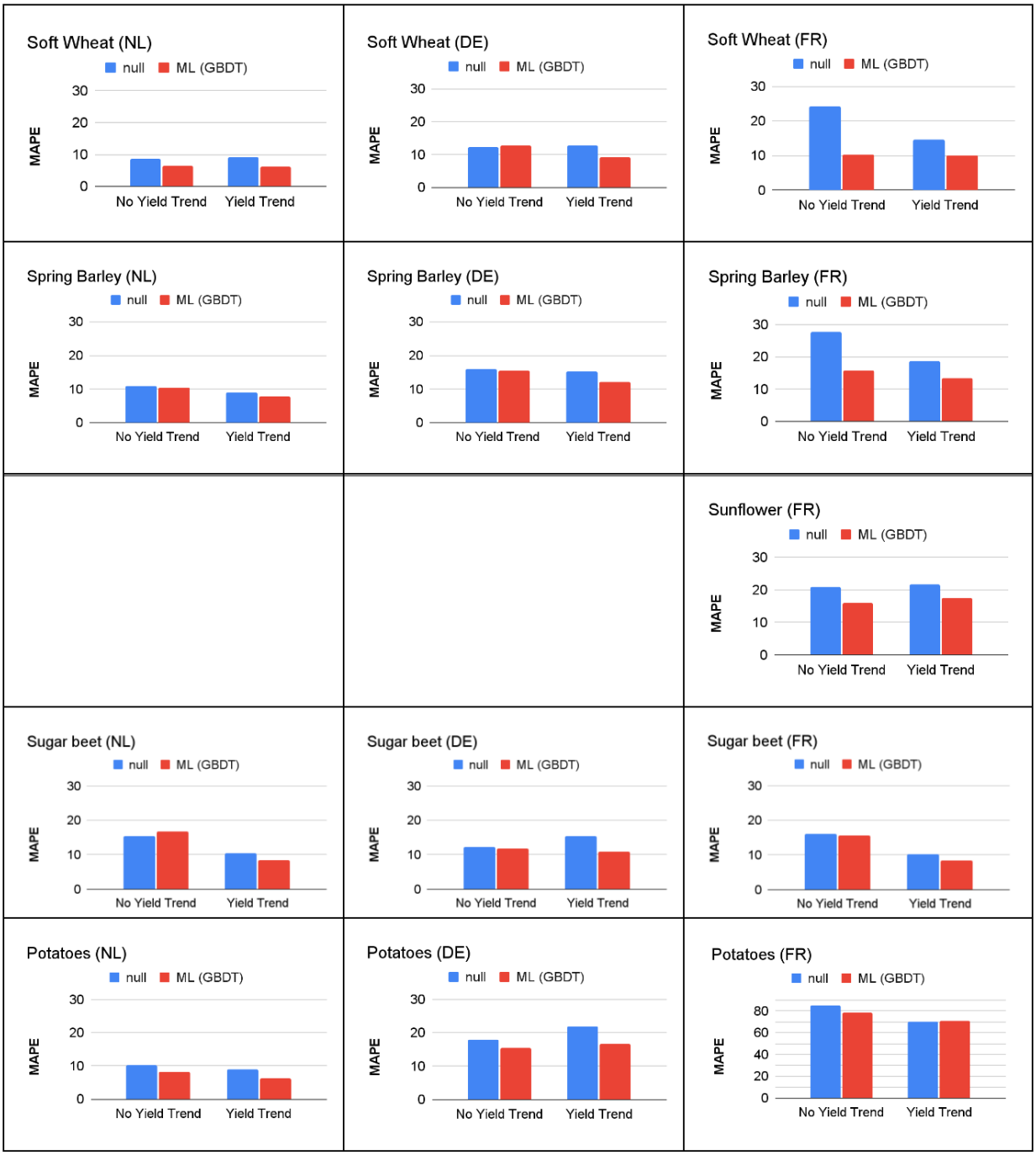


Figure A.3: Yield Trend vs. No Yield Trend. The MAPE of Gradient Boosted Decision Trees was compared with the null method.

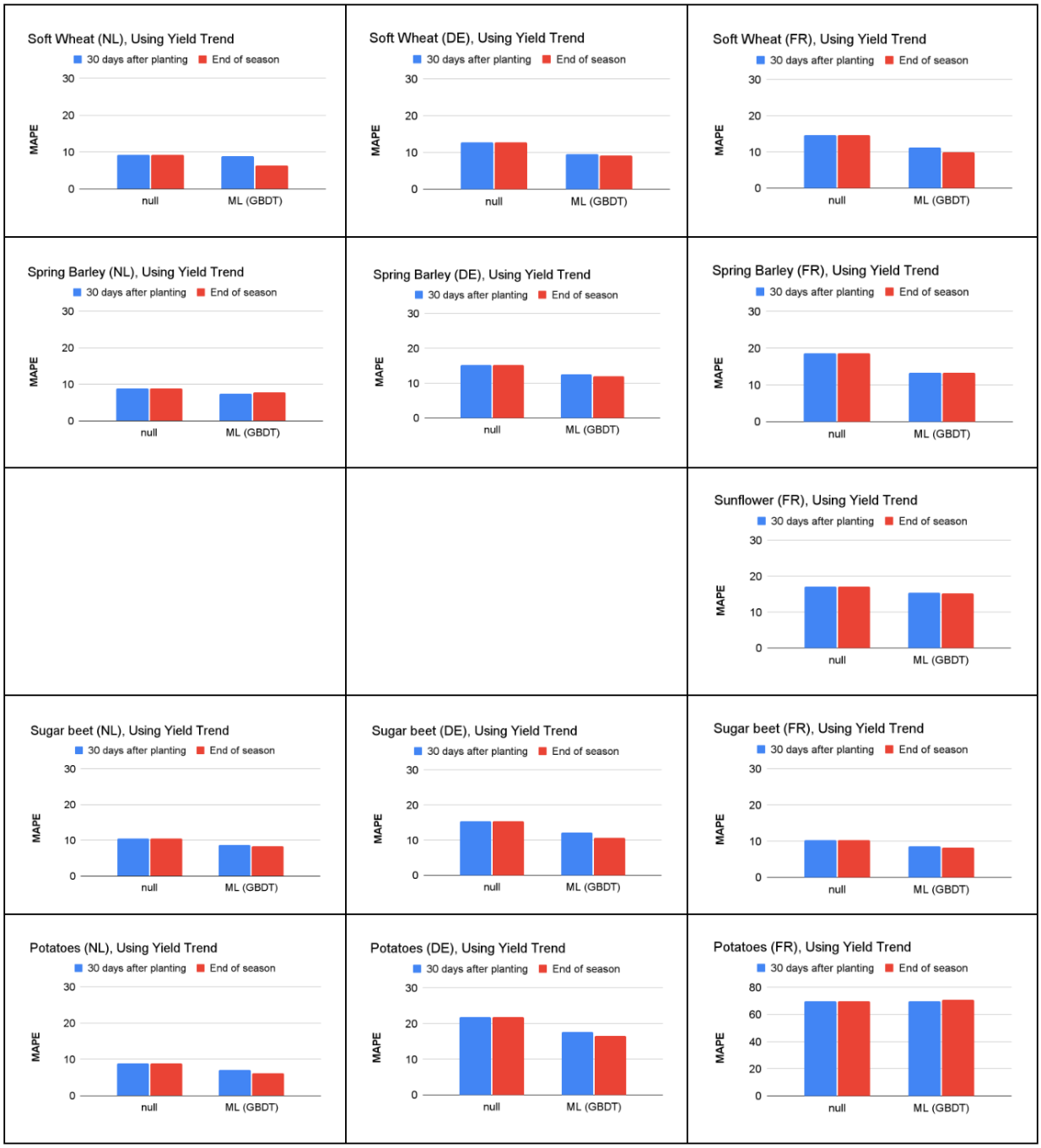
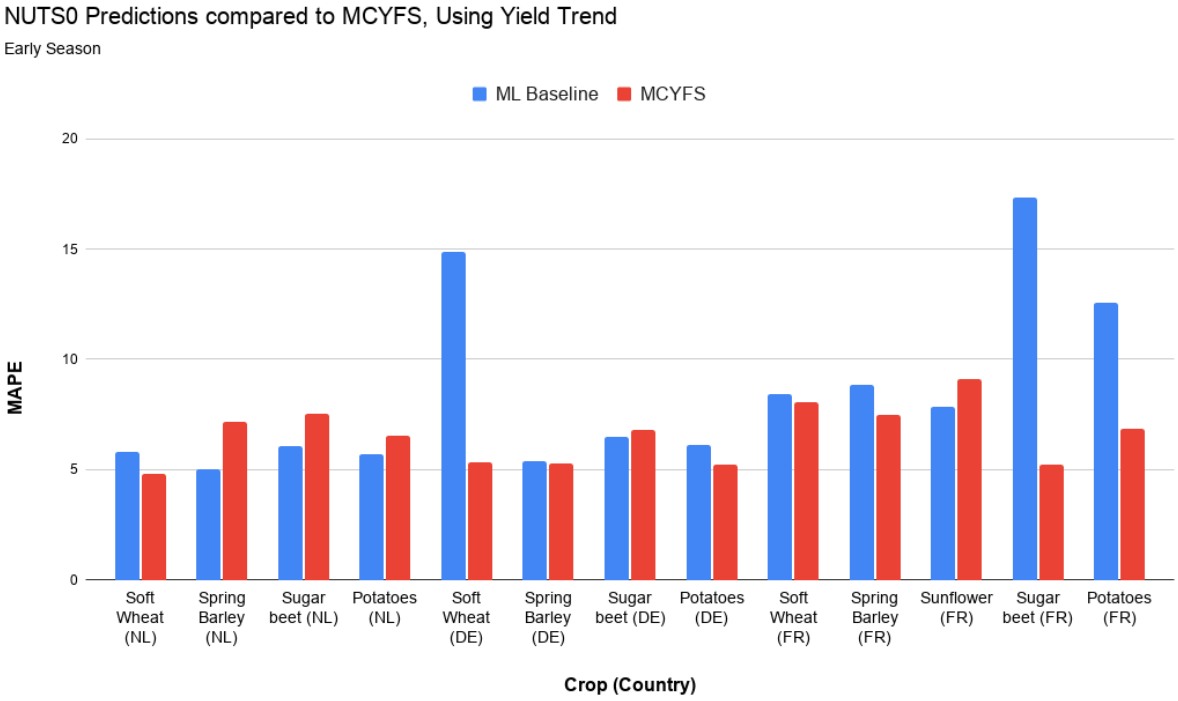
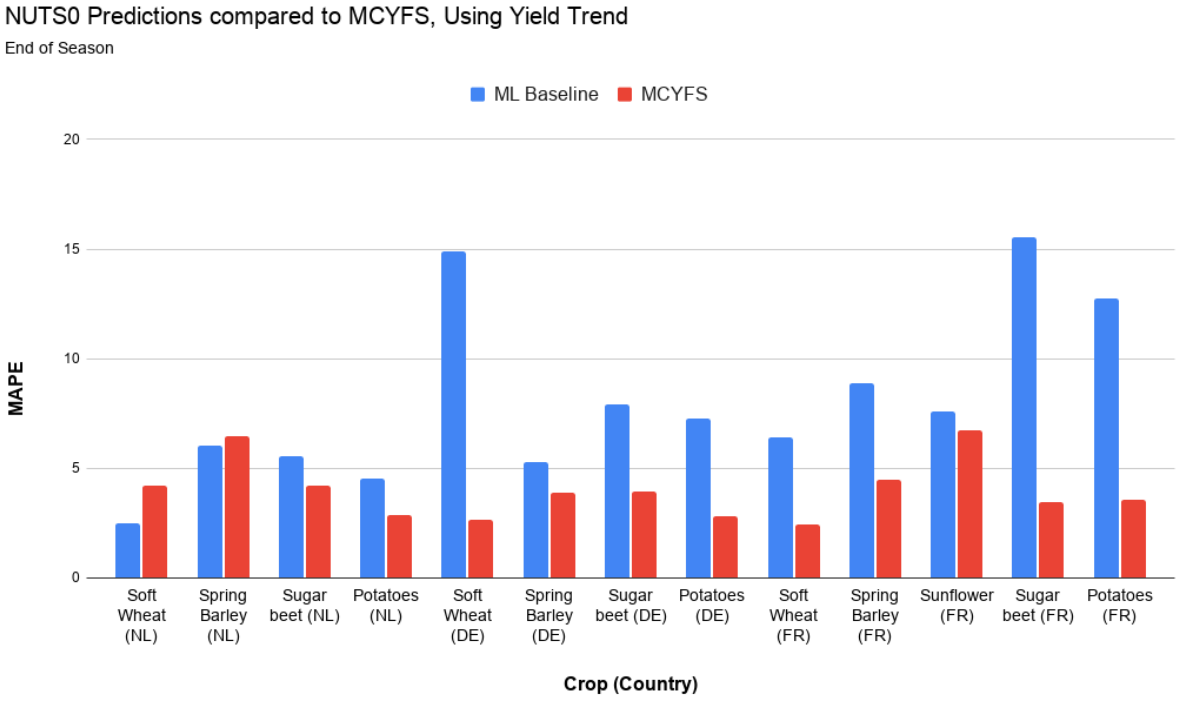


Figure A.4: Early season prediction using a 5-year yield trend The MAPE of Gradient Boosted Decision Trees for early and end of season predictions.



(a) Early season (30 days after planting)



(b) End of season

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

Appendix B

B.1 Case Studies

Table B.1: Case studies for evaluation. We had a total of 35 crop and country combinations from six crops and nine countries (colored yellow). Some crop and country combinations did not have (sufficient) data at the selected NUTS level (colored red). For sunflower (BG) and spring barley (PL), data size was small and the trend window was reduced to 3 to be able to run machine learning.

Crop, Country	Soft wheat	Spring barley	Sunflower	Grain maize	Sugar beets	Potatoes
BG	(no yield data)	(no yield data)	NUTS2 (trend window = 3)	NUTS2	(no yield data)	(no yield data)
DE	NUTS3	NUTS3	(no yield data)	(no yield data)	NUTS3	NUTS3
ES	NUTS3	NUTS3	NUTS3	NUTS3	(2011-2015)	(2010-2015)
FR	NUTS3	NUTS3	NUTS3	NUTS3	NUTS3	NUTS3
HU	(no yield data)	NUTS3	NUTS3	NUTS3	NUTS3	NUTS3
IT	NUTS3	(no yield data)	(no yield data)	NUTS3	(no yield data)	NUTS3
NL	NUTS2	NUTS2	(no yield data)	(2008-2017)	NUTS2	NUTS2
PL	NUTS2	NUTS2 (trend window = 3)	(many 0 values)	(2003-2012)	NUTS2	NUTS2
RO	(no yield data)	(no yield data)	NUTS3	NUTS3	(2011-2016)	NUTS3

B.2 Improvements to the machine learning baseline

Here we focus on parts of the machine learning baseline (Paudel et al., 2021) that were updated in the optimized configuration.

B.2.1 Data sources

The baseline used WOFOST simulation outputs, Soil (soil water holding capacity), Meteo, Remote Sensing (Fraction of Absorbed Photosynthetically Active Radiation (Smoothed)), Regional yield statistics. In the baseline, the features passed to machine learning did not explicitly indicate different regions. We expected machine learning models to distinguish different NUTS regions based on soil water holding capacity, which does not change year over year. In the updated workflow, we added data about elevation, slope, crop area, irrigated crop area and field size. These data sources, with the exception of crop area, are static and help capture spatial differences. Crop area includes some temporal variation as well.

B.2.2 Data preprocessing

Data preprocessing aligned data for each source to the same spatial and temporal resolution. Spatial resolution refers to the selected NUTS level. Temporal resolution was dekadal for WOFOST and Remote Sensing, daily for Meteo and yearly for yield statistics. We did not do any data cleaning or imputation. Some crops, such as winter wheat, have growing seasons that cross the calendar year boundary. In order to make the workflow support such crops, we transformed the data to align with the campaign year instead of the calendar year. A campaign year started after the end of the previous growing season and stretched up to the end of the current growing season.

In the updated workflow, we identified sequences of duplicate yield values and missing years. In the case of long or multiple such sequences, the entire region was removed from analysis. When there was only one short sequence, we removed the data points, but not the region. A long sequence consisted of 5 values. A short sequence consisted of 2-4 values.

B.2.3 Feature Design

First, we updated the table used for feature design (*Table 2.2*, *Table B.2*). We added RAD to the vegetative phase and the yield formation phase. We changed CWB features to average the cumulative values. We also added information from the previous campaign year by taking features from the previous harvest window. The main change in the feature design table is the way extreme conditions are captured. In the machine learning baseline, features for extreme conditions counted the number of days or dekads with values crossing certain thresholds. Such features had many data points with zero values. We replaced them with the standard score or z-score based on the long-term average and standard deviation. Z-score features were less sparse and also captured the magnitude of the extremes from the long term average. We calculated z-scores based on long-term average and standard deviations of agri-environmental zones instead of the whole country. We mapped NUTS regions to the GAES agri-environmental zones with the maximum overlapping area. Second, we made the crop calendar (inferred from WOFOST simulation outputs) per region and per year. In contrast, the machine learning baseline used the same crop calendar for the whole country. We expected the precise crop calendar to help in countries with multiple agro-environmental zones. Third, we added data to capture spatial differences in elevation, slope, field size, crop area and irrigated crop area (*Table B.3*).

B.2.4 Feature Selection

In the baseline we used three methods for feature selection: (i) Random Forest (Breiman, 2001), (ii) Recursive Feature Elimination (RFE) based on Lasso (Tibshirani, 1996), and (iii) Mutual Information Regression (MI). RFE (e.g. Granitto et al. (2006), recursively eliminates unimportant features by evaluating a machine learning algorithm which provides feature weights or feature importances. MI (Shannon, 1948) is a univariate feature selection method, similar to Pearson's r (see Benesty et al. (2009)), that calculates the information content of individual features. We had an implicit combined method that unioned the features selected by the other methods.

Table B.2: Feature design in the updated workflow. See *Table 2.1* for crop calendar definition and *Table 2.2* for feature design in the baseline. We added RAD to the vegetative and yield formation phase and made CWB features take the average of cumulative values. In addition, we added features for PREC from the previous season's harvest window.

Period	**Maximum Values**	**Average Values, Cumulative Sums***	**z-scores based on long term avg, std**
Pre-planting		TAVG, PREC, CWB*	
Planting		TAVG, PREC	RSM, TMIN, PREC
Vegetative	WLIM_YB, TWC, WLAI	RSM, TAVG, CWB*, RAD, FAPAR	RSM
Flowering		PREC	RSM, PREC, TMAX
Yield Formation	WLIM_YB, WLIM_YS, TWC, WLAI	RSM, CWB, RAD, FAPAR	RSM
Harvest		PREC	PREC

In the updated workflow, we removed feature selection based on mutual information and the combined method. We replaced mutual information by selection based on mutual correlation. We dropped features with a mutual correlation of > 0.9 (kept one and removed others) before passing them to machine learning. We removed the combined method mainly for explainability.

B.2.5 Hyperparameter Search

We combined hyperparameter optimization of feature selection and model training steps. We used a more robust and adaptive hyperparameter search based on Bayesian optimization (Brochu et al., 2010; Shahriari et al., 2015). Bayesian optimization optimizes an acquisition function that estimates the quality of the next hyperparameter configuration based on the configurations tried before and trades off between exploring unexplored parts of configuration space and exploiting parts of configuration space known to improve performance. We evaluated the performance of hyperparameter configurations on the k-fold validation error (mean squared error of predictions).

We also explored random search (Bergstra & Bengio, 2012) and BOHB (Falkner et al., 2018), a method that combines the benefits of Bayesian optimization and Hyperband (Li et al., 2017). Hyperband leverages a bandit strategy that dynamically allocates resources to a set of hyperparameter configurations and uses successive halving (Jamieson & Talwalkar, 2016) to terminate poorly performing ones. In the end, we decided to use the Bayesian optimization package called scikit-optimize (Scikit-optimize Contributors, 2021) that is compatible with scikit-learn.

B.2.6 Validation splits and Model Refitting

Updates to the custom 5-fold sliding validation and per-test-year model refitting are described in *Section 3.2.3*.

B.3 Data Sources

File format: CSV. **File name format**: <source>_<NUTS level>_<country>.csv

Section A.2 provides details about data sources.

Table B.3: Data sources summary. Most of the data sources came from the MCYFS database or Eurostat.

Data (indicators)	Source (Years)
WOFOST outputs aggregated to NUTS regions (WLIM_YB, WLIM_YS, WLAI, DVS, RSM, TWC)	MCYFS (see Lecerf et al. (2019)). (1979-2018)
Daily gridded weather observations aggregated to NUTS regions (only arable land area). (TAVG, TMIN, TMAX, PREC, ET0, CWB)	JRC (see EC-JRC (2022)). (1979-2018)
Remote sensing indicator (FAPAR) aggregated to NUTS regions (only arable land area)	MCYFS (see Copernicus GLS (2020)). (1999-2018)
Soil data aggregated to NUTS regions (only arable land area) (SM_WHC)	MCYFS (see Lecerf et al. (2019))
Reported regional (NUTS2 or NUTS3) yield statistics	NL-CBS (2020), FR-Agreste (2020), DE-RegionalStatistiks (2020). Others: Eurostat (2021a); EC-JRC (2022). (BG: 2001-2017, DE: 1999-2017, ES: 1998-2016, FR: 1989-2018, HU: 2000-2016, IT: 1995-2018, NL: 1994-2018, PL: 1995-2016, RO: 1998-2017)
Eurostat national (NUTS0) yield statistics	Eurostat (2021a); EC-JRC (2022). (ES, FR, IT, NL: 1971-2018; DE: 1975-2018; BG, HU, PL, RO: 1987-2018)
Crop Areas at NUTS2 or NUTS3	DE (Eurostat, 2021a). Others (EC-JRC, 2022). (BG: 1991-2017, DE: 1999-2018, ES: 1998-2016, FR: 1989-2018, HU: 1996-2016, IT: 1995-2018, NL: 1975-2017, PL: 1995-2017, RO: 1998-2017)
Crop Area Fractions (of parent NUTS region)	MCYFS (EC-JRC, 2022)). (1979-2018)
Elevation, Slope (AVG, STD)	USGS EROS Archive (USGS-EROS, 2021)
Irrigated Crop Area (total and per-crop)	MCYFS (EC-JRC, 2022))
Estimated Field Sizes (AVG and STD)	Lesiv et al. (2019)
Past MCYFS Yield Forecasts (date, Yield forecast)	MCYFS (see van der Velde & Nisini (2019)). (BG, HU, PL, RO: 1999-2018; DE, ES, FR, IT, NL: 1993-2018)

B.4 Software and data 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.

B.5 Supplementary Results

Here we have included supplementary figures and tables referenced in the main text.

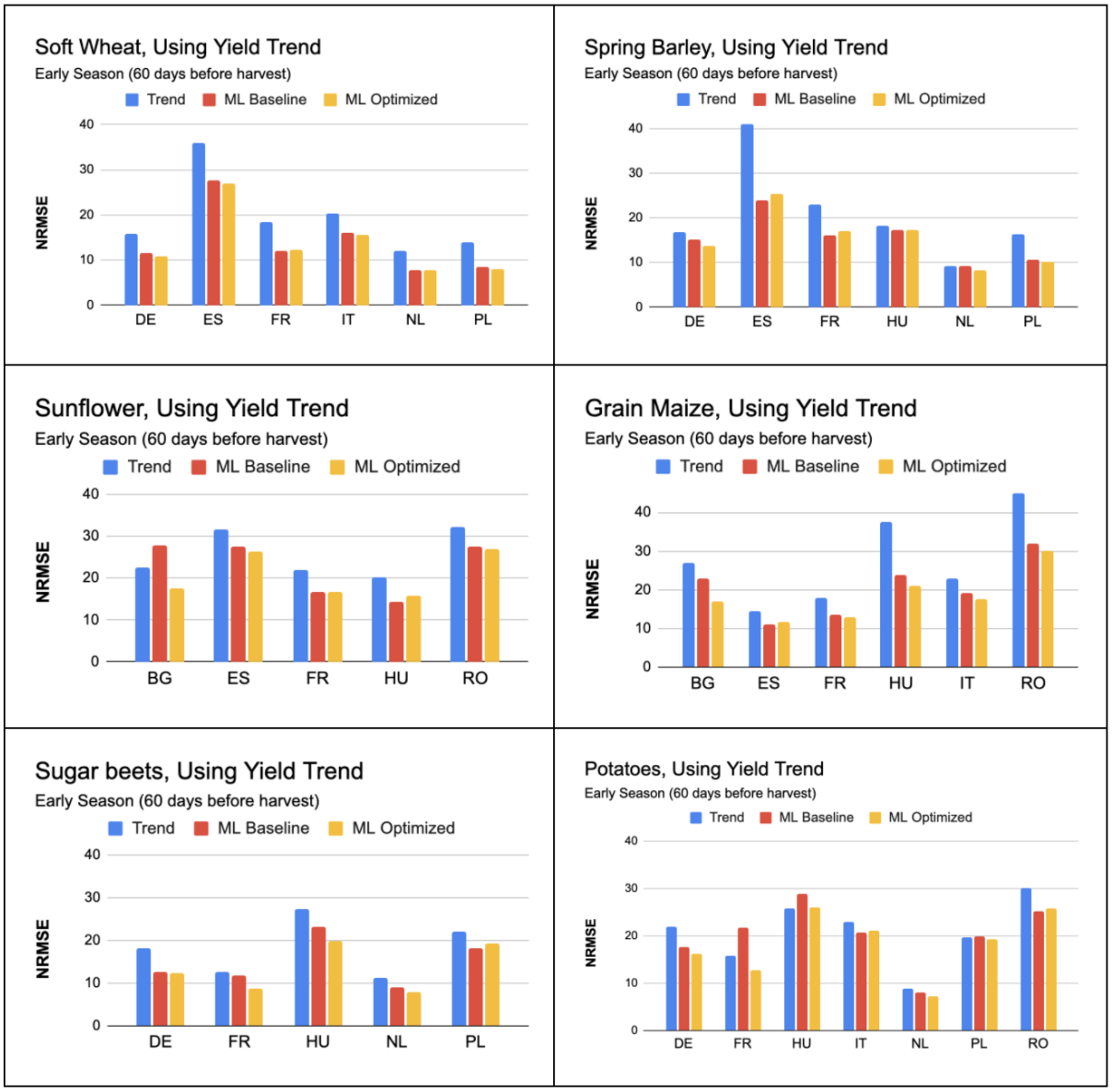


Figure B.1: (a) Regional Normalized RMSE 60 days before harvest. We compared the machine learning baseline and optimized machine learning models with a linear trend model. For ML Baseline and ML Optimized, we show the algorithm with the lowest normalized RMSE.

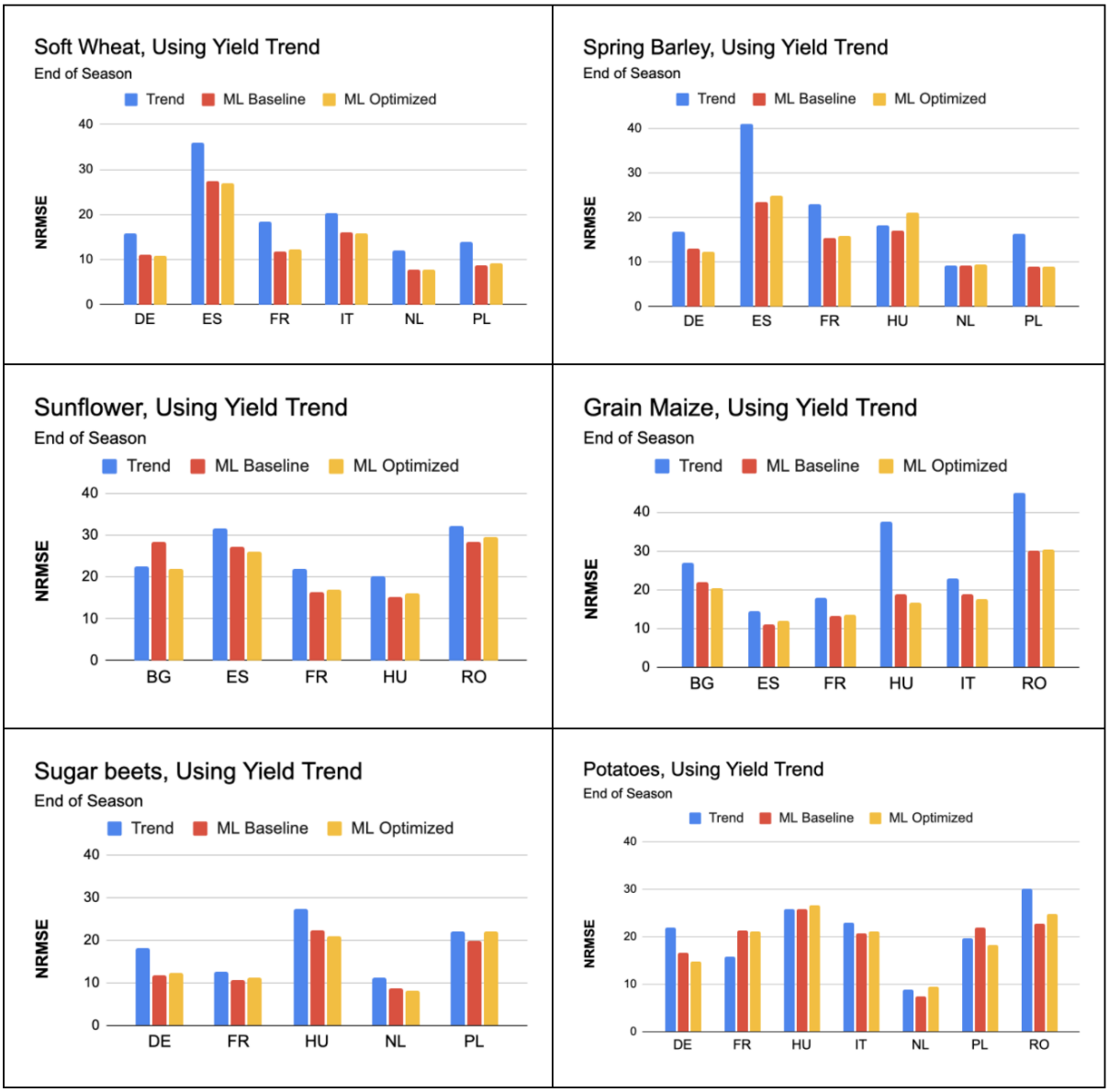


Figure B.1: (b) Regional Normalized RMSE for end of season. We compared the machine learning baseline and optimized machine learning models with a linear trend model. For ML Baseline and ML Optimized, we show the algorithm with the lowest normalized RMSE.

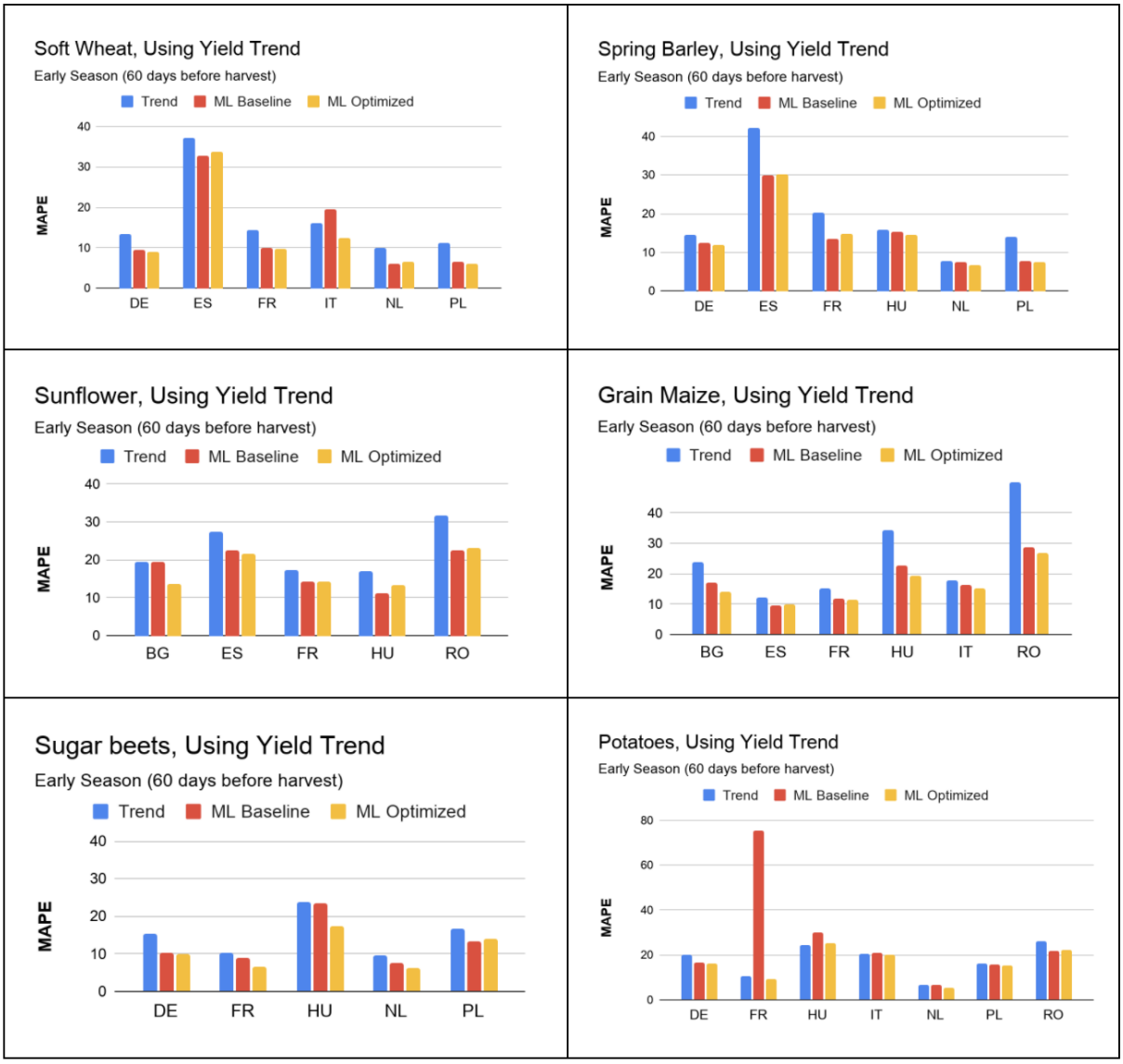


Figure B.2: (a) Regional MAPE for 60 days before harvest. We compared the machine learning baseline and optimized machine learning models with a linear trend model. For ML Baseline and ML Optimized, we show the algorithm with the lowest normalized RMSE.

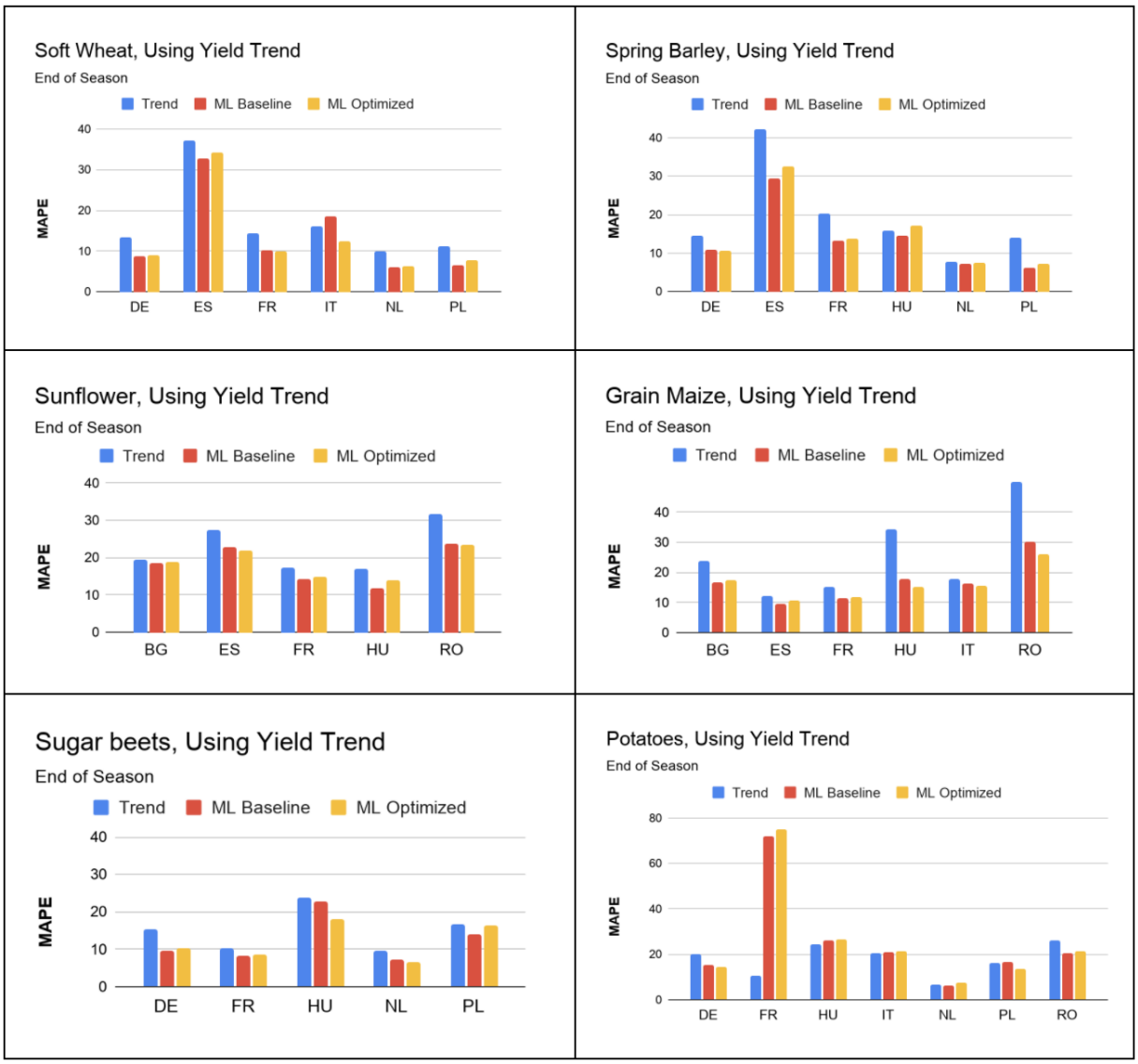


Figure B.2: (b) Regional MAPE for end of season. We compared the machine learning baseline and optimized machine learning models with a linear trend model. For ML Baseline and ML Optimized, we show the algorithm with the lowest normalized RMSE.

Table B.4: Normalized RMSEs for regional forecasts during early season and end of season. For ML Optimized, we selected the algorithm with the lowest normalized RMSE.

Crop (Country)	Trend	ML Optimized 120 days early	ML Optimized 60 days early	ML Optimized End of season
Soft wheat (DE)	15.29	10.73	10.84	10.62
Soft wheat (ES)	37.27	27.63	26.97	26
Soft wheat (FR)	18.66	13.13	12.2	12.2
Soft wheat (IT)	20.97	15.2	15.56	15.5
Soft wheat (NL)	11.36	8	7.77	8.08
Soft wheat (PL)	13.71	8.32	8.02	8.64
Spring barley (DE)	16.78	13.51	13.73	12.92
Spring barley (ES)	41	31.47	25.45	24.84
Spring barley (FR)	23.04	16.93	17.05	15.88
Spring barley (HU)	18.24	17.13	17.19	17.3
Spring barley (NL)	9.16	8.32	8.31	9.15
Spring barley (PL)	16.39	13.59	10.25	9.88
Sunflower (BG)	22.62	16.27	17.54	20.95
Sunflower (ES)	31.68	27.09	26.16	26.18
Sunflower (FR)	21.85	17.54	16.57	17.1
Sunflower (HU)	20.22	16.52	15.8	17
Sunflower (RO)	32.12	27.64	27.75	26.67
Grain maize (BG)	27.09	25.37	16.99	17.89
Grain maize (ES)	14.38	11.74	11.57	11.58
Grain maize (FR)	17.87	13.52	12.95	13.06
Grain maize (HU)	37.56	25.31	21.08	18.24
Grain maize (IT)	22.92	18.4	17.7	17.58
Grain maize (RO)	44.95	34.26	30.2	29.77
Sugar beets (DE)	18.16	13	12.37	12.37
Sugar beets (FR)	12.55	8.77	8.61	8.88
Sugar beets (HU)	27.27	19.61	19.9	18.32
Sugar beets (NL)	11.32	8.67	7.83	8.23
Sugar beets (PL)	21.99	19.33	19.26	19.56
Potatoes (DE)	21.86	16.62	16.11	15.11
Potatoes (FR)	15.81	12.37	12.69	11.76
Potatoes (HU)	25.76	25.56	26.07	24.4
Potatoes (IT)	22.9	20.66	20.66	20.47
Potatoes (NL)	8.82	6.34	7.22	6.46
Potatoes (PL)	19.63	18.79	19.35	15.41
Potatoes (RO)	30.09	24.48	25.81	25.36
Wins/losses	-	350	341	350
Wilcoxon p-value	-	2e-7	3e-7	2e-7

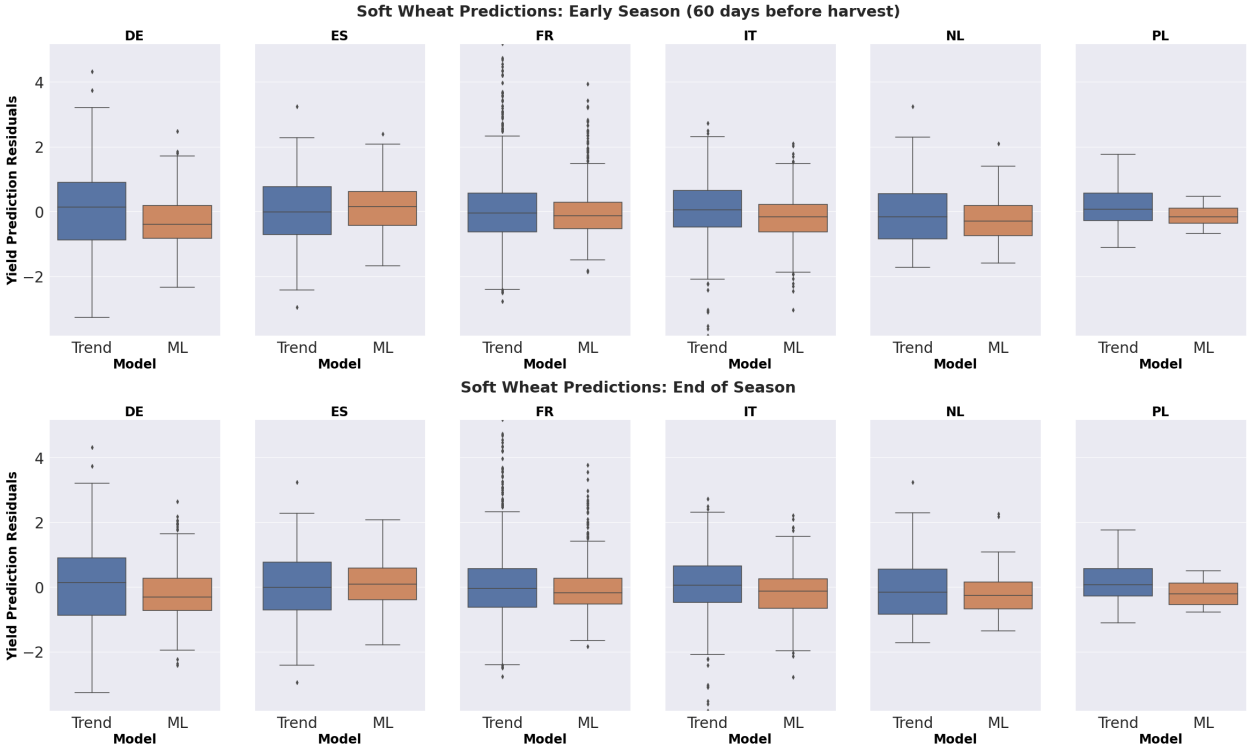


Figure B.3: (a) Box plots of regional prediction residuals (Soft wheat). For machine learning, we show the algorithm with the lowest normalized RMSE for the optimized configuration.

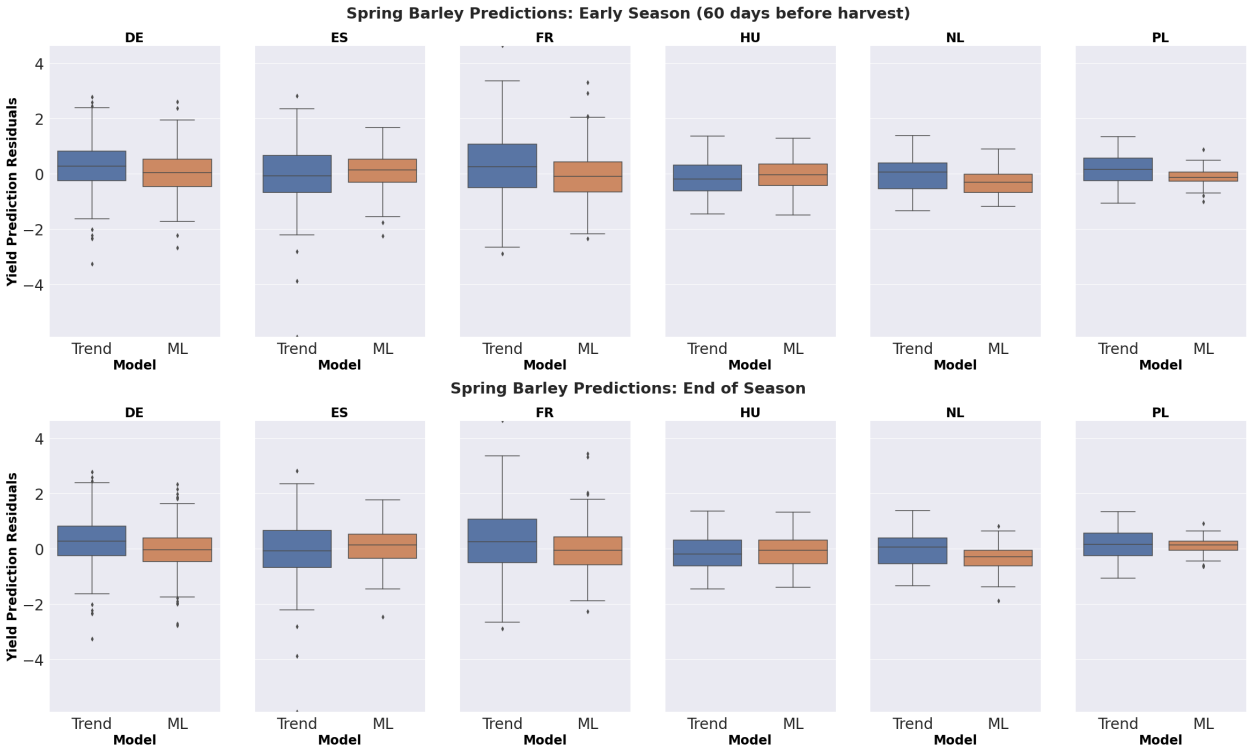


Figure B.3: (b) Box plots of regional prediction residuals (Spring barley). For machine learning, we show the algorithm with the lowest normalized RMSE for the optimized configuration.

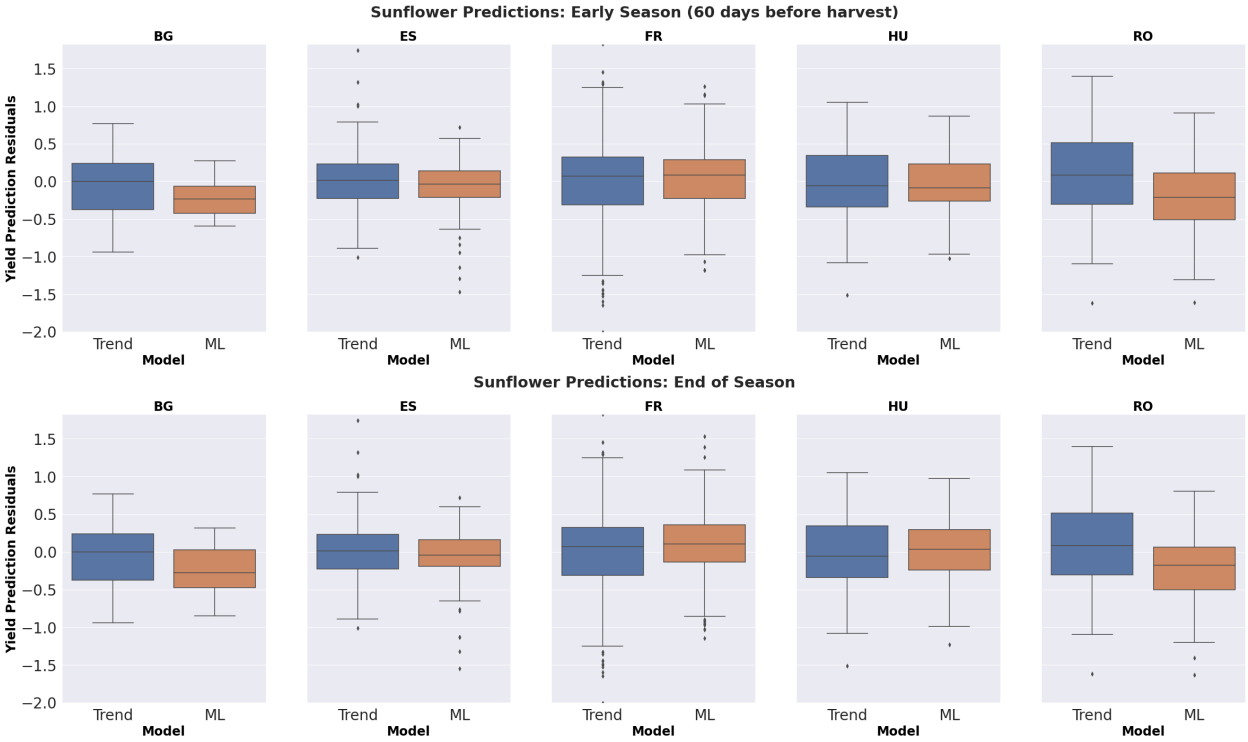


Figure B.3: (c) Box plots of regional prediction residuals (Sunflower). For machine learning, we show the algorithm with the lowest normalized RMSE for the optimized configuration.

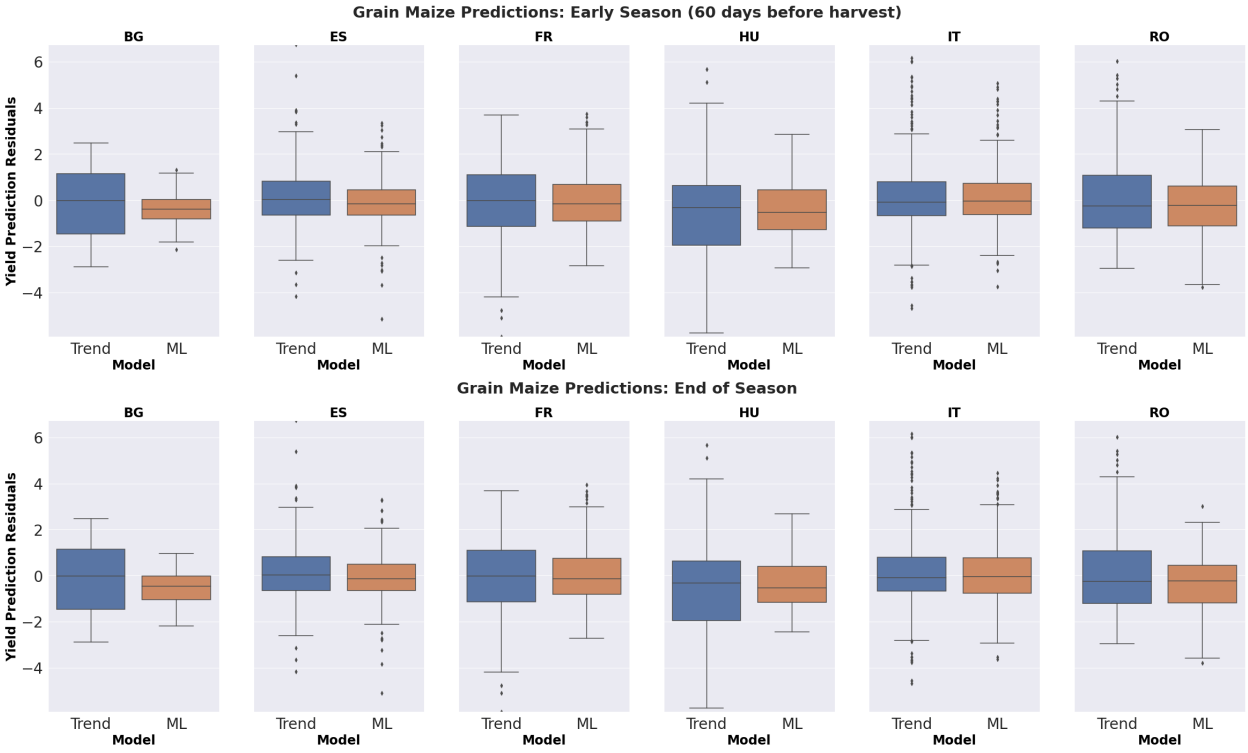


Figure B.3: (d) Box plots of regional prediction residuals (Grain maize). For machine learning, we show the algorithm with the lowest normalized RMSE for the optimized configuration.

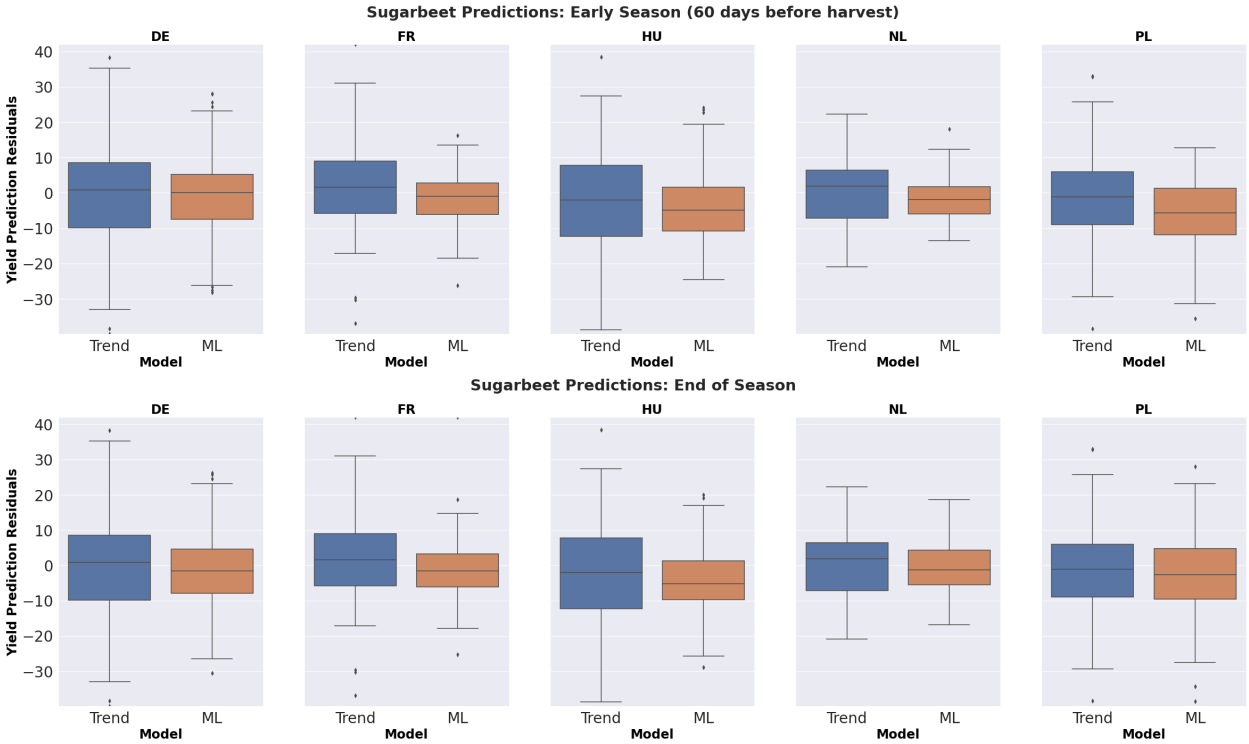


Figure B.3: (e) Box plots of regional prediction residuals (Sugar beets). For machine learning, we show the algorithm with the lowest normalized RMSE for the optimized configuration.

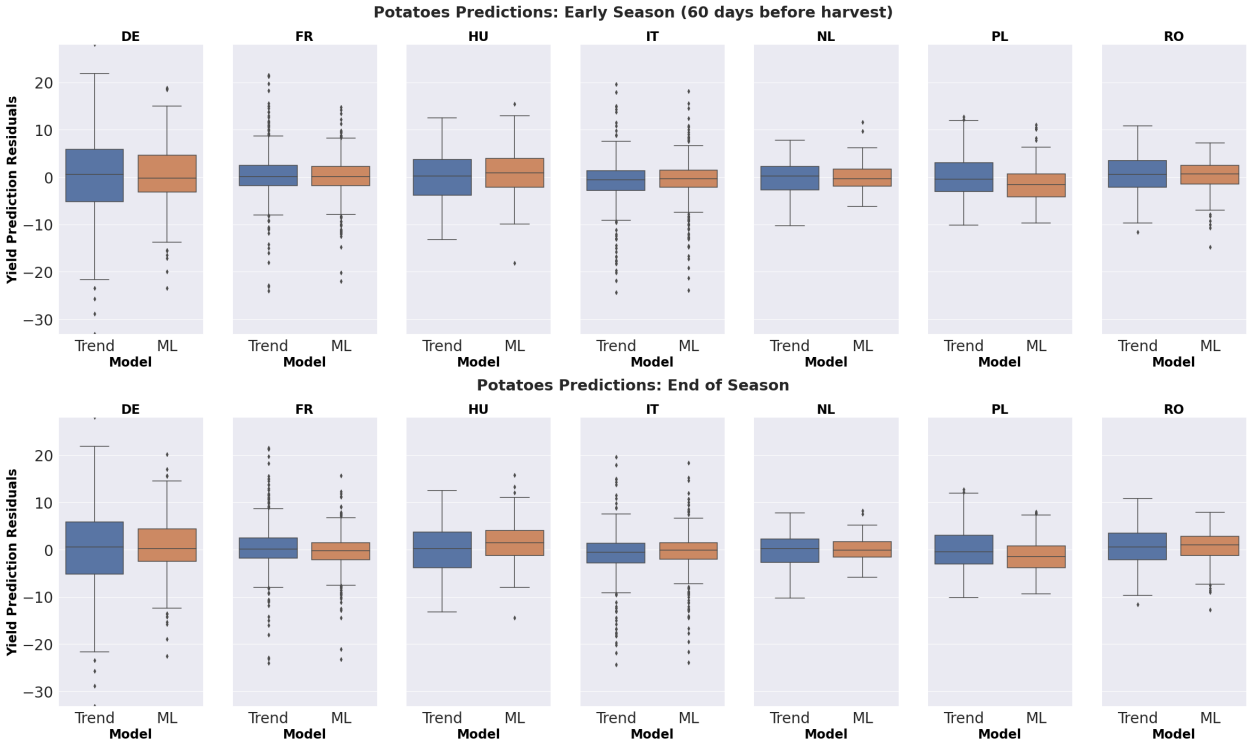


Figure B.3: (f) Box plots of regional prediction residuals (Potatoes). For machine learning, we show the algorithm with the lowest normalized RMSE for the optimized configuration.

Table B.5: Comparison of coefficient of variation for yield trend and machine learning. For cases where yield trend residuals had a low average (<= 10%) and high variance (standard deviation > 25%), optimized machine learning prediction residuals almost always had a lower coefficient of variation, indicating lower uncertainty.

Crop (Country)	Test Year	Trend AVG	Trend STD	ML AVG	ML STD	Trend CV	ML CV
Soft wheat (ES)	2015	0.24	28.93	11.19	29.96	122.55	2.41
Spring barley (ES)	2015	-0.35	28.64	7.29	28.80	82.46	2.99
Sunflower (ES)	2011	-3.27	31.17	-5.62	25.57	9.54	4.55
Sunflower (ES)	2015	7.97	23.38	1.79	17.04	3.43	9.54
Grain maize (ES)	2012	0.98	28.06	3.46	15.56	28.49	4.50
Grain maize (IT)	2009	5.05	28.54	7.76	25.48	5.65	3.28
Sugar beets (HU)	2010	-2.23	31.44	7.97	33.95	14.09	4.26
Potatoes (DE)	2016	5.01	69.81	19.21	115.71	13.94	6.02
Potatoes (IT)	2012	7.75	35.59	15.27	65.41	4.59	4.28
Potatoes (IT)	2013	3.43	77.85	15.31	81.28	22.68	5.31
Potatoes (IT)	2014	-6.07	62.8	7.85	40.86	10.34	5.21

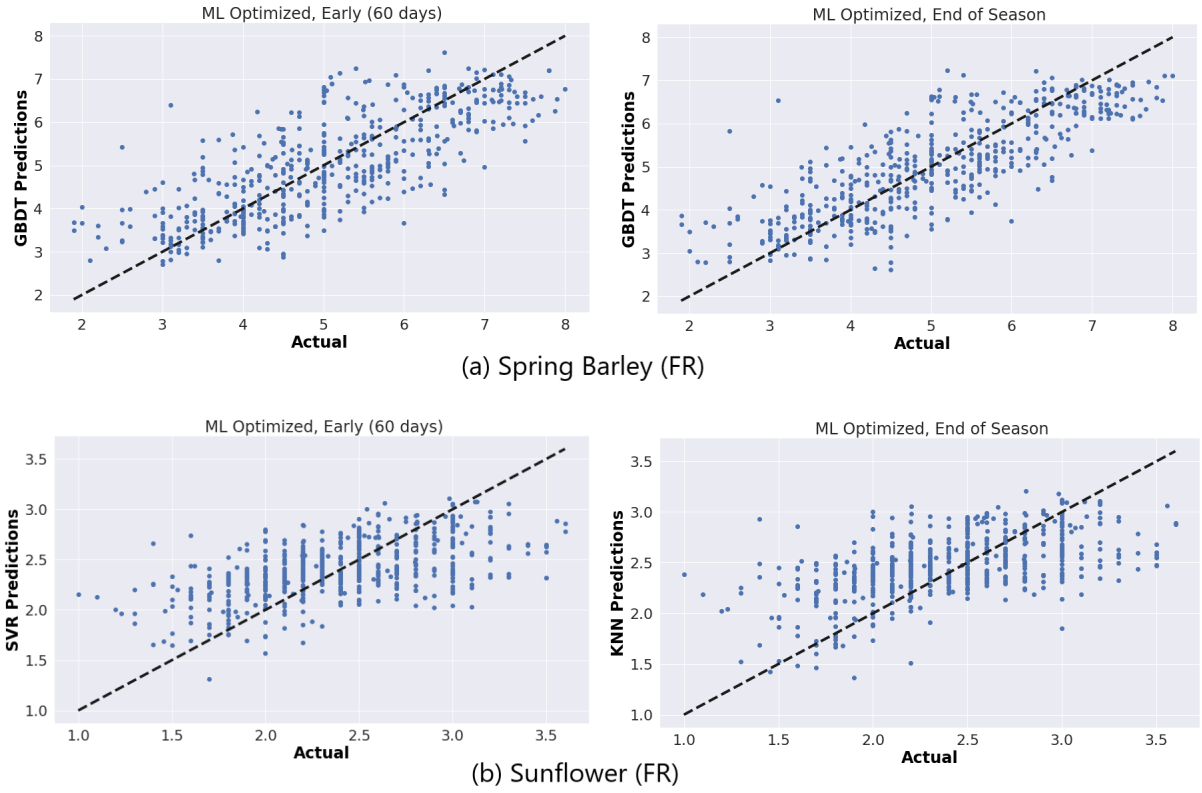


Figure B.4: Cases showing visible data quality issues. (a) Spring Barley (FR) and (b) Sunflower (FR) show many near duplicate values of yield labels. We show predictions of the algorithm with the lowest normalized RMSE.

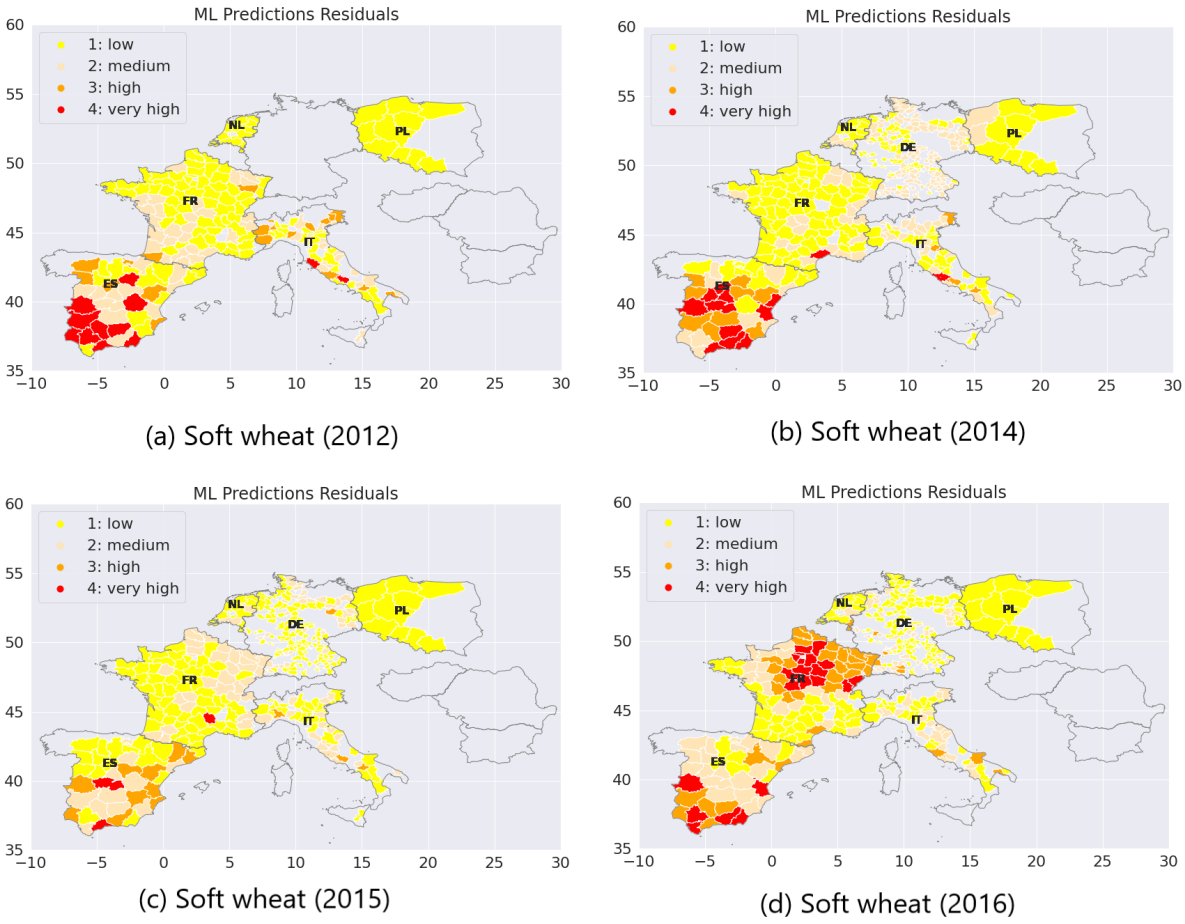


Figure B.5: Spatial patterns of prediction residuals for soft wheat 60 days before harvest. The plots show four classes for absolute percentage residuals. Low: <= 10%. Medium: 10-25%. High: 25-50%. Very high: >50%.

	P1	P2	P3	P4	P5		Country	Off by 1	Off by 2
T1	11	9					DE	33	
T2	2	94	30	1			FR	20	1
T3		20	106	11			HU	8	
T4			16	22	2		IT	13	1
T5			1	3	3		NL	4	
							PL	3	
							RO	12	

(a) Potatoes (2013)

	P1	P2	P3	P4	P5		Country	Off by 1	Off by 2
T1	0	7	8				BG	1	2
T2	3	24	30	4			ES	9	
T3		7	30	18			FR	35	10
T4			2	30	2		HU	10	1
T5				4	9		RO	18	

(b) Grain Maize (2015)

	P1	P2	P3	P4	P5		Country	Off by 1	Off by 2	Off by > 2
T1	6	5	4	2	1		DE	51	1	
T2	5	34	37	11	0		ES	18	1	
T3		12	91	43	1		FR	33	14	1
T4			22	41	2		IT	21		
T5				8	1		NL	8		1
							PL	3		

(c) Soft Wheat (2016)

Figure B.6: Confusion Matrices for reported vs predicted yield classes 60 days before harvest. P1-P5 are predicted classes 1 to 5. T1-T5 are true or reported classes 1 to 5. The machine learning model and yield classes are per country. Very low: <= 20% of min-max range; Low: 20-40%; Medium: 40-60%; High: 60-80%; Very high >80%.

Table B.6: Comparison of national Normalized RMSEs. We compared past MCYFS forecasts and predictions of the machine learning algorithm with the lowest normalized RMSE from the baseline (orange) or the optimized configuration (blue).

Crop (Country)	MCYFS 120 days early	ML optimized 120 days early	MCYFS 60 days early	ML optimized 60 days early
Soft wheat (DE)	6.54	7.9	6.03	8.41
Soft wheat (ES)	12.74	9.79	10.44	4.15
Soft wheat (FR)	11.44	10.55	10.66	9.72
Soft wheat (IT)	4.90	7.28	5.27	6.86
Soft wheat (NL)	6.35	5.68	6.42	4.26
Soft wheat (PL)	6.02	5.24	5.97	5.06
Spring barley (DE)	8.05	4.84	6.21	5.93
Spring barley (ES)	18.05	16.1	17.8	9.83
Spring barley (FR)	10.21	12.48	9.86	11.99
Spring barley (HU)	13.18	10.13	11.46	10.91
Spring barley (NL)	8.91	7.56	8.94	6.45
Spring barley (PL)	11.26	13.02	11.77	4.73
Sunflower (BG)	11.89	14.92	5.16	16.18
Sunflower (ES)	13.1	12.27	8.81	7.84
Sunflower (FR)	10.46	13.82	10.15	12.13
Sunflower (HU)	10.49	11.86	11.13	12.26
Sunflower (RO)	17.06	20.34	13.22	24.34
Grain maize (BG)	21.94	26.85	11.11	11.59
Grain maize (ES)	3.77	3.3	3.03	3.08
Grain maize (FR)	9.09	7.62	6.48	7.33
Grain maize (HU)	29.27	24.02	16.95	18.82
Grain maize (IT)	8.51	7.65	5.65	6.83
Grain maize (RO)	25.08	29.55	24.49	24.83
Sugar beets (DE)	8.78	6.79	8.72	6.27
Sugar beets (FR)	7.45	3.54	4.65	5.43
Sugar beets (HU)	13.78	11.71	14.7	14.23
Sugar beets (NL)	9.14	6.37	5.56	5.45
Sugar beets (PL)	9.18	8.47	8.93	8.53
Potatoes (DE)	6.23	4.19	5.23	4.48
Potatoes (FR)	7.83	12.84	7.79	13.22
Potatoes (HU)	12.38	12.65	15.65	12.83
Potatoes (IT)	5.39	4.79	4.51	4.93
Potatoes (NL)	8.14	3.43	4.4	3.7
Potatoes (PL)	15.09	12.41	13.46	13.18
Potatoes (RO)	17.85	15.25	16.8	16.86
Wins/losses	-	23/12	-	18/17
Wilcoxon p-value	-	0.31	-	0.95

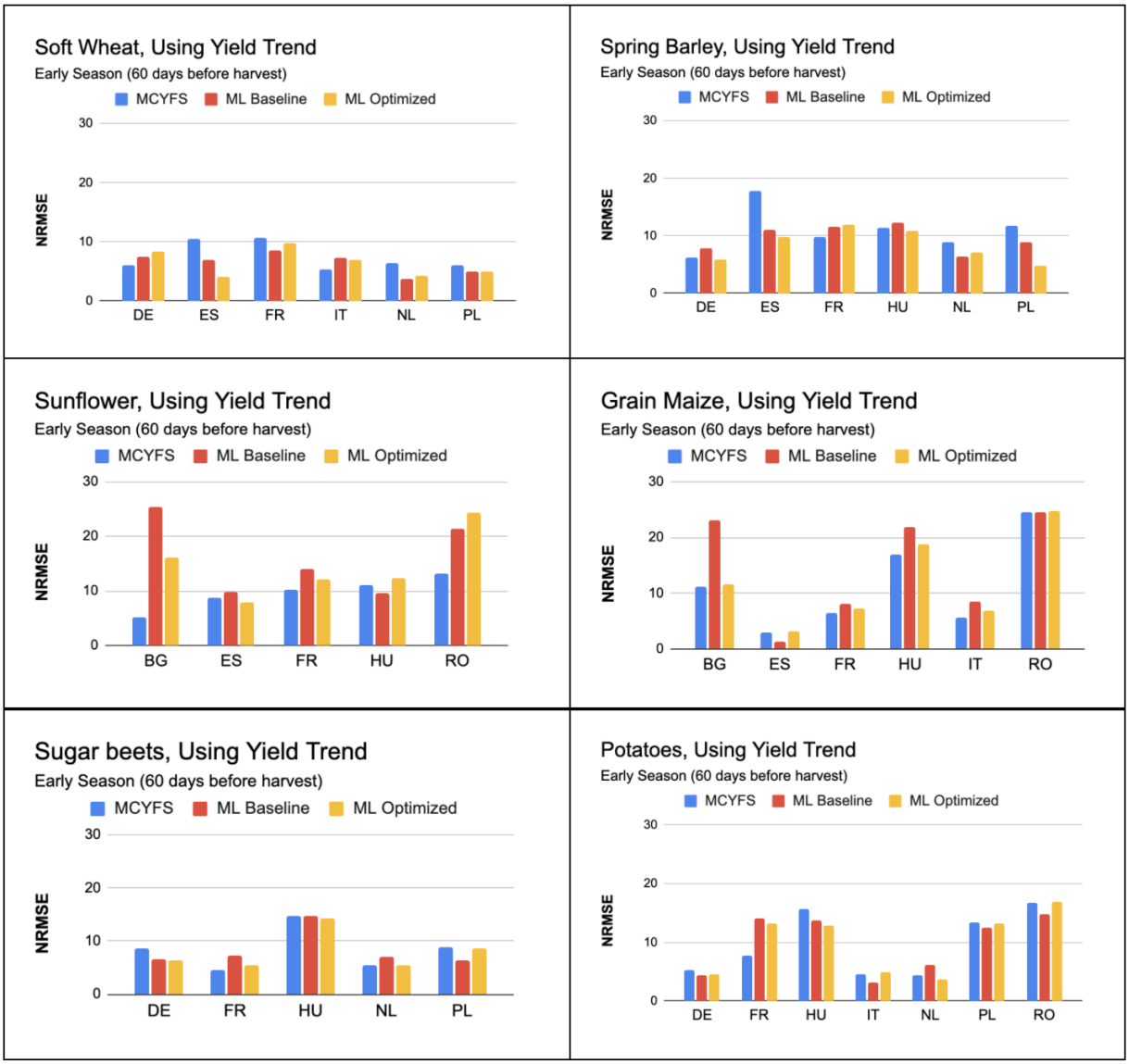


Figure B.7: (a) National Normalized RMSE 60 days before harvest. The machine learning baseline and the optimized machine learning models were compared to MCYFS. For both the ML Baseline and the ML Optimized, we show the algorithm with the lowest normalized RMSE.

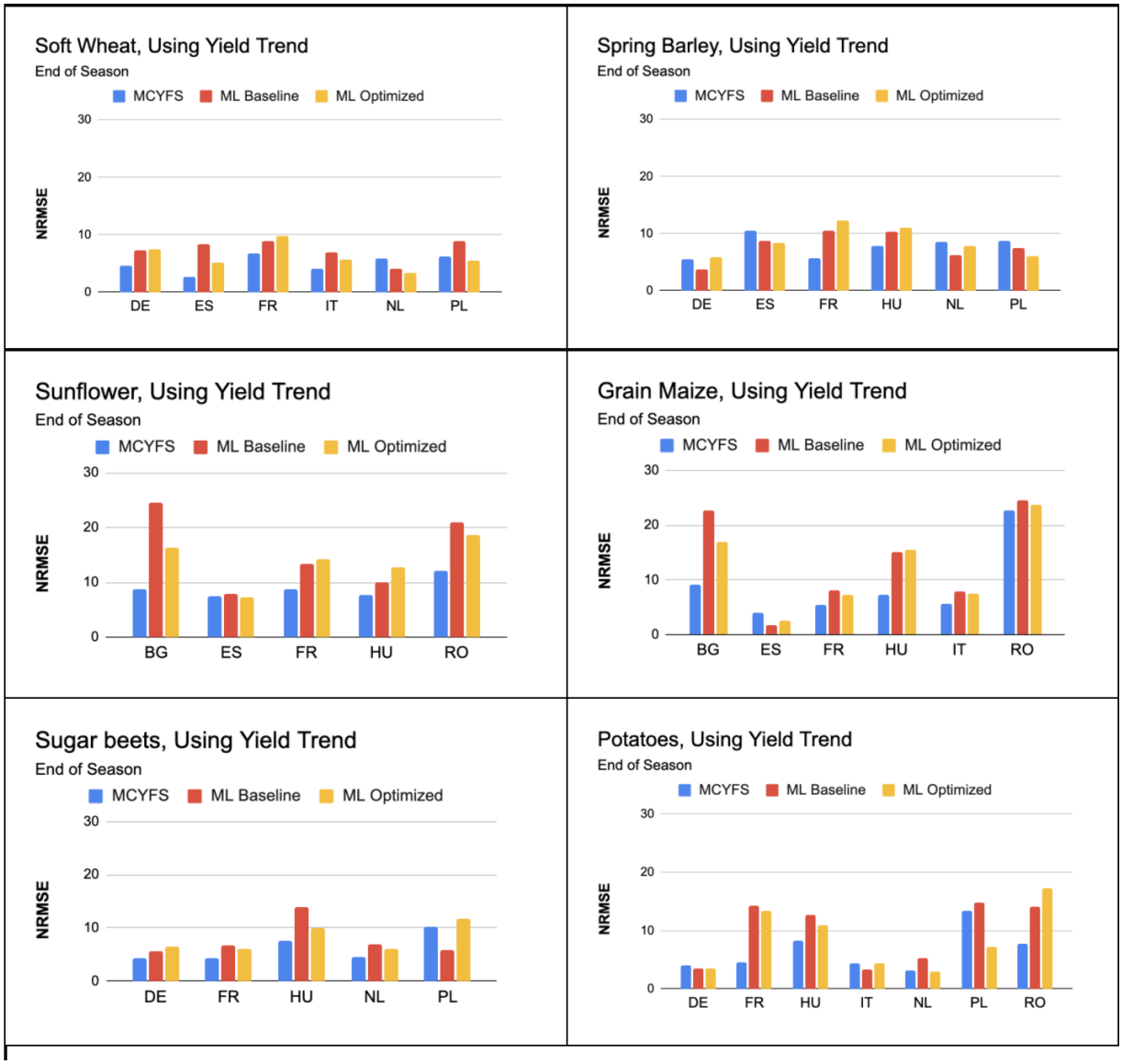


Figure B.7: (b) National Normalized RMSE for end of season. The machine learning baseline and the optimized machine learning models were compared to MCYFS. For both the ML Baseline and the ML Optimized, we show the algorithm with the lowest normalized RMSE.

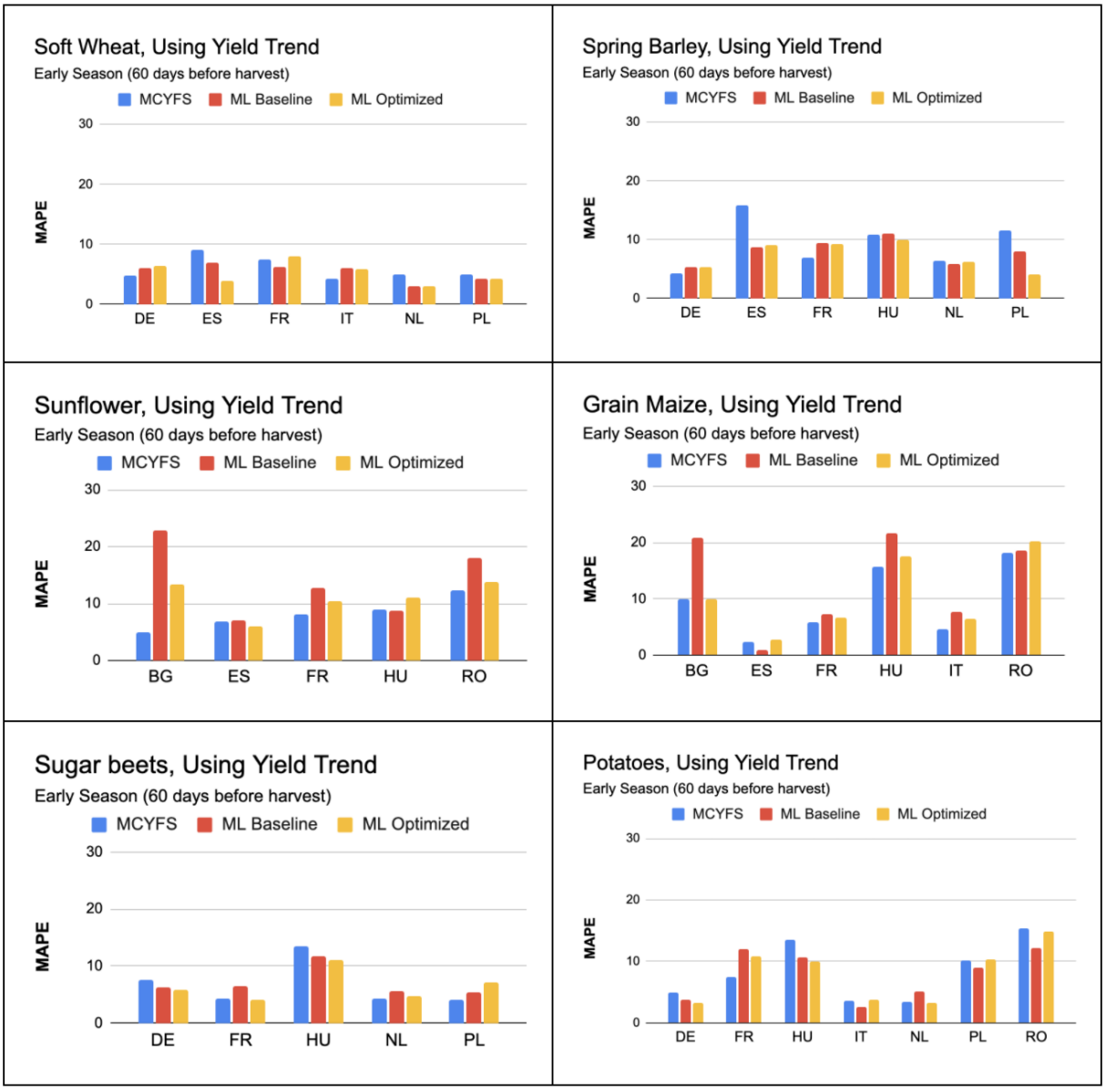


Figure B.8: (a) National MAPE for 60 days before harvest. The machine learning baseline and the optimized machine learning models were compared to MCYFS. For both the ML Baseline and the ML Optimized, we show the algorithm with the lowest normalized RMSE.

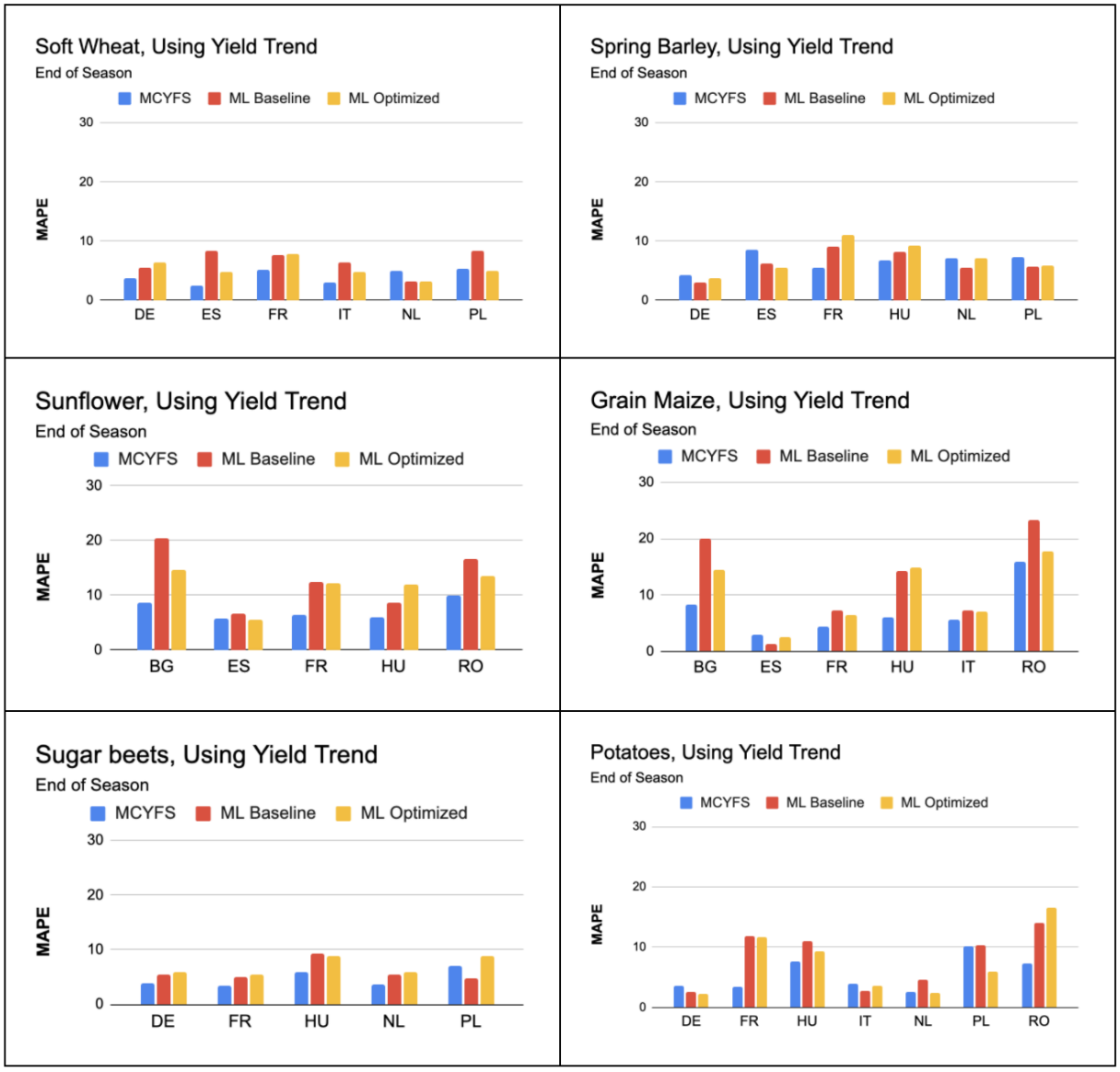


Figure B.8: (b) National MAPE for end of season. The machine learning baseline and the optimized machine learning models were compared to MCYFS. For both the ML Baseline and the ML Optimized, we show the algorithm with the lowest normalized RMSE.

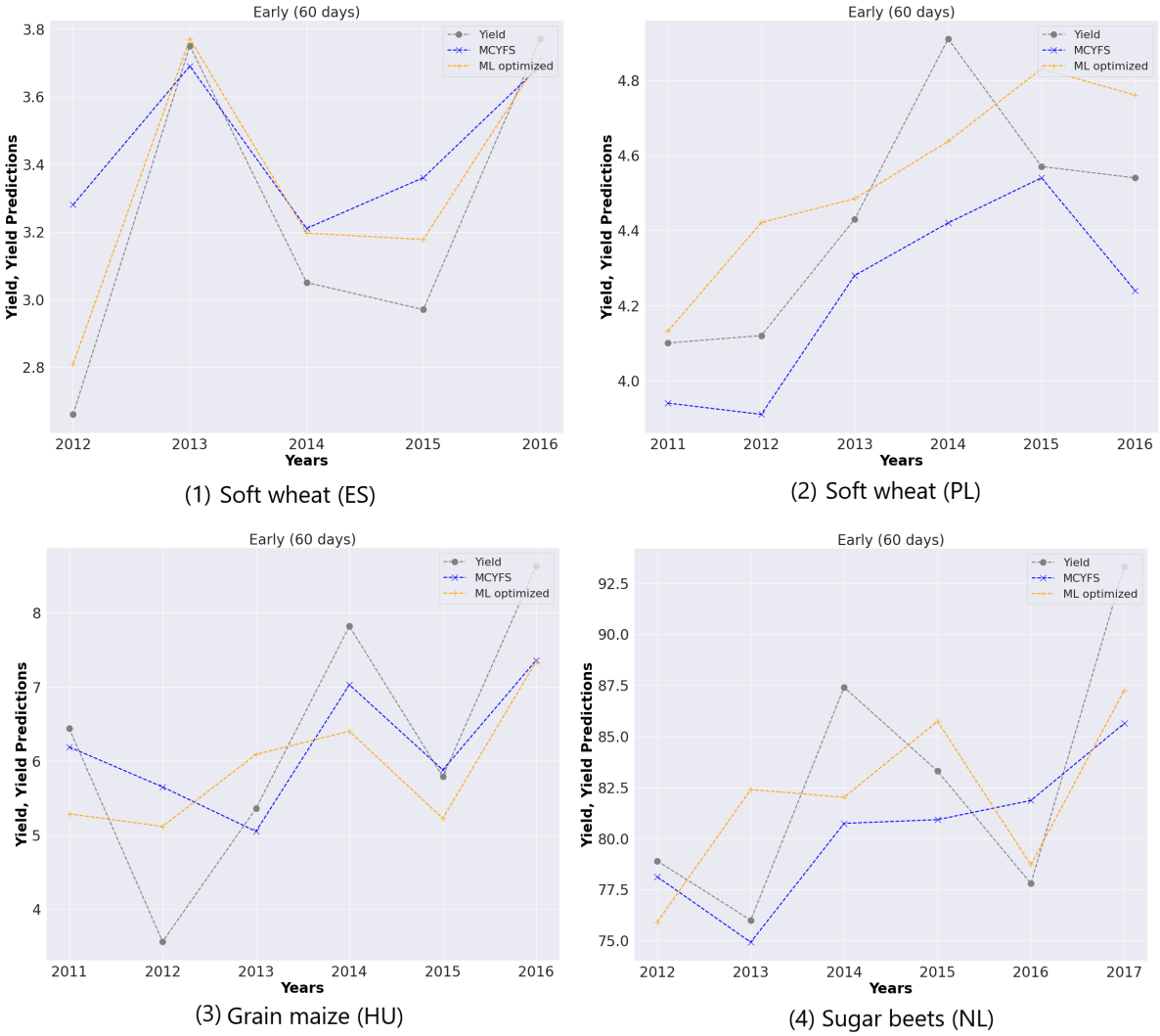


Figure B.9: (a) Reported vs predicted national yields for machine learning and MCYFS. Cases where both MCYFS and machine learning follow the reported yields. Results are for 60 days before harvest.

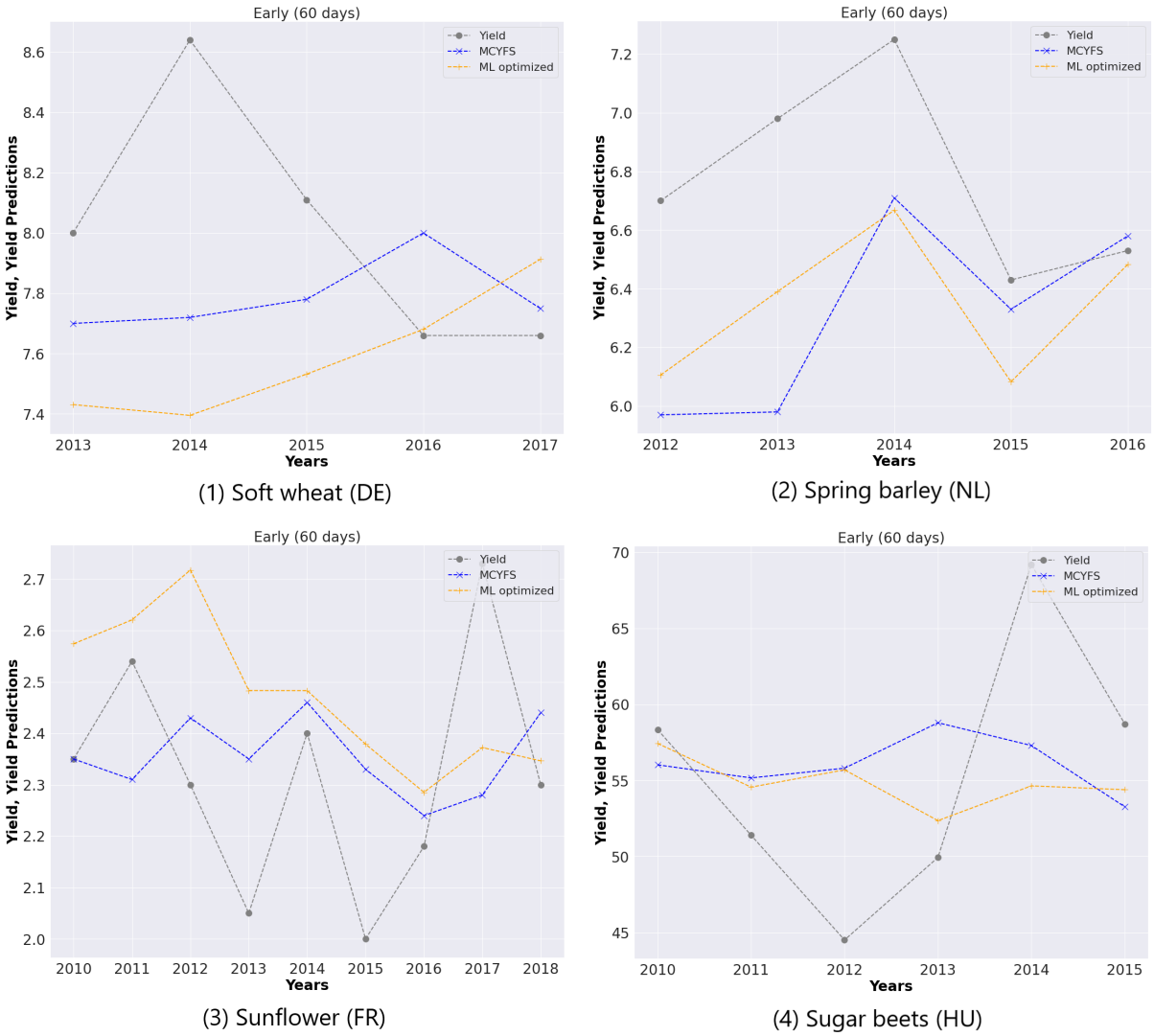


Figure B.9: (b) Reported vs predicted national yields for machine learning and MCYFS. Cases where both MCYFS and machine learning do not capture the yield variability very well. Results are for 60 days before harvest.

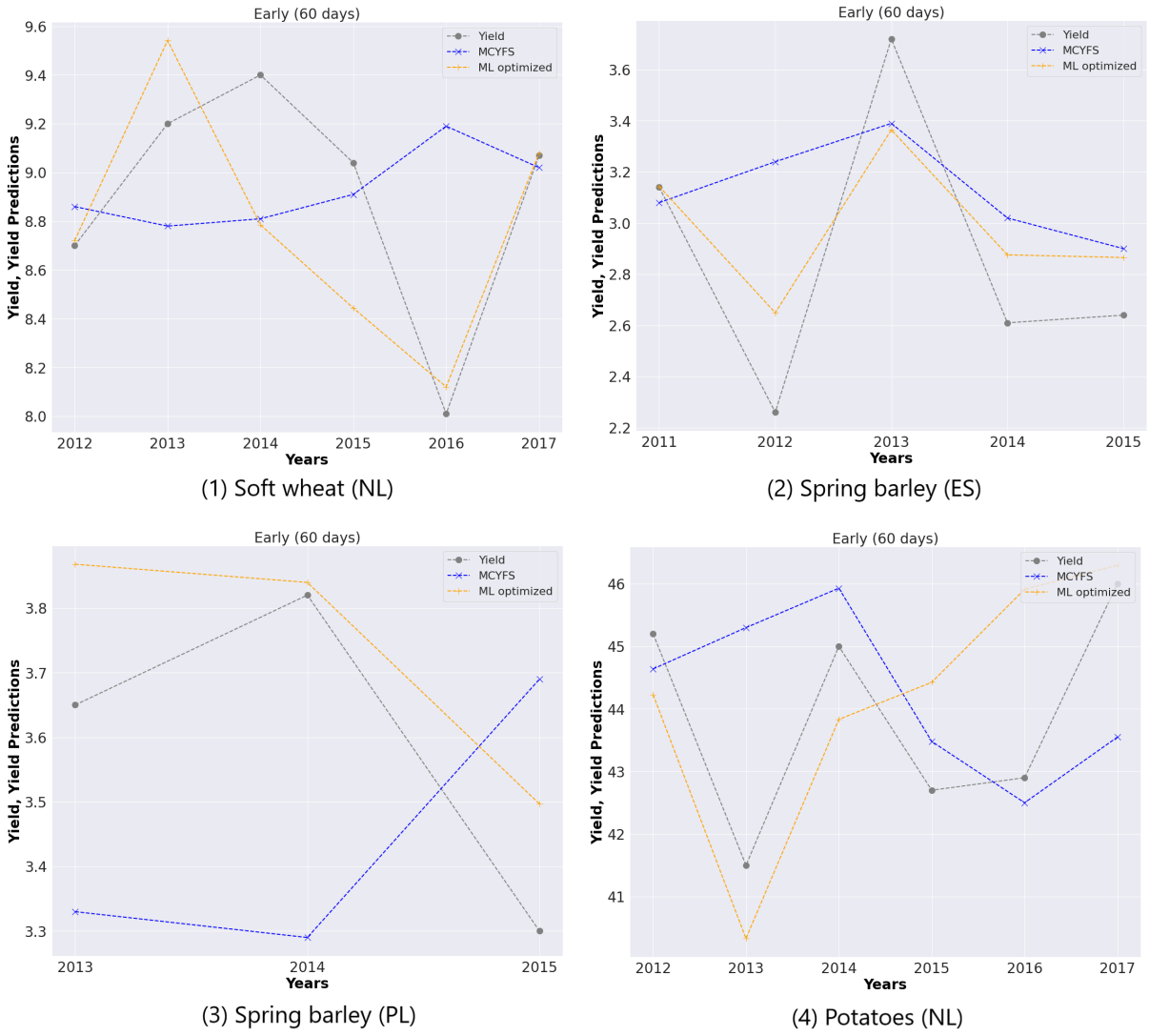


Figure B.9: (c) Reported vs predicted national yields for machine learning and MCYFS. Cases where machine learning performs better than MCYFS. Results are for 60 days before harvest.

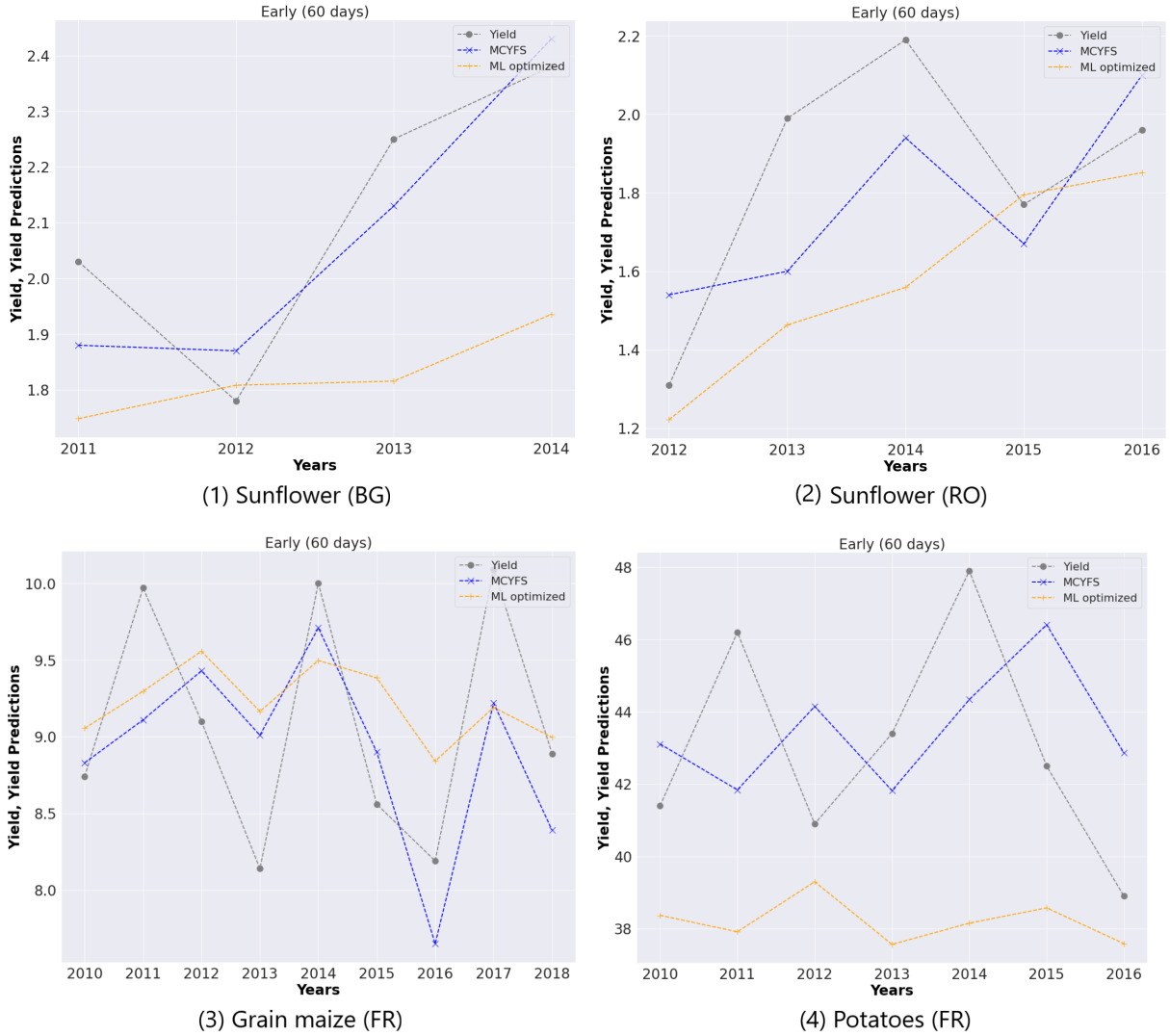


Figure B.9: (d) Reported vs predicted national yields for machine learning and MCYFS. Cases where MCYFS performs better than machine learning. Results are for 60 days before harvest.

Appendix C

C.1 Case studies and crop calendars

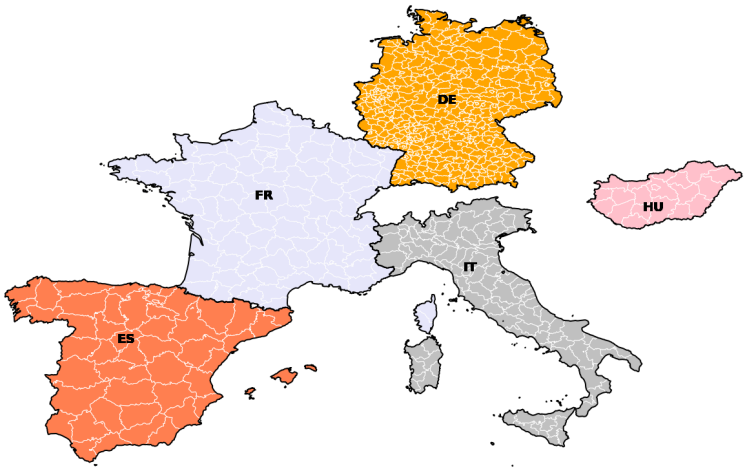
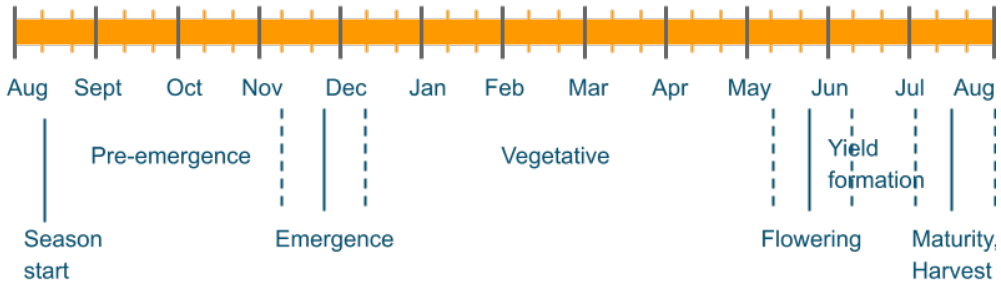
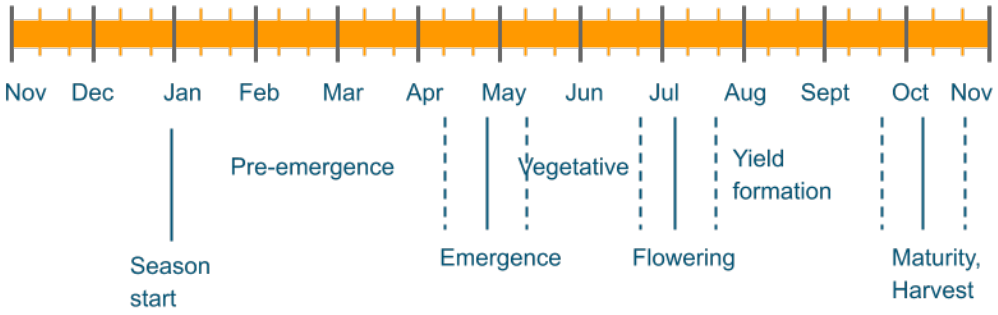


Figure C.1: Selected countries and NUTS regions. Data for soft wheat came from DE, ES, FR and IT, and for grain maize from ES, FR, HU and IT.



(a) Soft wheat, France



(b) Grain maize, France

Figure C.2: Crop calendars for soft wheat and grain maize in France based on the development stages simulated by the WOFOST crop model.

C.2 GBDT Model

Input data and training, validation and test splits were identical between the GBDT model and the neural network models. The GBDT model used in this paper has a small difference compared to Paudel et al. (2022a): agro-environmental zones were added as categorical features.

The GBDT model is based on GradientBoostingRegressor() from scikit-learn (Pedregosa et al., 2011). Hyperparameters including GBDT parameters, feature selector and number of features optimized using BayesSearchCV from scikit-optimize package (Scikit-optimize Contributors, 2021). Feature selectors included a Random Forests model (SelectFromModel) and Recursive FeatureElimination using a Lasso Regression model.

C.3 Deep learning models

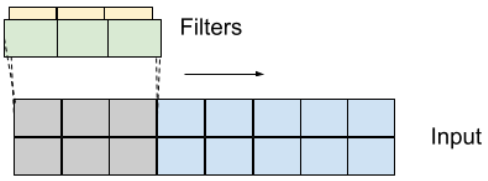


Figure C.3: 1-dimensional convolutional neural network. A set of filters are used to learn different features from input data. Each filter (shown here with width 3) slides through the input to scan the full sequence. The number of steps to slide is called stride.

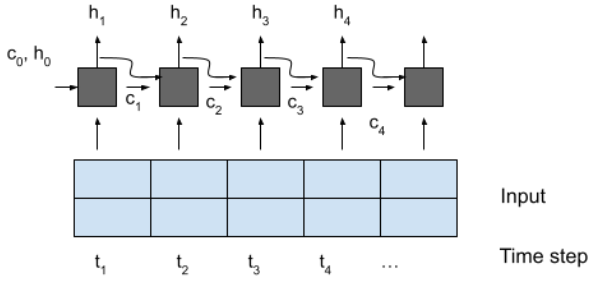


Figure C.4: LSTM network. Each dark gray box is called an LSTM cell. For each time step t_i, h_i represents the current output of the network. The cell state, c_i, acts as the memory of the network and is controlled by gates that update, remove or reset the saved information.

C.3.1 1DCNN Architecture

CNN Layer1 : Conv1d(10, 16, kernel_size=(3,), stride=(1,), padding=(1,))
CNN Layer 2: Conv1d(16, 32, kernel_size=(3,), stride=(2,), padding=(1,))
CNN Layer 3: Conv1d(32, 8, kernel_size=(3,), stride=(2,), padding=(1,))
Batch Normalization, ReLU Activation and Dropout(p=0.1) were added after each CNN layer.
Output Layer: Linear(in_features=83, out_features=2, bias=True)

C.3.2 LSTM Architecture

LSTM Layer : LSTM(10, 64, batch_first=True)
Output Layer : Linear(in_features=83, out_features=2, bias=True)

Both LSTM and 1DCNN were implemented using pytorch (https://pytorch.org/).

C.4 Software implementation

Software implementation of interpretability of deep learning models can be accessed here: https://github.com/BigDataWUR/MLforCropYieldForecasting/tree/dlinterpret

C.5 Validation set results for Architecture Selection

The decision to use per-country data from four countries was based on validation set performance of NUTS3 models for soft wheat and grain maize. We also chose per-country data because agro-climatic differences between countries could influence interpretability of models. In general, LSTM had lower and more stable NRMSEs on the validation set. Therefore, the results reported in the main text are from the LSTM models.

We also experimented with

- LSTM architecture with one network for each seasonal indicator.
- **Interpretable LSTM**: Based on Guo et al. (2019). Uses per indicator hidden states and attention.
- LSTM Autoencoder with one head predicting input time series and another predicting crop yield.

All of these architectures added complexity in terms of number of weights to tune and running time without significant gain in performance.

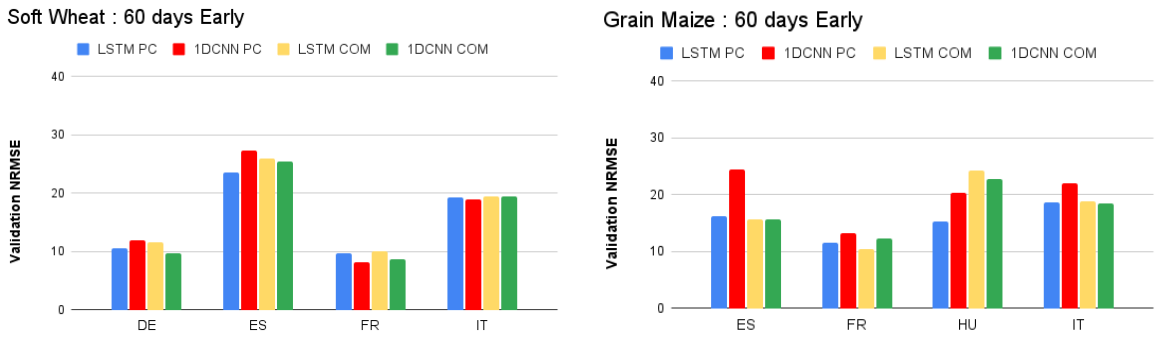


Figure C.5: Validation NRMSEs of NUTS3 forecasts.
PC stands for per-country models; COM stands for the model using data from all countries.

C.6 Supplementary results

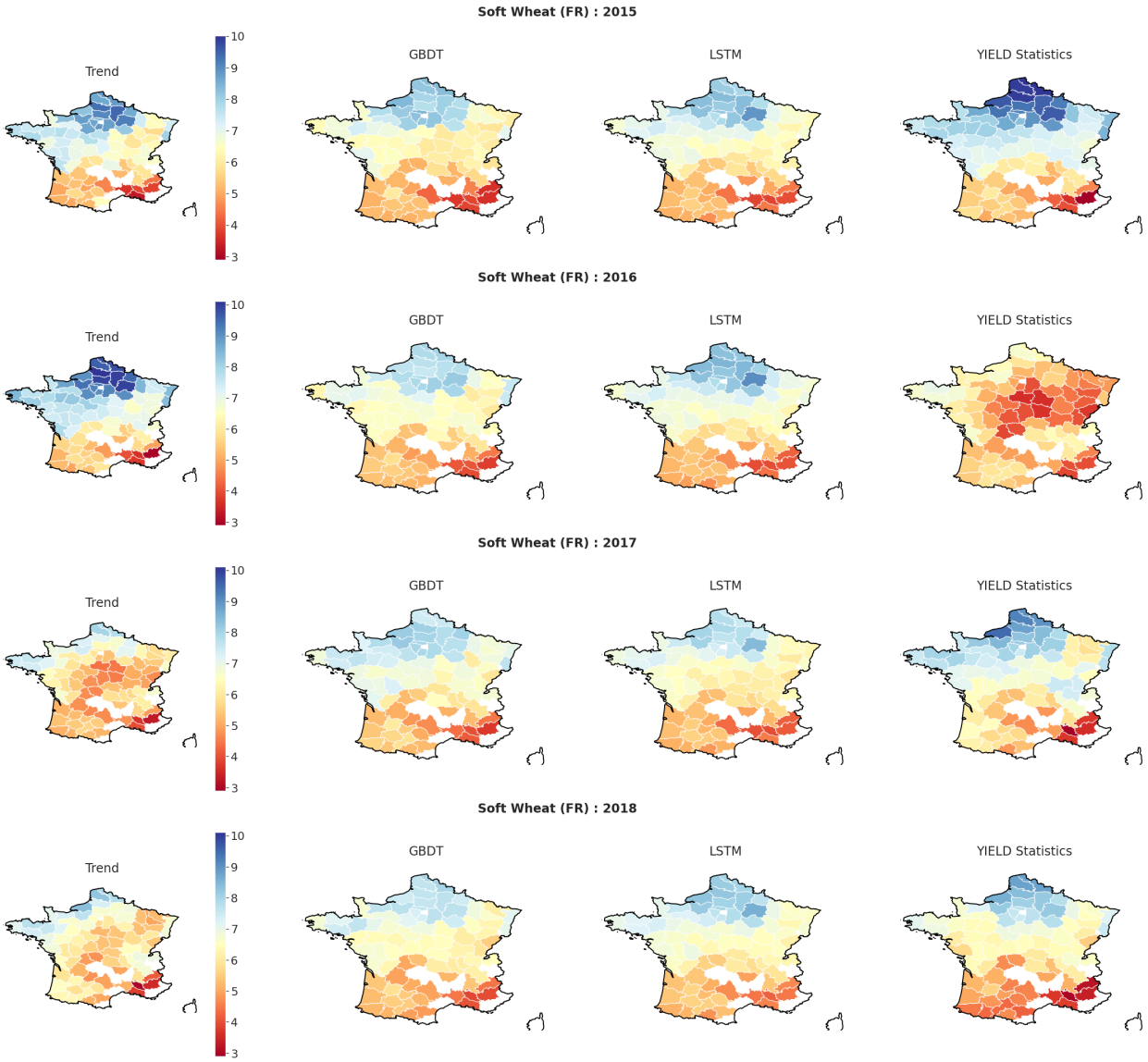


Figure C.6: Predicted vs reported yields for selected test years. (Soft wheat FR, 2015-2018).

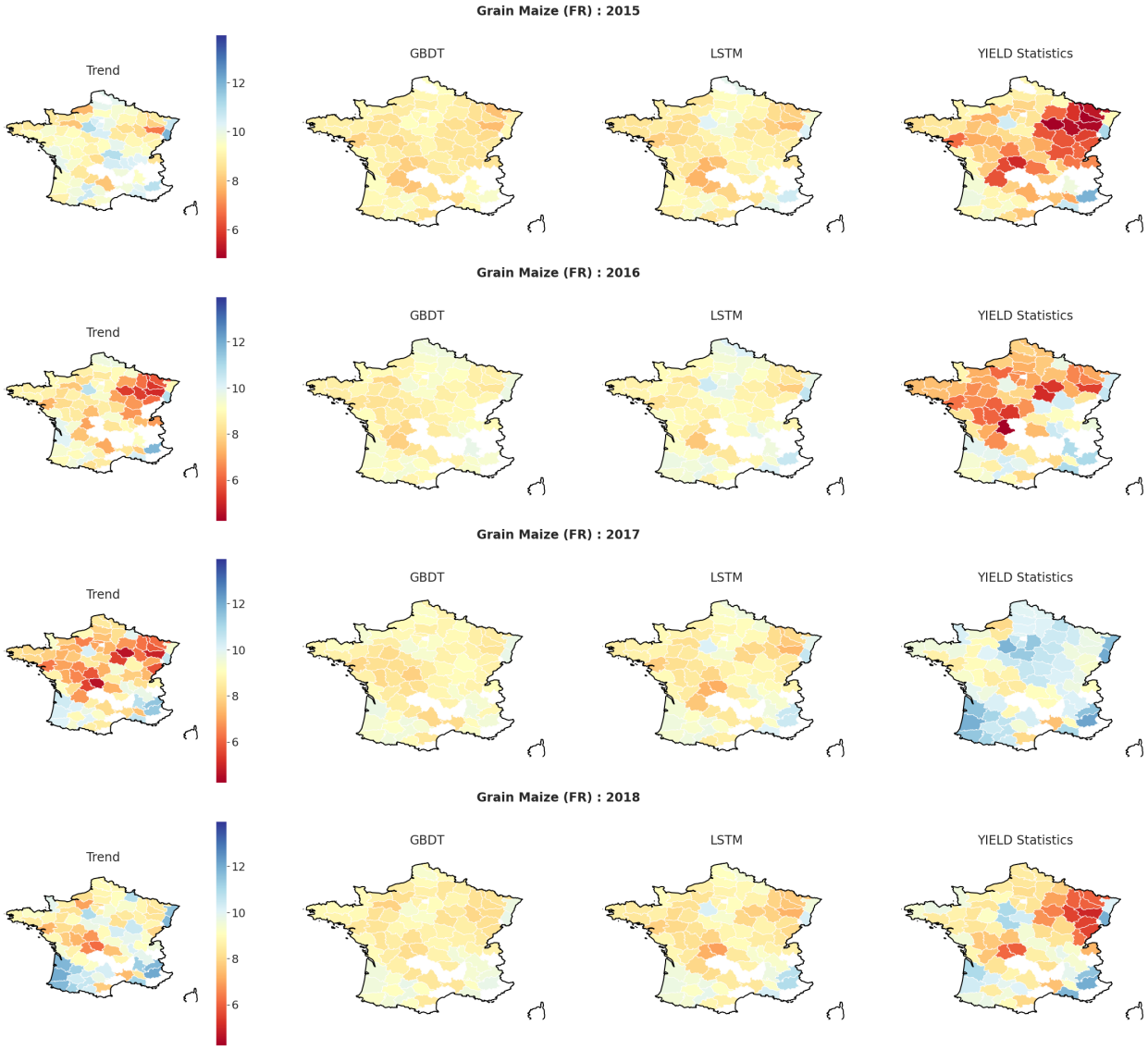


Figure C.7: Predicted vs reported yields for selected test years. (Grain maize FR, 2015-2018).

Table C.1: Summary of Mann-Whitney U test for soft wheat. Trend p-value and GBDT p-value show how similar LSTM forecasts were to trend and GBDT forecasts. The test was run with prediction residuals (predicted yield - reported yield) averaged across 10 models.

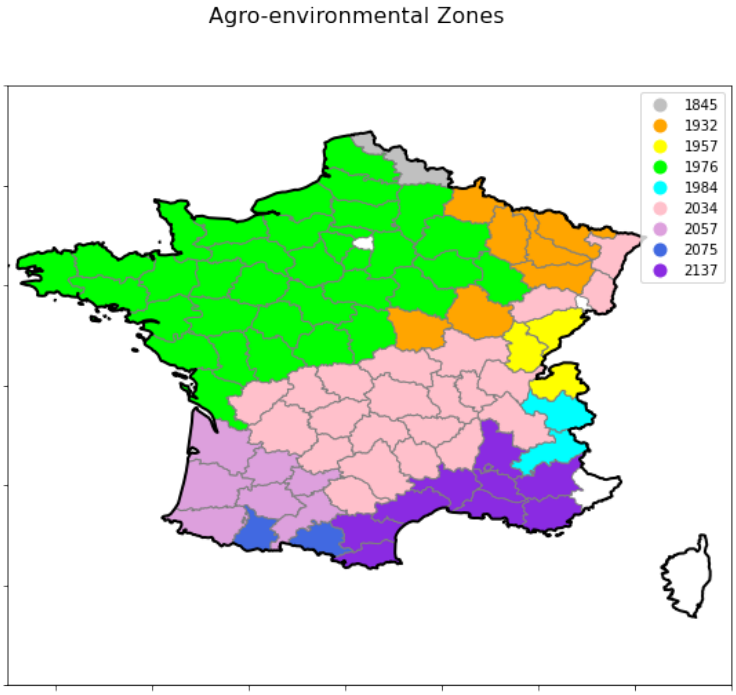
Case study	Trend p-value	GBDT p-value	Trend median	GBDT median	LSTM median
Soft wheat (DE)	0.000	0.001	0.135	-0.460	-0.342
Soft wheat (ES)	0.033	0.397	-0.010	0.181	0.224
Soft wheat (FR)	0.007	0.373	-0.007	-0.138	-0.153
Soft wheat (IT)	0.000	0.710	0.025	-0.224	-0.254

Table C.2: Summary of Mann-Whitney U test for grain maize. Trend p-value and GBDT p-value show how similar LSTM forecasts were to trend and GBDT forecasts. The test was run with prediction residuals (predicted yield - reported yield) averaged across 10 models.

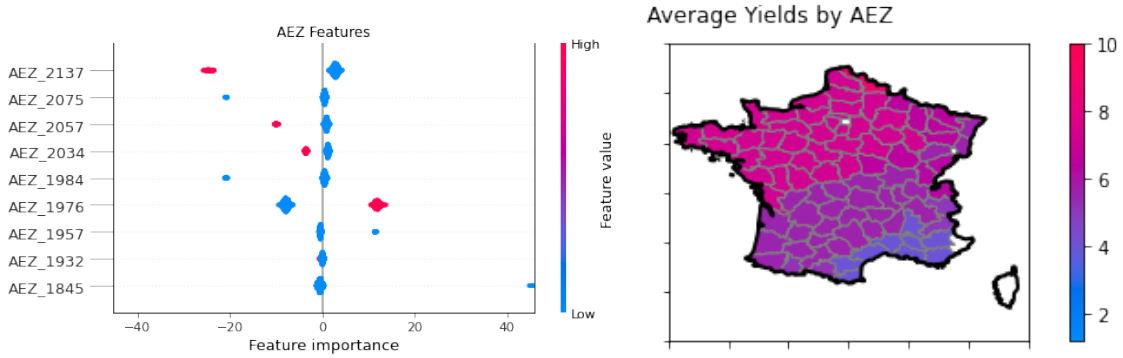
Case study	Trend p-value	GBDT p-value	Trend median	GBDT median	LSTM median
Grain maize (ES)	0.672	0.298	0.000	-0.138	-0.048
Grain maize (FR)	0.396	0.457	-0.020	-0.232	-0.149
Grain maize (HU)	0.996	0.566	-0.330	-0.708	-0.613
Grain maize (IT)	0.788	0.353	-0.040	-0.055	-0.107

Table C.3: Average standard deviation of importance scores across a hundred runs.

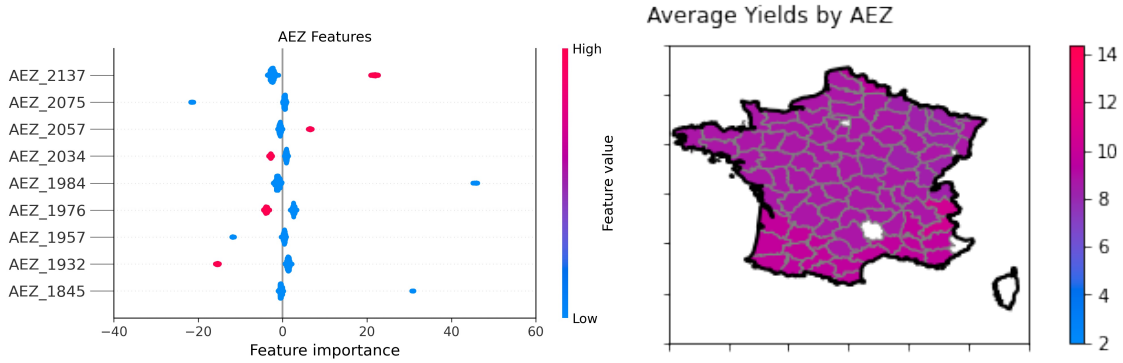
Crop	Occlusion	Integrated Gradients	GradientShap
Soft wheat	Static: 0.078; Trend: 0.154; Seasonal: 0.033	Static: 0.078; Trend: 0.154; Seasonal: 0.021	Static: 0.046; Trend: 0.088; Seasonal: 0.017
Grain maize	Static: 0.064; Trend: 0.150; Seasonal: 0.005	Static: 0.064; Trend: 0.151; Seasonal: 0.001	Static: 0.041; Trend: 0.082; Seasonal: 0.001



(a) Agro-environmental zones

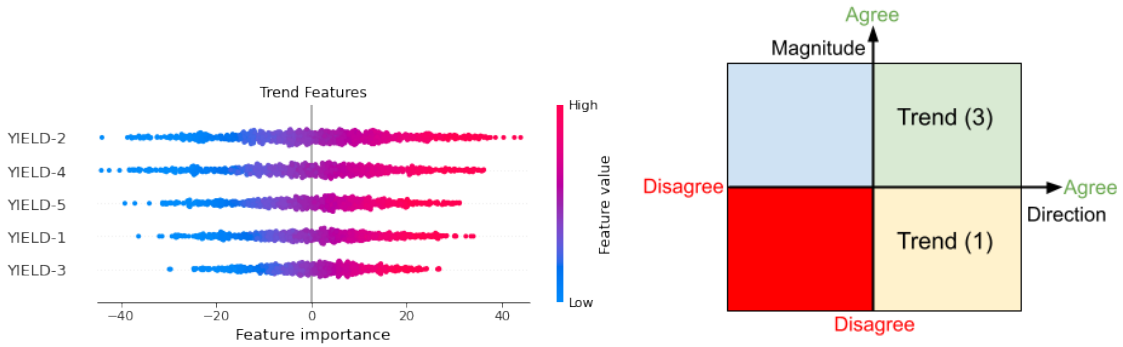


(b) Soft wheat

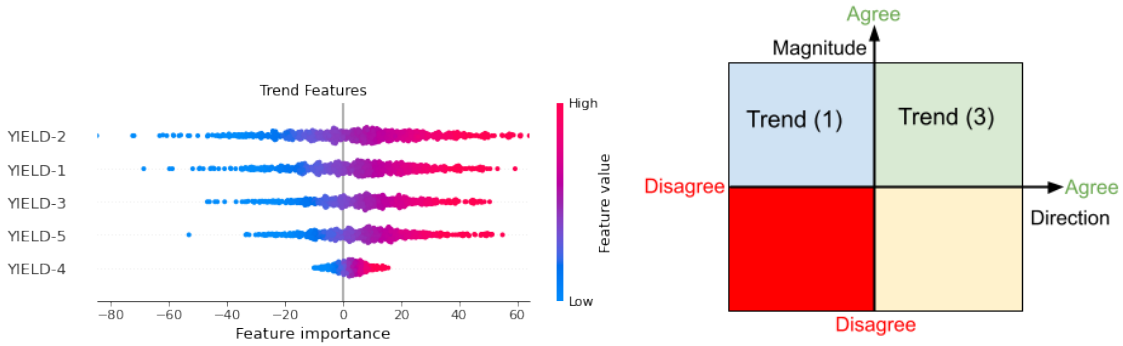


(c) Grain maize

Figure C.8: Importance of agro-environmental zones and average yields. Importance scores show the magnitude and direction (positive or negative) of feature influence on yield. Feature values going low to high (blue to red) from left to right represent a positive influence and vice versa. Importance plots captured the north-south variation in average yields for both crops.

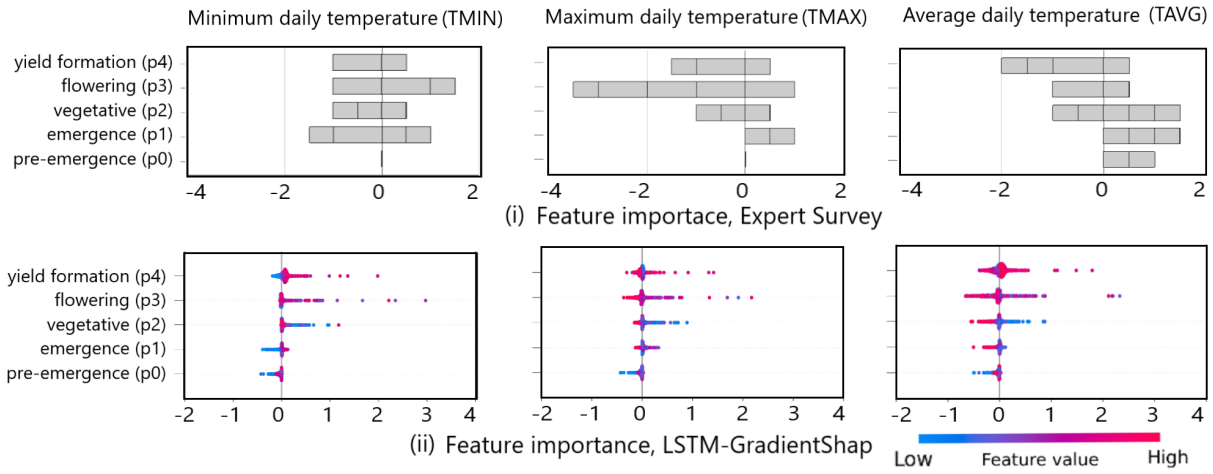


(a) Soft wheat

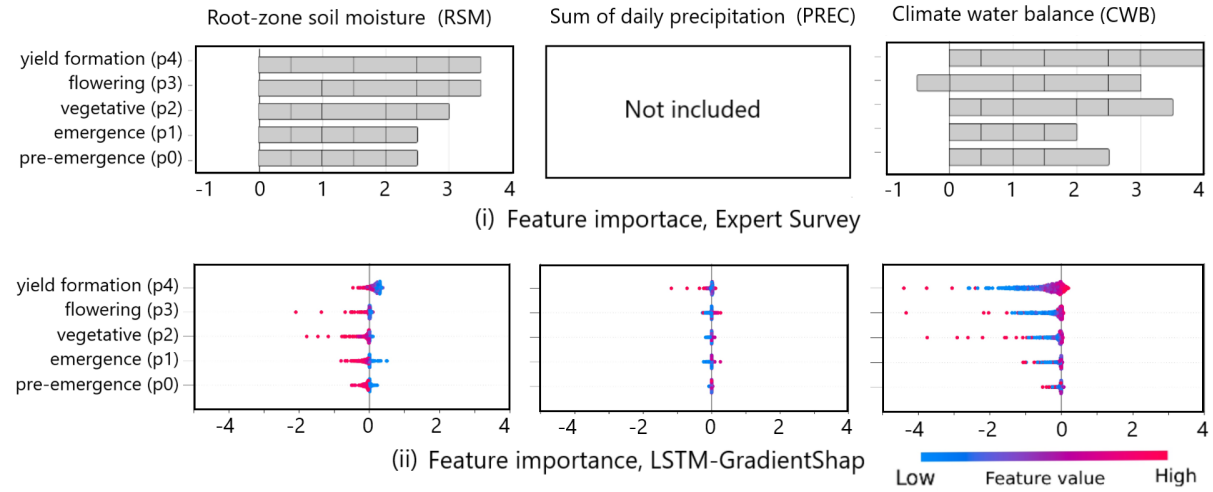


(b) Grain maize

Figure C.9: Importance and interpretability of trend features. Importance scores show the magnitude and direction (positive or negative) of feature influence on yield. Feature values going low to high (blue to red) from left to right represent a positive relation and vice versa. In the interpretability quadrants, the numbers in brackets are the number of experts, and the axes are relations (positive or negative) with yield (Direction) and relative importance (Magnitude).



(a) Temperature features



(b) Moisture features

Figure C.10: Importance of temperature and moisture features for grain maize 60 days before harvest. In (a)(i) and (b)(i), the scale used for the expert survey was: strong negative influence (-1), mild negative influence (-0.5), no influence (0), mild positive influence (0.5), strong positive influence (1). The divisions within each bar represent how different experts voted. Importance scores from deep learning show the magnitude and direction (positive or negative) of feature influence on yield. Feature values going low to high (blue to red) from left to right represent a positive relation and vice versa.

Appendix D

D.1 Data Sources for the US

Most of the input data for 10km grids in the US was acquired from the Climate Data Store of Copernicus Climate Change Service (Copernicus CDS (2022), *Table D.1*). The European Copernicus Programme and the European Centre for Medium-Range Weather Forecast (ECMWF) provide ERA5 (ECMWF Reanalysis Version 5) data (Hersbach et al., 2020), which includes global historical and near real-time weather information. AgERA5 (Boogaard et al., 2022) makes the ERA5 data available for the agricultural domain. Crop productivity indicators, such as total above-ground production and total weight of storage organs, come from de Wit et al. (2022). Their algorithm combines earth observation data of plant light interception, meteorological variables from reanalysis (ERA-Interim or AgERA5) and a simple crop model. The crop model model uses a radiation-use efficiency concept to convert intercepted light into crop biomass and a growing-degree day concept to determine crop phenology and length cropping season. Soil data from WISE30sec database (Batjes, 2016) was processed to extract soil water holding capacity for each 10km grid. World Inventory of Soil Emissions Potentials (WISE) provides homogenized sets of soil property estimates for the whole world.

Table D.1: Data sources summary for the US

Data	**Indicators, Source**
Crop productivity indicators	total above-ground production ($kg\ ha^{-1}$), total weight of storage organs ($kg\ ha^{-1}$), development stage (0-2). **Source**: (de Wit et al., 2022)
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), climate water balance = (PREC - ET0) (mm). **Source**: (Boogaard et al., 2022).
Remote Sensing	Fraction of Absorbed Photosynthetically Active Radiation (Smoothed) (FAPAR). **Source**: (Copernicus GLS, 2020).
Crop Areas	County-level planted areas (ha). **Source**: NASS (USDA-NASS, 2022).
Soil	SM_WHC (water holding capacity). **Source**: WISE Soil Property Database (Batjes, 2016)
County yields	County-level yield statistics ($bushels/acre$). **Source**: NASS (USDA-NASS, 2022).
Grid yields	Grid-level modeled yields ($t\ ha^{-}1$). **Source**: Deines et al. (2021).

D.2 GBDT and LSTM Models

Input data and training, validation and test splits were identical between the GBDT model and the weakly supervised model, except for trend features. Feature design from seasonal indicators was identical to Paudel et al. (2022a). The GBDT model used in this paper has two small differences compared to Paudel et al. (2022a). First, a combined model was built

for four countries. Second, agro-environmental zones and countries were added as categorical features.

The GBDT model is based on GradientBoostingRegressor() from scikit-learn (Pedregosa et al., 2011). Hyperparameters including GBDT parameters, feature selector and number of features optimized using BayesSearchCV from scikit-optimize package (Scikit-optimize Contributors, 2021). Feature selectors included a Random Forests model (SelectFromModel) and Recursive Feature Elimination using a Lasso Regression model.

The LSTM models were trained with the same input data as GBDT models, except that seasonal indicator values were passed to LSTM without feature design.

D.3 Weakly Supervised Model

The weakly supervised model was supervised using NUTS2 crop areas and yields in Europe and county crop areas and yields in the US. The combined loss was the sum of the two losses (crop area loss and yield loss) normalized by the training set average of the corresponding labels. The hyperparameters learning rate and L2-penalty lambda were optimized using custom sliding validation (*Figure 3.3*; Paudel et al. (2022a)). In the US, hyperparameters were optimized using a single validation set with 5 years of data, instead of a 5-fold validation. After optimizing hyperparameters, the model was trained on the entire validation set (no 5-fold) with early stopping: training stopped after validation error increased for two successive epochs. The optimized hyperparameters and early stopping epochs were used to evaluate the model on the test set.

D.3.1 1DCNN Architecture

CNN Layer1 : Conv1d(11, 16, kernel_size=(3,), stride=(1,), padding=(1,))
CNN Layer 2: Conv1d(16, 32, kernel_size=(3,), stride=(2,), padding=(1,))
CNN Layer 3: Conv1d(32, 8, kernel_size=(3,), stride=(2,), padding=(1,))
Batch Normalization, ReLU Activation and Dropout(p=0.1) were added after each CNN layer.
Output Layer: Linear(in_features=100, out_features=2, bias=True)

D.3.2 LSTM Architecture

LSTM Layer : LSTM(11, 64, batch_first=True)
Output Layer : Linear(in_features=100, out_features=2, bias=True)

Both LSTM and 1DCNN were implemented using pytorch (Paszke et al., 2019).

D.4 Software and data availability

Software implementation of the weakly supervised framework can be accessed here:
https://github.com/BigDataWUR/MLforCropYieldForecasting/tree/weaksup

Sample data for the US is available at DOI: https://doi.org/10.5281/zenodo.7751191

D.5 Validation set results for Architecture Selection

The decision to use combined data from four countries was based on validation set performance of strongly supervised NUTS3 models for soft wheat. Since NRMSEs were similar for both cases, we chose to use combined data because the larger data size would limit overfitting issues. CV comparisons for weak supervision using 1DCNN and LSTM showed that LSTM had lower NRMSEs on the validation set (*Figure D.1*).

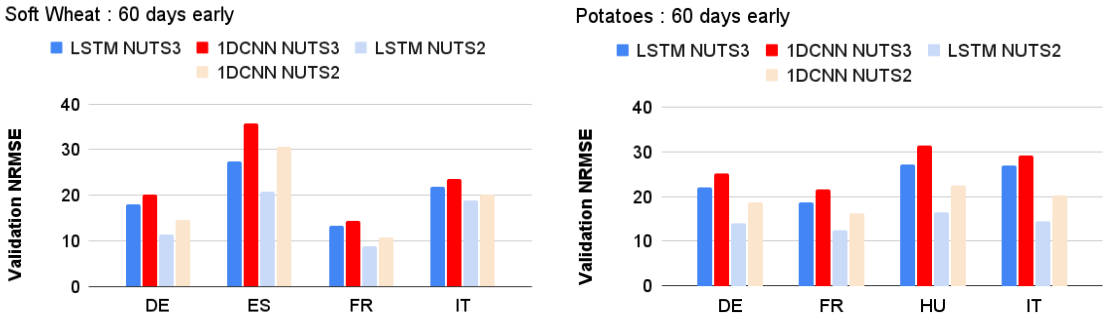


Figure D.1: Validation NRMSEs of weakly supervised NUTS3 and NUTS2 forecasts.

D.6 Supplementary Results

Table D.2: Mann-Whitney test results for NUTS2 forecasts: soft wheat (Europe). Mann-Whitney U test was run using prediction residuals (predicted yield - reported yield).

	LR Trend	LR GBDT	LR LSTM	WS Trend
LR Trend	-			
LR GBDT	0.3046	-		
LR LSTM	0.1302	0.5524	-	
WS Trend	0.1424	0.6471	0.8534	-
Median	-0.005	-0.0922	-0.117	-0.076

Table D.3: Mann-Whitney test results for NUTS2 forecasts: potatoes (Europe). Mann-Whitney U test was run using prediction residuals (predicted yield - reported yield).

	LR Trend	LR GBDT	LR LSTM	WS Trend
LR Trend	-			
LR GBDT	0.0037	-		
LR LSTM	0.0128	0.7025	-	
WS Trend	0.002	0.8471	0.5274	-
Median	-0.1	0.5297	0.6854	0.5547

Table D.4: Mann-Whitney test results for county forecasts: corn (US). Mann-Whitney U test was run using prediction residuals (predicted yield - reported yield).

	LR Trend	LR GBDT	LR LSTM	WS Trend
LR Trend	-			
LR GBDT	0.0	-		
LR LSTM	0.0	0.0	-	
WS Trend	0.0	0.0	0.0	-
Median	0.371	-0.956	-0.559	-0.227

Table D.5: Mann-Whitney test results for NUTS3 forecasts: soft wheat (Europe). Mann-Whitney U test was run using prediction residuals (predicted yield - reported yield).

	Naive Trend	Naive GBDT	Naive LSTM	WS Trend	HR Trend	HR GBDT	HR LSTM
Naive Trend	-						
Naive GBDT	0.0001	-					
Naive LSTM	0.0	0.8805	-				
WS Trend	0.0	0.6133	0.731	-			
HR Trend	0.369	0.0	0.0	0.0	-		
HR GBDT	0.0006	0.2797	0.2116	0.1202	0.0	-	
HR LSTM	0.0	0.0741	0.1007	0.2117	0.0	0.0015	-
Median	0.0	-0.157	-0.152	-0.161	0.01	-0.121	-0.237

Table D.6: Mann-Whitney test results for NUTS3 forecasts: potatoes (Europe). Mann-Whitney U test was run using prediction residuals (predicted yield - reported yield).

	Naive Trend	Naive GBDT	Naive LSTM	WS Trend	HR Trend	HR GBDT	HR LSTM
Naive Trend	-						
Naive GBDT	0.0737	-					
Naive LSTM	0.3994	0.3428	-				
WS Trend	0.0228	0.5686	0.1266	-			
HR Trend	0.3523	0.251	0.8644	0.0756	-		
HR GBDT	0.067	0.7892	0.4815	0.3103	0.2055	-	
HR LSTM	0.0269	0.9417	0.2854	0.515	0.0989	0.6934	-
Median	-0.28	0.3048	0.0236	0.3588	-0.005	0.1835	0.2123

Table D.7: Mann-Whitney test results for 10-km grids forecasts: corn (US). Mann-Whitney U test was run using prediction residuals (predicted yield - reported yield).

	Naive Trend	Naive GBDT	Naive LSTM	WS Trend	HR Trend	HR GBDT	HR LSTM
Naive Trend	-						
Naive GBDT	0.0	-					
Naive LSTM	0.0	0.0	-				
WS Trend	0.0	0.0	0.0	-			
HR Trend	0.0	0.0	0.0	0.0	-		
HR GBDT	0.0	0.0004	0.0	0.0	0.0	-	
HR LSTM	0.0	0.0	0.1955	0.0	0.0	0.0	-
Median	0.0	-1.756	-1.031	-0.621	0.47	-1.755	-1.014

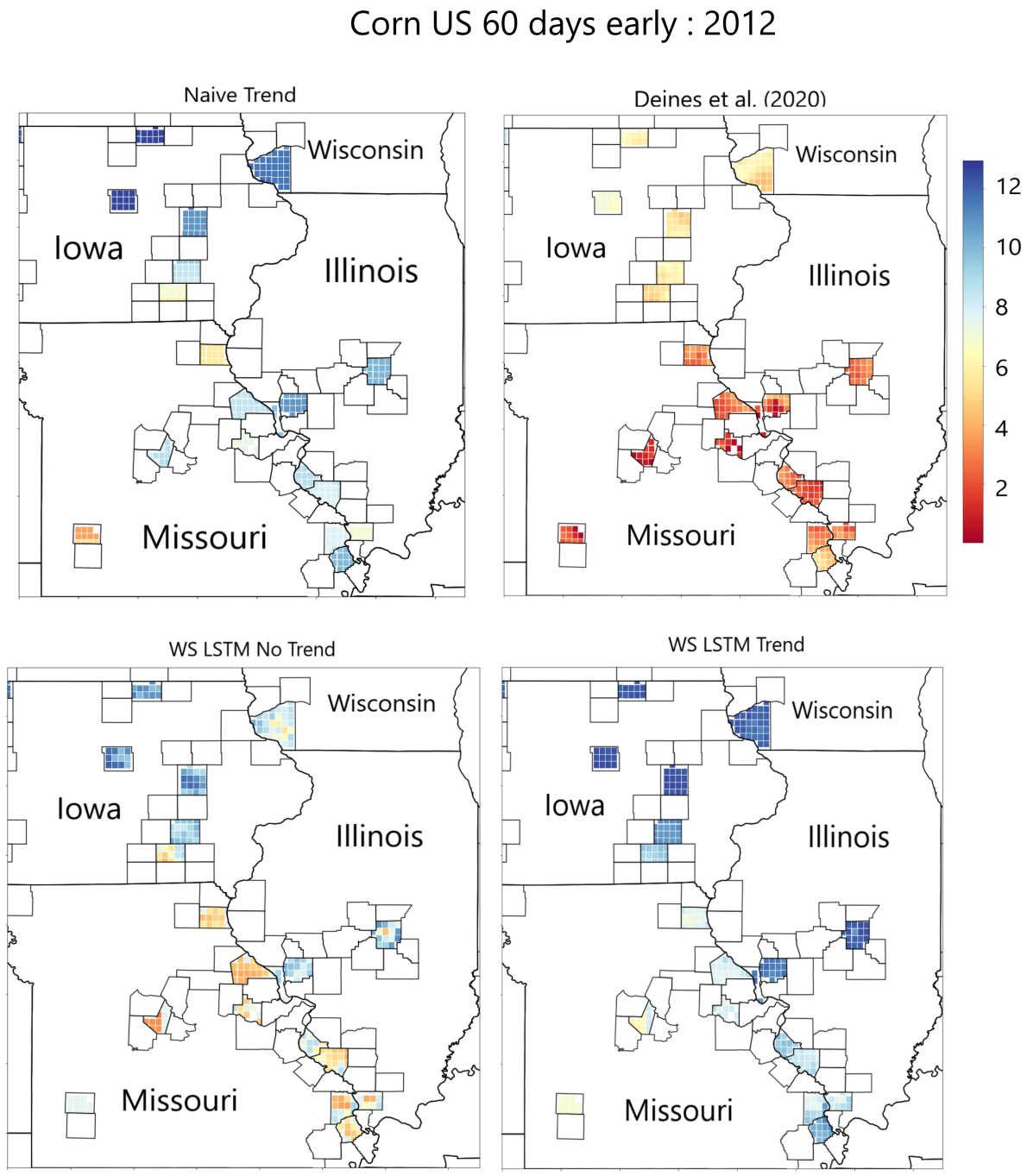


Figure D.2: Spatial variability of corn yield forecasts for 2012. The WS model with county yield produced more accurate results, but shows less variability within counties.

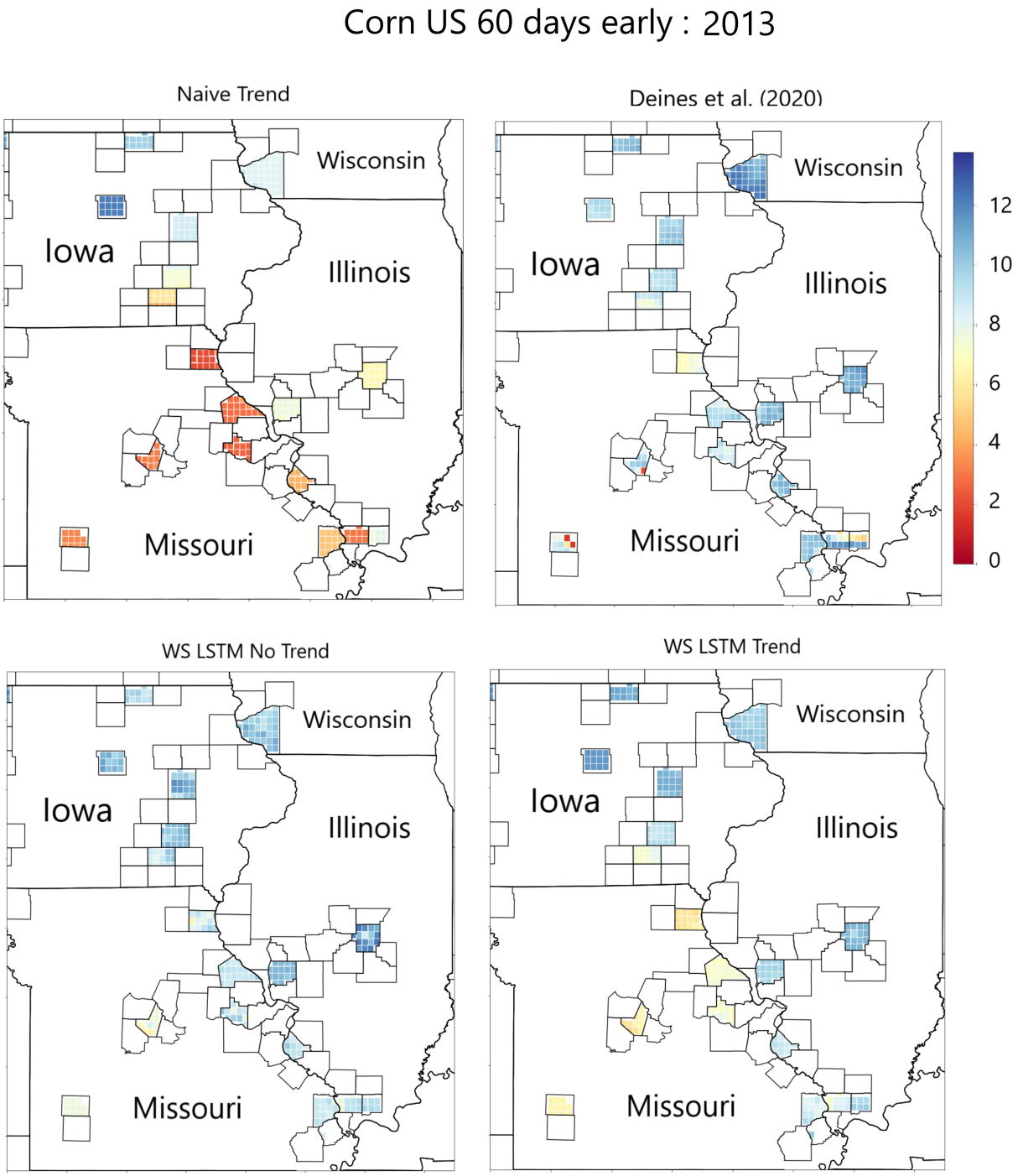


Figure D.3: Spatial variability of corn yield forecasts for 2013. The WS model with county yield produced more accurate results, but shows less variability within counties.

References

David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991. doi:10.1007/BF00153759.

Woubet G Alemu and Geoffrey M Henebry. 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 Sensing*, 8(12):1016, 2016. doi:10.3390/rs8121016.

Richard G Allen, Luis S Pereira, Dirk Raes, Martin Smith, et al. Crop evapotranspiration – Guidelines for computing crop water requirements. In *Irrigation and Drainage, paper 56*. FAO, Rome, 1998. ISBN 92-5-104219-5.

M. Ancona, E. Ceolini, C. Oztireli, and M Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations*. ICLR 2018, 2018. doi:10.48550/arXiv.1711.06104.

Javad Ansarifar, Lizhi Wang, and Sotirios V Archontoulis. An interaction regression model for crop yield prediction. *Scientific reports*, 11(1):17754, 2021. doi:10.1038/s41598-021-97221-7.

Junaid Bajwa, Usman Munir, Aditya Nori, and Bryan Williams. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthcare Journal*, 8(2):e188, 2021. doi:10.7861/fhj.2021-0095.

Hannah Barrett and David Christian Rose. Perceptions of the fourth agricultural revolution: What's in, what's out, and what consequences are anticipated? *Sociologia Ruralis*, 62(2): 162–189, 2022. doi:10.1111/soru.12324.

Etienne Bartholome and Allan S Belward. GLC2000: a new approach to global land cover mapping from Earth observation data. *International Journal of Remote Sensing*, 26(9): 1959–1977, 2005. doi:10.1080/01431160412331291297.

Bettina Baruth, A Royer, Anja Klisch, and G Genovese. The use of remote sensing within the MARS crop yield monitoring system of the European Commission. In *Proceedings of the 21st Congress of the International Society for Photogrammetry and Remote Sensing. Vol. 37 Part B8*, pp. 935–940, 2008. `https://www.isprs.org/proceedings/XXXVII/congress/8_pdf/10_WG-VIII-10/02.pdf`, Last accessed: July 27, 2022.

Igor I Baskin, Gilles Marcou, Dragos Horvath, and Alexandre Varnek. Benchmarking machine-learning methods. *Tutorials in Chemoinformatics*, pp. 209–222, 2017. doi:10.1002/9781119161110.ch13.

Bruno Basso and Lin Liu. Seasonal crop yield forecast: Methods, applications, and accuracies. In *Advances in Agronomy*, volume 154, pp. 201–255. Elsevier, 2019. doi:10.1016/bs.agron.2018.11.002.

Bruno Basso, Davide Cammarano, and Elisabetta Carfagna. Review of crop yield forecasting methods and early warning systems. In *Report presented to the First Meeting of the Scientific Advisory Committee of the Gloal Strategy to Improve Agricultural and Rural Statistics*. FAO Headquarters, Rome, Italy, 18–19 July, 2013.

Niels H Batjes. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. *Geoderma*, 269:61–68, 2016. doi:10.1016/j.geoderma.2016.01.034.

Jan Behmann, Anne-Katrin Mahlein, Till Rumpf, Christoph Römer, and Lutz Plümer. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision Agriculture*, 16:239–260, 2015. doi:10.1007/s11119-014-9372-7.

Tamara Ben-Ari, Julien Boé, Philippe Ciais, Remi Lecerf, Marijn Van der Velde, and David Makowski. Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nature Communications*, 9(1):1–10, 2018. doi:10.1038/s41467-018-04087-x.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, pp. 1–4. Springer, 2009. `https://link.springer.com/content/pdf/10.1007/978-3-642-00296-0_5.pdf`, Last accessed: May 11, 2020.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013. doi:10.1109/TPAMI.2013.50.

Lefteris Benos, Aristotelis C Tagarakis, Georgios Dolias, Remigio Berruto, Dimitrios Kateris, and Dionysis Bochtis. Machine learning in agriculture: A comprehensive updated review. *Sensors*, 21(11):3758, 2021. doi:10.3390/s21113758.

James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2), 2012. `https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf`, Last accessed: May 20, 2021.

H. Boogaard, J. Schubert, A. De Wit, J. Lazebnik, R. Hutjes, and G. Van der Grijn. Agrometeorological indicators from 1979 to present derived from reanalysis. Climate Data Store - Copernicus Climate Change Service, `https://doi.org/10.24381/cds.6c68c9bb`, 2022.

Claire Boryan, Zhengwei Yang, Rick Mueller, and Mike Craig. Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto International*, 26(5):341–358, 2011. doi:10.1080/10106049.2011.562309.

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

Azzedine Boukerche, Lining Zheng, and Omar Alfandi. Outlier Detection: Methods, Models, and Classification. *ACM Computing Surveys*, 53(3):1–37, 2020. doi:10.1145/3381028.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Nadine Brisson, Christian Gary, Eric Justes, Romain Roche, Bruno Mary, Dominique Ripoche, Daniel Zimmer, Jorge Sierra, Patrick Bertuzzi, Philippe Burger, et al. An overview of the crop model STICS. *European Journal of agronomy*, 18(3-4):309–332, 2003. doi:10.1016/S1161-0301(02)00110-7.

Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010. doi:arXiv:1012.2599.

DJ Brus, H Boogaard, T Ceccarelli, TG Orton, S Traore, and M Zhang. Geostatistical disaggregation of polygon maps of average crop yields by area-to-point kriging. *European Journal of Agronomy*, 97:48–59, 2018.

Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998. doi:10.1023/A:1009715923555.

Attila Bussay, Marijn van der Velde, Davide Fumagalli, and Lorenzo Seguini. Improving operational maize yield forecasting in hungary. *Agricultural Systems*, 141:94–106, 2015. doi:10.1016/j.agsy.2015.10.001.

George Büttner, Jan Feranec, Gabriel Jaffrain, László Mari, Gergely Maucha, and Tomas Soukup. The CORINE land cover 2000 project. *EARSeL eProceedings*, 3 (3):331–346, 2004. `http://eproceedings.uni-oldenburg.de/website/vol03_3/03_3_buttner2.pdf`, Last accessed: May 18, 2021.

Yaping Cai, Kaiyu Guan, David Lobell, Andries B Potgieter, Shaowen Wang, Jian Peng, Tianfang Xu, Senthold Asseng, Yongguang Zhang, Liangzhi You, et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and forest meteorology*, 274:144–159, 2019. doi:10.1016/j.agrformet.2019.03.010.

Yiqing Cai, Kristen Moore, Adam Pellegrini, Aymn Elhaddad, Jerrod Lessel, Christianna Townsend, Hayley Solak, and Nemo Semret. Crop yield predictions-high resolution statistical model for intra-season forecasts applied to corn in the US. In *2017 Fall Meeting*. Gro Intelligence Inc., 2017.

Pierre Cantelaube and Jean-Michel Terres. Seasonal weather forecasts for crop yield modelling in europe. *Tellus A: Dynamic Meteorology and Oceanography*, 57(3):476–487, 2005. doi:10.3402/tellusa.v57i3.14669.

Juan Cao, Zhao Zhang, Yuchuan Luo, Liangliang Zhang, Jing Zhang, Ziyue Li, and Fulu Tao. Wheat yield predictions at a county and field scale with deep learning, machine learning, and Google Earth Engine. *European Journal of Agronomy*, 123:126204, 2021. doi:10.1016/j.eja.2020.126204.

Andrej Ceglar, Andrea Toreti, Rémi Lecerf, Marijn Van der Velde, and Frank Dentener. Impact of meteorological drivers on regional inter-annual crop yield variability in france. *Agricultural and Forest Meteorology*, 216:58–67, 2016. doi:10.1016/j.agrformet.2015.10.004.

I. Cerrani and R. López Lozano. Algorithm for the dissaggregation of crop area statistics in the MARS crop yield forecasting system, 2017. `https://agri4cast.jrc.ec.europa.eu/DataPortal/Resource_Files/PDF_Documents/31_rationale.pdf`, Last accessed: Oct 8, 2020.

I. Cerrani and R. López Lozano. Algorithm for the dissaggregation of crop area statistics in the MARS crop yield forecasting system - update 2022, 2022. `https://agri4cast.jrc.ec.europa.eu/DataPortal/Resource_Files/PDF_Documents/31_rationale.pdf`, Last accessed: Jan 18, 2023.

Andrew Challinor. Towards the development of adaptation options using climate and crop yield forecasting at seasonal to multi-decadal timescales. *Environmental Science & Policy*, 12(4):453–465, 2009. doi:10.1016/j.envsci.2008.09.008.

Akshay L Chandra, Sai Vikas Desai, Vineeth N Balasubramanian, Seishi Ninomiya, and Wei Guo. Active learning with point supervision for cost-effective panicle detection in cereal crops. *Plant Methods*, 16(1):1–16, 2020. doi:10.1186/s13007-020-00575-8.

Aston Chipanshi, Yinsuo Zhang, Louis Kouadio, Nathaniel Newlands, Andrew Davidson, Harvey Hill, Richard Warren, Budong Qian, Bahram Daneshfar, Frederic Bedard, et al. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the canadian agricultural landscape. *Agricultural and Forest Meteorology*, 206:137–150, 2015. doi:10.1016/j.agrformet.2015.03.007.

Ringson J Chitsiko, Onisimo Mutanga, Timothy Dube, and Dumisani Kutywayo. Review of the current models and approaches used for maize crop yield forecasting in sub-saharan africa, and their potential use in early warning systems. *Physics and Chemistry of the Earth, Parts A/B/C*, pp. 103199, 2022. doi:10.1016/j.pce.2022.103199.

Anna Chlingaryan, Salah Sukkarieh, and Brett Whelan. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151:61–69, 2018. doi:10.1016/j.compag.2018.05.012.

Hee Chang Chung, Dong Il Kim, and Ae Kyung Moon. Overview of Smart Farming based on networks. In *Proceedings of the Korean Institute of Information and Commucation Sciences Conference*, pp. 617–618, 2015. `https://koreascience.kr/article/CFKO201531751945980.page`, Last accessed: July 27, 2022.

Sun-Ok Chung, Moon-Chan Choi, Kyu-Ho Lee, Yong-Joo Kim, Soon-Jung Hong, and Minzan Li. Sensing technologies for grain crop yield monitoring systems: A review. *Journal of Biosystems Engineering*, 41(4):408–417, 2016. doi:10.5307/JBE.2016.41.4.408.

Aidan Connolly. 8 Disruptive Digital Technologies... with the Power to Transform Agriculture, 2016. https://www.linkedin.com/pulse/disruptive-digital-technologies-power-transform-aidan-connolly-7k-?trk=mp-author-card, Last accessed: Feb 24, 2023.

Copernicus CDS. Copernicus Climate Data Store. Copernicus Climate Change Service, https://cds.climate.copernicus.eu/, 2022.

Copernicus ESA. Sentinel Earth Observation Data. Copernicus Open Access Hub, https://scihub.copernicus.eu/, 2022.

Copernicus GLS. Fraction of Absorbed Photosynthetically Active Radiation. Copernicus Global Land Service, https://doi.org10.24381/cds.7e59b01a, 2020.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3): 273–297, 1995. doi:10.1007/BF00994018.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi:10.1109/TIT.1967.1053964.

Andrew Crane-Droesch. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, 13(11):114003, 2018. doi:10.1088/1748-9326/aae159.

P Dagnelie, Rodolphe Palm, and A Istasse. Prévision de productions agricoles dans les dix pays de la communauté economique européenne. Technical report, Faculté des Sciences Agronomiques de l'Etat, Gembloux , Belgium, 1983.

CST Daughtry, KP Gallo, SN Goward, SD Prince, and WP Kustas. Spectral estimates of absorbed radiation and phytomass production in corn and soybean canopies. *Remote Sensing of Environment*, 39(2):141–152, 1992. doi:10.1016/0034-4257(92)90132-4.

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

AJW de Wit and CA van Diepen. Crop model data assimilation with the ensemble kalman filter for improving regional crop yield forecasts. *Agricultural and Forest Meteorology*, 146 (1-2):38–56, 2007. doi:10.1016/j.agrformet.2007.05.004.

A.J.W. de Wit, A. Elhaddad, S. Meyer zum Alten Borgloh, U.D. Turdukulov, and R.W.A. Hutjes. Crop productivity and evapotranspiration indicators from 2000 to present derived from satellite observations. Climate Data Store - Copernicus Climate Change Service, https://doi.org/10.24381/cds.b2f6f9f6, 2022.

Allard de Wit, Hendrik Boogaard, Davide Fumagalli, Sander Janssen, Rob Knapen, Daniel van Kraalingen, Iwan Supit, Raymond van der Wijngaart, and Kees van Diepen. 25 years of the WOFOST cropping systems model. *Agricultural Systems*, 168:154–167, 2019. doi:10.1016/j.agsy.2018.06.018.

Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008. doi:10.1145/1327452.1327492.

Pierre Defourny, Sophie Bontemps, Nicolas Bellemans, Cosmin Cara, Gérard Dedieu, Eric Guzzonato, Olivier Hagolle, Jordi Inglada, Laurentiu Nicola, Thierry Rabaute, et al. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. *Remote sensing of environment*, 221:551–568, 2019. doi:10.1016/j.rse.2018.11.007.

Jillian M Deines, Rinkal Patel, Sang-Zi Liang, Walter Dado, and David B Lobell. 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 Sensing of Environment*, 253:112174, 2021. doi:10.1016/j.rse.2020.112174.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006. `https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf`, Last accessed: Sept 20, 2021.

Andreas Dengel. Special issue on artificial intelligence in agriculture, 2013.

Cüneyt Dirican. The impacts of robotics, artificial intelligence on business and economics. *Procedia-Social and Behavioral Sciences*, 195:564–573, 2015. doi:10.1016/j.sbspro.2015.06.134.

Randall J Donohue, Roger A Lawes, Gonzalo Mata, David Gobbett, and Jackie Ouzman. Towards a national, remote-sensing-based model for predicting field-scale crop yield. *Field Crops Research*, 227:79–90, 2018. doi:10.1016/j.fcr.2018.08.005.

Paul C Doraiswamy, Sophie Moulin, Paul W Cook, and Alan Stern. Crop yield assessment from remote sensing. *Photogrammetric engineering & remote sensing*, 69(6):665–674, 2003. doi:10.14358/PERS.69.6.665.

PC Doraiswamy, JL Hatfield, TJ Jackson, B Akhmedov, J Prueger, and Alan Stern. Crop condition and yield simulations using landsat and modis. *Remote sensing of environment*, 92(4):548–559, 2004. doi:10.1016/j.rse.2004.05.017.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of Interpretable Machine Learning, 2017.

EC-JRC. JRC MARS Bulletins, 2021. `https://ec.europa.eu/jrc/en/mars/bulletins`, Last accessed: May 11, 2021.

EC-JRC. JRC Agri4Cast Data Portal, 2022. `https://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx`, Last accessed: Feb 11, 2022.

Dhivya Elavarasan, Durai Raj Vincent, Vishal Sharma, Albert Y Zomaya, and Kathiravan Srinivasan. Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Computers and electronics in agriculture*, 155:257–282, 2018. doi:10.1016/j.compag.2018.10.024.

ESDAC. European soil database, 2021. `https://esdac.jrc.ec.europa.eu/resource-type/datasets`, Last accessed: April 28, 2021.

European Commission. Destination Earth, 2019. https://digital-strategy.ec.europa.eu/en/policies/destination-earth, Last accessed: July 27, 2022.

European Commission. A European strategy for data. Technical report, European Commission, 2020. https://ec.europa.eu/info/sites/default/files/communication-european-strategy-data-19feb2020_en.pdf, Last accessed: May 11, 2021.

European Commission. A European Green Deal, 2021. https://ec.europa.eu/info/strategy/priorities-2019-2024/european-green-deal_en, Last accessed: July 27, 2022.

Eurostat. Irrigated area of semi-intensive crops, updated 2016, 2016a.

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

Eurostat. Annual crop statistics handbook, 2020a. https://ec.europa.eu/eurostat/cache/metadata/Annexes/apro_cp_esms_an1.pdf, Last accessed: May 11, 2020.

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

Eurostat. Eurostat - Agricultural Production - crops, 2021a. https://ec.europa.eu/eurostat/statistics-explained/index.php/Agricultural_production_-_crops, Last accessed: May 11, 2021.

Eurostat. Eurostat Database, 2021b. https://ec.europa.eu/eurostat/web/agriculture/data/database, Last accessed: April 28, 2021.

Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning*, pp. 1437–1446. PMLR, 2018. http://proceedings.mlr.press/v80/falkner18a, Last accessed: April 28, 2021.

Joshua Fan, Junwen Bai, Zhiyun Li, Ariel Ortiz-Bobea, and Carla P. Gomes. A GNN-RNN approach for harnessing geospatial and temporal information: Application to crop yield prediction. *arXiv preprint arXiv:2111.08900v2*, 2021. doi:10.48550/arXiv.2111.08900.

Puyu Feng, Bin Wang, De Li Liu, Cathy Waters, and Qiang Yu. Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern australia. *Agricultural and Forest Meteorology*, 275:100–113, 2019. doi:10.1016/j.agrformet.2019.05.018.

Puyu Feng, Bin Wang, De Li Liu, Cathy Waters, Dengpan Xiao, Lijie Shi, and Qiang Yu. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agricultural and Forest Meteorology*, 285:107922, 2020. doi:10.1016/j.agrformet.2020.107922.

Patrick Filippi, Edward J Jones, Niranjan S Wimalathunge, Pallegedara DSN Somarathna, Liana E Pozza, Sabastine U Ugbaje, Thomas G Jephcott, Stacey E Paterson, Brett M Whelan, and Thomas FA Bishop. An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agriculture*, 20:1015–1029, 2019. doi:10.1007/s11119-018-09628-4.

RA Fischer. Definitions and determination of crop yield, yield gaps, and of rates of change. *Field Crops Research*, 182:9–18, 2015. doi:10.1016/j.fcr.2014.12.006.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. `http://jmlr.org/papers/v20/18-760.html`, Last accessed: Feb 22, 2023.

Christian Folberth, Artem Baklanov, Juraj Balkovič, Rastislav Skalskỳ, Nikolay Khabarov, and Michael Obersteiner. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agricultural and forest meteorology*, 264:1–15, 2019. doi:10.1016/j.agrformet.2018.09.021.

FR-Agreste. Agreste web data portal, 2020. `https://agreste.agriculture.gouv.fr/agreste-web/`, Last accessed: May 11, 2020.

Belén Franch, Juanma Cintas, Inbal Becker-Reshef, María José Sanchez-Torres, Javier Roger, Sergii Skakun, José Antonio Sobrino, Kristof Van Tricht, Jeroen Degerickx, Sven Gilliams, et al. Global crop calendars of maize and wheat in the framework of the WorldCereal project. *GIScience & Remote Sensing*, 59(1):885–913, 2022. doi:10.1080/15481603.2022.2079273.

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

Steffen Fritz, Linda See, Juan Carlos Laso Bayas, François Waldner, Damien Jacques, Inbal Becker-Reshef, Alyssa Whitcraft, Bettina Baruth, Rogerio Bonifacio, Jim Crutchfield, et al. A comparison of global agricultural monitoring systems and current gaps. *Agricultural systems*, 168:258–272, 2019. doi:10.1016/j.agsy.2018.05.010.

David García-León, Raúl López-Lozano, Andrea Toreti, and Matteo Zampieri. Local-scale cereal yield forecasting in Italy: Lessons from different statistical models and spatial aggregations. *Agronomy*, 10(6):809, 2020. doi:10.3390/agronomy10060809.

Deborah V. Gaso, Dilli R. Paudel, Allard de Wit, Laila Puntel, Adugna Mullissa, and Lammert Kooistra. Beyond assimilation of leaf area index: leveraging additional spectral information usingmachine learning for site-specific soybean yield prediction. *International Journal of Applied Earth Observation and Geoinformation* (Under Review), 2023.

Keyhan Gavahi, Peyman Abbaszadeh, and Hamid Moradkhani. Deepyield: A combined convolutional neural network with long short-term memory for crop yield forecasting. *Expert Systems with Applications*, 184:115511, 2021. doi:10.1016/j.eswa.2021.115511.

G Genovese and M Bettio. *Methodology of the MARS Crop Yield Forecasting System. Vol 4. Statistical Data Collection, Data Processing and Analysis*. Official Publications of the European Communities, Luxembourg, 2004. ISBN 92-894-8183-8.

Giampiero Genovese, C Vignolles, T Nègre, and G Passera. A methodology for a combined use of normalised difference vegetation index and corine land cover data for crop yield monitoring and forecasting. a case study on spain. *Agronomie*, 21 (1):91–111, 2001. `https://web.archive.org/web/201707060433334id_/https://hal.archives-ouvertes.fr/hal-00886104/document/`, Last accessed: July 27, 2022.

Sambuddha Ghosal, Bangyou Zheng, Scott C Chapman, Andries B Potgieter, David R Jordan, Xuemin Wang, Asheesh K Singh, Arti Singh, Masayuki Hirafuji, Seishi Ninomiya, et al. A weakly supervised deep learning framework for sorghum head detection and counting. *Plant Phenomics*, 2019, 2019. doi:10.34133/2019/1525874.

Anatoly A Gitelson, Yi Peng, and Karl F Huemmrich. Relationship between fraction of radiation absorbed by photosynthesizing maize and soybean canopies and NDVI from remotely sensed data taken at close range and from MODIS 250 m resolution data. *Remote Sensing of Environment*, 147:108–120, 2014. doi:10.1016/j.rse.2014.02.014.

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

Alberto González Sánchez, Juan Frausto Solís, Waldo Ojeda Bustamante, et al. Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research*, 12, 2014.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`, Last accessed: May 11, 2020.

David Gouache, Anne-Sophie Bouchon, Elodie Jouanneau, and Xavier Le Bris. Agrometeorological analysis and prediction of wheat yield at the departmental level in France. *Agricultural and Forest Meteorology*, 209:1–10, 2015. doi:10.1016/j.agrformet.2015.04.027.

Pablo M Granitto, Cesare Furlanello, Franco Biasioli, and Flavia Gasperi. Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2):83–90, 2006. doi:10.1016/j.chemolab.2006.01.007.

Patricio Grassini, Lenny GJ van Bussel, Justin Van Wart, Joost Wolf, Lieven Claessens, Haishun Yang, Hendrik Boogaard, Hugo de Groot, Martin K van Ittersum, and Kenneth G Cassman. How good is good enough? Data requirements for reliable crop yield simulations and yield-gap analysis. *Field Crops Research*, 177:49–63, 2015. doi:10.1016/j.fcr.2015.03.004.

Andrey Gritsenko, Anton Akusok, Yoan Miche, Kaj-Mikael Björk, Stephen Baek, and Amaury Lendasse. Combined nonlinear visualization and classification: Elmvis++ c. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2617–2624. IEEE, 2016. doi:10.1109/IJCNN.2016.7727527.

Qiang Gu, Anup Kumar, Simon Bray, Allison Creason, Alireza Khanteymoori, Vahid Jalili, Björn Grüning, and Jeremy Goecks. Galaxy-ml: An accessible, reproducible, and scalable machine learning toolkit for biomedicine. *PLoS computational biology*, 17(6):e1009014, 2021. doi:10.1371/journal.pcbi.1009014.

Tian Guo, Tao Lin, and Nino Antulov-Fantulin. Exploring interpretable LSTM neural networks over multi-variable data. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2494–2504. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/guo19b.html`.

Isabelle Guyon, Vladimir Vapnik, Bernhard Boser, Leon Bottou, and Sara A Solla. Structural risk minimization for character recognition. In *Advances in Neural Information Processing Systems*, pp. 471–479, 1992.

Jichong Han, Zhao Zhang, Juan Cao, Yuchuan Luo, Liangliang Zhang, Ziyue Li, and Jing Zhang. Prediction of winter wheat yield based on multi-source data and machine learning in china. *Remote Sensing*, 12(2):236, 2020. doi:10.3390/rs12020236.

Jingye Han, Liangsheng Shi, Christos Pylianidis, Qi Yang, and Ioannis N Athanasiadis. DeepOryza: A Knowledge guided machine learning model for rice growth simulation. In *2nd AAAI Workshop on AI for Agriculture and Food Systems*, 2023. `https://openreview.net/pdf?id=L9ankU4Ge-v`, Last accessed: Feb 27, 2023.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Tomislav Hengl, Jorge Mendes de Jesus, Gerard BM Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2):e0169748, 2017. doi:10.1371/journal.pone.0169748.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020. doi:10.1002/qj.3803.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi:10.1080/00401706.1970.10488634.

Dean P Holzworth, Neil I Huth, Peter G deVoil, Eric J Zurcher, Neville I Herrmann, Greg McLean, Karine Chenu, Erik J van Oosterom, Val Snow, Chris Murphy, et al. APSIM–evolution towards a new generation of agricultural systems simulation. *Environmental Modelling & Software*, 62:327–350, 2014. doi:10.1016/j.envsoft.2014.07.009.

Dean P Holzworth, Val Snow, Sander Janssen, Ioannis N Athanasiadis, Marcello Donatelli, Gerrit Hoogenboom, Jeffrey W White, and Peter Thorburn. Agricultural production systems modelling and software: current status and future prospects. *Environmental Modelling & Software*, 72:276–286, 2015. doi:10.1016/j.envsoft.2014.12.013.

Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020. doi:10.1145/3392878.

T Horie, M Yajima, and H Nakagawa. Yield forecasting. *Agricultural Systems*, 40(1-3): 211–236, 1992. doi:10.1016/0308-521X(92)90022-G.

IGC. About Us: International Grains Council, 2022. `http://www.igc.int/`, Last accessed: August 9, 2022.

Alan Inglis, Andrew Parnell, and Catherine B Hurley. Visualizing variable importance and variable interaction effects in machine learning models. *Journal of Computational and Graphical Statistics*, pp. 1–13, 2022. doi:10.1080/10618600.2021.2007935.

Ryan HL Ip, Li-Minn Ang, Kah Phooi Seng, JC Broster, and JE Pratley. Big data and machine learning for crop protection. *Computers and Electronics in Agriculture*, 151: 376–383, 2018. doi:10.1016/j.compag.2018.06.008.

Nathan Jacobs, Adam Kraft, Muhammad Usman Rafique, and Ranti Dev Sharma. A weakly supervised approach for estimating spatial density functions from high-resolution satellite imagery. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 33–42, 2018. doi:10.1145/3274895.3274934.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Kevin Jamieson and Ameet Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pp. 240–248. PMLR, 2016. `http://proceedings.mlr.press/v51/jamieson16.html`, Last accessed: May 20, 2021.

Sander JC Janssen, Cheryl H Porter, Andrew D Moore, Ioannis N Athanasiadis, Ian Foster, James W Jones, and John M Antle. Towards a new generation of agricultural system data, models and knowledge products: Information and communication technology. *Agricultural Systems*, 155:200–212, 2017. doi:10.1016/j.agsy.2016.09.017.

Jig Han Jeong, Jonathan P Resop, Nathaniel D Mueller, David H Fleisher, Kyungdahm Yun, Ethan E Butler, Dennis J Timlin, Kyo-Moon Shim, James S Gerber, Vangimalla R Reddy, et al. Random forests for global and regional crop yield predictions. *PLoS One*, 11(6): e0156571, 2016. doi:10.1371/journal.pone.0156571.

Seungtaek Jeong, Jonghan Ko, Taehwan Shin, and Jong-min Yeom. Incorporation of machine learning and deep neural network approaches into a remote sensing-integrated crop model for the simulation of rice growth. *Scientific Reports*, 12:9030, 2022. doi:10.1038/s41598-022-13232-y.

Zehui Jiang, Chao Liu, Baskar Ganapathysubramanian, Dermot J Hayes, and Soumik Sarkar. Predicting county-scale maize yields with publicly available data. *Scientific Reports*, 10(1): 1–12, 2020. doi:10.1038/s41598-020-71898-8.

David M Johnson. An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment*, 141:116–128, 2014. doi:10.1016/j.rse.2013.10.027.

Michael D Johnson, William W Hsieh, Alex J Cannon, Andrew Davidson, and Frédéric Bédard. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and forest meteorology*, 218:74–84, 2016. doi:10.1016/j.agrformet.2015.11.003.

James W Jones, Gerrit Hoogenboom, Cheryl H Porter, Ken J Boote, William D Batchelor, LA Hunt, Paul W Wilkens, Upendra Singh, Arjan J Gijsman, and Joe T Ritchie. The DSSAT cropping system model. *European journal of agronomy*, 18(3-4):235–265, 2003. doi:10.1016/S1161-0301(02)00107-7.

James W Jones, John M Antle, Bruno Basso, Kenneth J Boote, Richard T Conant, Ian Foster, H Charles J Godfray, Mario Herrero, Richard E Howitt, Sander Janssen, et al. Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. *Agricultural systems*, 155:269–288, 2017. doi:10.1016/j.agsy.2016.09.021.

Arlind Kadra, Marius Lindauer, Frank Hutter, and Josif Grabocka. Regularization is all you need: Simple neural nets can excel on tabular data. *arXiv preprint arXiv:2106.11189*, 2021. `https://arxiv.org/pdf/2106.11189.pdf`, Last accessed: Sept 20, 2021.

Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018. doi:10.1016/j.compag.2018.02.016.

John Kang, Russell Schwartz, John Flickinger, and Sushil Beriwal. Machine learning approaches for predicting radiation therapy outcomes: a clinician's perspective. *International Journal of Radiation Oncology*Biology*Physics*, 93(5):1127–1135, 2015. doi:10.1016/j.ijrobp.2015.07.2286.

Yanghui Kang and Mutlu Özdoğan. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. *Remote Sensing of Environment*, 228:144–163, 2019. doi:10.1016/j.rse.2019.04.005.

Ahmed Kayad, Marco Sozzi, Simone Gatto, Francesco Marinello, and Francesco Pirotti. Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. *Remote Sensing*, 11(23):2873, 2019. doi:10.3390/rs11232873.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. doi:10.2307/2332226.

Saeed Khaki and Lizhi Wang. Crop yield prediction using deep neural networks. *Frontiers in plant science*, 10:621, 2019. doi:10.3389/fpls.2019.00621.

Saeed Khaki, Lizhi Wang, and Sotirios V Archontoulis. A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, 10:1750, 2020. doi:10.3389/fpls.2019.01750.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

Teuvo Kohonen. *Self-organizing maps*. Springer, 2001.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for PyTorch, 2020.

PL Kooman and AJ Haverkort. Modelling development and growth of the potato crop influenced by temperature and daylength: LINTUL-POTATO. In *Potato ecology and modelling of crops under conditions limiting growth*, pp. 41–59. Springer, 1995. doi:10.1007/978-94-011-0051-9_3.

Alana Lajoie-O'Malley, Kelly Bronson, Simone van der Burg, and Laurens Klerkx. The future (s) of digital agriculture and sustainable food systems: An analysis of high-level policy documents. *Ecosystem Services*, 45:101183, 2020. doi:10.1016/j.ecoser.2020.101183.

Roger Lawes, Gonzalo Mata, Jonathan Richetti, Andrew Fletcher, and Chris Herrmann. Using remote sensing, process-based crop models, and machine learning to evaluate crop rotations across 20 million hectares in Western Australia. *Agronomy for Sustainable Development*, 42:120, 2022. doi:10.1007/s13593-022-00851-y.

Rémi Lecerf, Andrej Ceglar, Raúl López-Lozano, Marijn Van Der Velde, and Bettina Baruth. Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe. *Agricultural Systems*, 168:191–202, 2019. doi:10.1016/j.agsy.2018.03.002.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi:10.1038/nature14539.

Guoyong Leng and Jim Hall. Crop yield sensitivity of global major agricultural countries to droughts and the projected changes in the future. *Science of the Total Environment*, 654: 811–821, 2019. doi:10.1016/j.scitotenv.2018.10.434.

Myroslava Lesiv, Juan Carlos Laso Bayas, Linda See, Martina Duerauer, Domian Dahlia, Neal Durando, Rubul Hazarika, Parag Kumar Sahariah, Mar'yana Vakolyuk, Volodymyr Blyshchyk, et al. Estimating the global distribution of field size using crowdsourcing. *Global Change Biology*, 25(1):174–186, 2019. doi:10.1111/gcb.14492.

Linchao Li, Bin Wang, Puyu Feng, Huanhuan Wang, Qinsi He, Yakai Wang, De Li Liu, Yi Li, Jianqiang He, Hao Feng, et al. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. *Agricultural and Forest Meteorology*, 308:108558, 2021. doi:10.1016/j.agrformet.2021.108558.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.

Konstantinos Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8):2674, 2018. doi:10.3390/s18082674.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001. doi:10.1002/asmb.446.

Gunnar Lischeid, Heidi Webber, Michael Sommer, Claas Nendel, and Frank Ewert. Machine learning in crop yield modelling: A powerful tool, but no surrogate for science. *Agricultural and Forest Meteorology*, 312:108698, 2022. doi:10.1016/j.agrformet.2021.108698.

L. Liu, S. Xu, J. Tang, K. Guan, T. J. Griffis, M. D. Erickson, A. L. Frie, X. Jia, T. Kim, L. T. Miller, B. Peng, S. Wu, Y. Yang, W. Zhou, V. Kumar, and Z. Jin. KGML-ag: a modeling framework of knowledge-guided machine learning to simulate agroecosystems: a case study of estimating n_2o emission using data from mesocosm experiments. *Geoscientific Model Development*, 15(7):2839–2858, 2022. doi:10.5194/gmd-15-2839-2022.

David Lobell. Crop responses to climate: time-series models. In *Climate Change and Food Security*, pp. 85–98. Springer, 2010. doi:10.1007/978-90-481-2953-9_5.

David B Lobell, David Thau, Christopher Seifert, Eric Engle, and Bertis Little. A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, 164:324–333, 2015. doi:10.1016/j.rse.2015.04.021.

Rob Lokers, Rob Knapen, Sander Janssen, Yke van Randen, and Jacques Jansen. Analysis of Big Data technologies for use in agro-environmental science. *Environmental Modelling & Software*, 84:494–504, 2016. doi:10.1016/j.envsoft.2016.07.017.

Raúl López-Lozano and Bettina Baruth. An evaluation framework to build a cost-efficient crop monitoring system. Experiences from the extension of the European crop monitoring system. *Agricultural Systems*, 168:231–246, 2019. doi:10.1016/j.agsy.2018.04.002.

Raúl López-Lozano, Gregory Duveiller, Lorenzo Seguini, Michele Meroni, Sara García-Condado, Josh Hooker, Olivier Leo, and Bettina Baruth. Towards regional grain yield forecasting with 1 km-resolution EO biophysical products: Strengths and limitations at pan-European level. *Agricultural and Forest Meteorology*, 206:12–32, 2015. doi:10.3390/s18082674.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017. `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`, Last accessed: June 29, 2022.

Liye Ma and Baohong Sun. Machine learning and AI in marketing–Connecting computing power to human insights. *International Journal of Research in Marketing*, 37(3):481–504, 2020. doi:10.1016/j.ijresmar.2020.04.005.

Sebastian D Mackowiak, Henrik Zauber, Chris Bielow, Denise Thiel, Kamila Kutz, Lorenzo Calviello, Guido Mastrobuoni, Nikolaus Rajewsky, Stefan Kempa, Matthias Selbach, et al. Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*, 16(1):179, 2015. doi:10.1186/s13059-015-0742-x.

Mahalanobis Centre. Mahalanobis National Crop Forecast Centre, 2022. `https://www.ncfc.gov.in/`, Last accessed: Oct 27, 2022.

Spyros Makridakis. The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, 90:46–60, 2017. doi:10.1016/j.futures.2017.03.006.

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pp. 50–60, 1947. doi:10.1214/aoms/1177730491.

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

MARSWiki. MARS Crop Yield Forecasting System, 2021. `https://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Welcome_to_WikiMCYFS`, Last accessed: May 11, 2021.

Anna Mateo-Sanchis, Maria Piles, Julia Amorós-López, Jordi Muñoz-Marí, Jose E Adsuara, Álvaro Moreno-Martínez, and Gustau Camps-Valls. Learning main drivers of crop progress and failure in Europe with interpretable machine learning. *International Journal of Applied Earth Observation and Geoinformation*, 104:102574, 2021. doi:10.1016/j.jag.2021.102574.

J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. A proposal for Dartmouth summer research project on artificial intelligence, 1955. `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`, Last accessed: July 27, 2022.

Amy McGovern, Ryan Lagerquist, David John Gagne, G Eli Jergensen, Kimberly L Elmore, Cameron R Homeyer, and Travis Smith. Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199, 2019. doi:10.1175/BAMS-D-18-0195.1.

Michele Meroni, François Waldner, Lorenzo Seguini, Hervé Kerdiles, and Felix Rembold. Yield forecasting with machine learning and small data: What gains for grains? *Agricultural and Forest Meteorology*, 308:108555, 2021. doi:10.1016/j.agrformet.2021.108555.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019. doi:10.1016/j.artint.2018.07.007.

Tom M Mitchell. *Machine Learning*. McGraw Hill, New York, 1997.

Christoph Molnar. *Interpretable Machine Learning*. Christoph Molnar, 2 edition, 2022. URL `https://christophm.github.io/interpretable-ml-book`.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018. doi:10.1016/j.dsp.2017.10.011.

JL Monteith. Solar radiation and productivity in tropical ecosystems. *Journal of applied ecology*, 9(3):747–766, 1972. doi:10.2307/2401901.

Sander Mücher, Lorenzo De Simone, Henk Kramer, Allard de Wit, Laure Roupioz, Gerard Hazeu, Hendrik Boogaard, Rini Schuiling, Steffen Fritz, John Latham, et al. A new global agro-environmental stratification (GAES). Technical report, Wageningen Environmental Research, 2016. `https://edepot.wur.nl/400815`, Last accessed: June 14, 2021.

Andreas C Müller and Sarah Guido. *Introduction to machine learning with Python: a guide for data scientists*. O'Reilly Media, Inc., 2016.

Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation, 2018. URL `https://arxiv.org/abs/1802.00682`.

Hari Sankar Nayak, Joao Vasco Silva, Chiter Mal Parihar, Timothy J Krupnik, Dipaka Ranjan Sena, Suresh K Kakraliya, Hanuman Sahay Jat, Harminder Singh Sidhu, Parbodh C Sharma, Mangi Lal Jat, et al. Interpretable machine learning methods to explain on-farm yield variability of high productivity wheat in Northwest India. *Field Crops Research*, 287: 108640, 2022. doi:10.1016/j.fcr.2022.108640.

Petteri Nevavuori, Nathaniel Narra, and Tarmo Lipping. Crop yield prediction with deep convolutional neural networks. *Computers and electronics in agriculture*, 163:104859, 2019. doi:10.1016/j.compag.2019.104859.

Nathaniel K Newlands, David S Zamar, Louis A Kouadio, Yinsuo Zhang, Aston Chipanshi, Andries Potgieter, Souleymane Toure, and Harvey SJ Hill. An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. *Frontiers in Environmental Science*, 2:17, 2014. doi:10.3389/fenvs.2014.00017.

NIFA. Artificial Intelligence (AI) Research Institutes, 2021. `https://www.nifa.usda.gov/grants/funding-opportunities/artificial-intelligence-ai-research-institutes`, Last accessed: July 27, 2022.

NL-CBS. CBS Open Data Portal, 2020. `https://opendata.cbs.nl/statline`, Last accessed: May 11, 2020.

Alexandros Oikonomidis, Cagatay Catal, and Ayalew Kassahun. Deep learning for crop yield prediction: a systematic literature review. *New Zealand Journal of Crop and Horticultural Science*, pp. 1–26, 2022. doi:10.1080/01140671.2022.2032213.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. *Distill*, 2017. doi:10.23915/distill.00007.

Sjoukje A Osinga, Dilli Paudel, Spiros A Mouzakitis, and Ioannis N Athanasiadis. Big data in agriculture: Between opportunity and solution. *Agricultural Systems*, 195:103298, 2022. doi:10.1016/j.agsy.2021.103298.

Valentina Pagani, Tommaso Guarneri, Lorenzo Busetto, Luigi Ranghetti, Mirco Boschetti, Ermes Movedi, Manuel Campos-Taberner, Francisco Javier Garcia-Haro, Dimitrios Katsantonis, Dimitris Stavrakoudis, et al. A high-resolution, integrated system for rice yield forecasting at district level. *Agricultural Systems*, 168:181–190, 2019. doi:10.1016/j.agsy.2018.05.007.

Xanthoula Eirini Pantazi, Dimitrios Moshou, Thomas Alexandridis, Rebecca L Whetton, and Abdul Mounem Mouazen. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, 121:57–65, 2016. doi:10.1016/j.compag.2015.11.018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf`.

Dilli Paudel, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjoukje Osinga, Christos Pylianidis, and Ioannis N Athanasiadis. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187:103016, 2021. doi:10.1016/j.agsy.2020.103016.

Dilli Paudel, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, and Ioannis N Athanasiadis. Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*, 276:108377, 2022a. doi:10.1016/j.fcr.2021.108377.

Dilli Paudel, Allard de Wit, Hendrik Boogaard, Diego Marcos, Sjoukje Osinga, and Ioannis N Athanasiadis. Interpretability of deep learning models for crop yield forecasting. *Computers and Electronics in Agriculture*, 206:107663, 2023a. doi:10.1016/j.compag.2023.107663.

Dilli R. Paudel, Diego M. Gonzalez, Allard de Wit, Hendrik Boogaard, and Ioannis N. Athanasiadis. A weakly supervised framework for high-resolution crop yield forecasts. In *AI for Earth and Space Science Workshop at ICLR 2022*. International Conference on Learning Representations, 2022b. doi:10.48550/ARXIV.2205.09016.

Dilli R. Paudel, Diego Marcos, Allard de Wit, Hendrik Boogaard, and Ioannis N. Athanasiadis. A weakly supervised framework for high resolution crop yield forecasts. *Environmental Research Letters* (Under Review), 2023b.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ben Phalan, Rhys Green, and Andrew Balmford. Closing yield gaps: perils and possibilities for biodiversity conservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1639):20120285, 2014. doi:10.1098/rstb.2012.0285.

Laura Poggio, Luis M De Sousa, Niels H Batjes, Gerard Heuvelink, Bas Kempen, Eloi Ribeiro, and David Rossiter. Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil*, 7(1):217–240, 2021. doi:10.5194/soil-7-217-2021.

Vera Porwollik, Christoph Müller, Joshua Elliott, James Chryssanthacopoulos, Toshichika Iizumi, Deepak K Ray, Alex C Ruane, Almut Arneth, Juraj Balkovič, Philippe Ciais, et al. Spatial and temporal uncertainty of crop yield aggregations. *European Journal of Agronomy*, 88:10–21, 2017.

Christos Pylianidis and Ioannis N. Athanasiadis. Learning latent representations for operational nitrogen response rate prediction. In *AI for Earth and Space Science Workshop at ICLR 2022*. International Conference on Learning Representations, 2022. doi:10.48550/arXiv.2205.09025.

QGIS Development Team. QGIS Geographic Information System, 2020. `http://qgis.osgeo.org`, Last accessed: May 11, 2020.

Budong Qian, Reinder De Jong, Jingyi Yang, Hong Wang, and Sam Gameda. Comparing simulated crop yields with observed and synthetic weather data. *Agricultural and Forest Meteorology*, 151(12):1781–1791, 2011. doi:10.1016/j.agrformet.2011.07.016.

J. Ross Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986. doi:10.1007/BF00116251.

Steven M Quiring and David R Legates. Application of CERES-maize for within-season prediction of rainfed corn yields in Delaware, USA. *Agricultural and forest meteorology*, 148(6-7):964–975, 2008. doi:10.1016/j.agrformet.2008.01.009.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016. URL `https://arxiv.org/abs/1606.05386`.

Bradley R Rippey. The us drought of 2012. *Weather and Climate Extremes*, 10:57–64, 2015. doi:10.1016/j.wace.2015.10.004.

JT Ritchie, U Singh, DC Godwin, and WT Bowen. Cereal growth, development and yield. In *Understanding options for agricultural production*, pp. 79–98. Springer, 1998. doi:10.1007/978-94-017-3624-4_5.

O Rojas, F Rembold, A Royer, and T Negre. Real-time agrometeorological crop yield monitoring in Eastern Africa. *Agronomy for Sustainable Development*, 25(1):63–77, 2005. `https://hal.archives-ouvertes.fr/hal-00886252/`, Last accessed: July 27, 2022.

David Christian Rose, Anna Barkemeyer, Auvikki de Boon, Catherine Price, and Dannielle Roche. The old, the new, or the old made new? everyday counter-narratives of the so-called fourth agricultural revolution. *Agriculture and Human Values*, pp. 1–17, 2022. doi:10.1007/s10460-022-10374-7.

José Luis Ruiz-Real, Juan Uribe-Toril, José Antonio Torres Arriaza, and Jaime de Pablo Valenciano. A look at the past, present and future research trends of Artificial Intelligence in agriculture. *Agronomy*, 10(11):1839, 2020. doi:10.3390/agronomy10111839.

Stefan Rüping. Learning interpretable models, 2006. `https://eldorado.tu-dortmund.de/bitstream/2003/23008/1/dissertation_rueping.pdf`, Last accessed: June 29, 2022.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

Bernhard Schauberger, Tamara Ben-Ari, David Makowski, Tomomichi Kato, Hiromi Kato, and Philippe Ciais. Yield trends, variability and stagnation analysis of major crops in france over more than a century. *Scientific Reports*, 8:16865, 2018. doi:10.1038/s41598-018-35351-1.

Bernhard Schauberger, Jonas Jägermeyr, and Christoph Gornott. A systematic review of local to regional yield forecasting approaches and frequently used data resources. *European Journal of Agronomy*, 120:126153, 2020. doi:10.1016/j.eja.2020.126153.

Jack Schieffer and Carl Dillon. The economic and environmental impacts of precision agriculture and interactions with agro-environmental policy. *Precision Agriculture*, 16(1): 46–61, 2015. doi:10.1007/s11119-014-9382-5.

Jonas Schmitt, Frank Offermann, Mareike Söder, Cathleen Frühauf, and Robert Finger. Extreme weather events cause significant crop yield losses at the farm level in German agriculture. *Food Policy*, 112:102359, 2022. doi:10.1016/j.foodpol.2022.102359.

Maja Schneider, Amelie Broszeit, and Marco Körner. EuroCrops: A pan-European dataset for time series crop type classification. In *Proceedings of the 2021 conference on Big Data from Space*. European Commission (Joint Research Centre) and SatCen. and European Space Agency, 2021. doi:10.2760/125905.

R Schnepf. NASS and US Crop Production Forecasts: Methods and Issues. Technical report, Congressional Research Service, 2017. `https://fas.org/sgp/crs/misc/R44814.pdf`, Last accessed: May 11, 2020.

Scikit-optimize Contributors. Scikit-optimize: Sequential model-based optimization, 2021. `https://scikit-optimize.github.io/stable/getting_started.html`, Last accessed: Sept 20, 2021.

L Seguini, A Bussay, and B Baruth. From extreme weather to impacts: The role of the areas of concern maps in the JRC MARS bulletin. *Agricultural systems*, 168:213–223, 2019. doi:10.1016/j.agsy.2018.07.003.

Mohsen Shahhosseini, Rafael A Martinez-Feria, Guiping Hu, and Sotirios V Archontoulis. Maize yield and nitrate loss prediction with machine learning algorithms. *Environmental Research Letters*, 14(12):124026, 2019. doi:10.1088/1748-9326/ab5268.

Mohsen Shahhosseini, Guiping Hu, Isaiah Huber, and Sotirios V Archontoulis. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific reports*, 11(1):1–15, 2021. doi:10.1038/s41598-020-80820-1.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015. doi:10.1109/JPROC.2015.2494218.

Claude E Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

L Shapley. Quota solutions op n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pp. 307–317. Princeton University Press, 1953. doi:10.1515/9781400881970.

Mark Shepherd, James A Turner, Bruce Small, and David Wheeler. Priorities for science to overcome hurdles thwarting the full promise of the 'digital agriculture' revolution. *Journal of the Science of Food and Agriculture*, 100(14):5083–5092, 2020. doi:10.1002/jsfa.9346.

Paresh B Shirsath, Vinay Kumar Sehgal, and Pramod K Aggarwal. Downscaling regional crop yields to local scale using remote sensing. *Agriculture*, 10(3):58, 2020. doi:10.3390/agriculture10030058.

Johnathon Shook, Tryambak Gangopadhyay, Linjiang Wu, Baskar Ganapathysubramanian, Soumik Sarkar, and Asheesh K Singh. Crop yield prediction integrating genotype and weather variables using deep learning. *Plos one*, 16(6):0252402, 2021. doi:10.1371/journal.pone.0252402.

Matthew J Smith. Getting value from artificial intelligence in agriculture. *Animal Production Science*, 60(1):46–54, 2018.

Abir Smiti. A critical overview of outlier detection methods. *Computer Science Review*, 38: 100306, 2020. doi:10.1016/j.cosrev.2020.100306.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 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*, pp. 1201–1211. Association for Computational Linguistics, 2012. doi:10.5555/2390948.2391084.

EG Souza, CL Bazzi, R Khosla, MA Uribe-Opazo, and Robin M Reich. Interpolation type and data computation of crop yield maps is important for precision crop production. *Journal of Plant Nutrition*, 39(4):531–538, 2016. doi:10.1080/01904167.2015.1124893.

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

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

Luc Steinbuch, Thomas G Orton, and Dick J Brus. Model-based geostatistics from a bayesian perspective: Investigating area-to-point kriging with small data sets. *Mathematical Geosciences*, 52(3):397–423, 2020. doi:10.1007/s11004-019-09840-6.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017. URL `https://proceedings.mlr.press/v70/sundararajan17a.html`.

Harald Sundmaeker, CN Verdouw, J Wolfert, and Luis Perez Freire. Internet of Food and Farm 2020. In *Digitising the industry*, volume 49, pp. 129–150. River Publishers, 2016. `https://library.wur.nl/WebQuery/wurpubs/507125`, Last accesssed: July 27, 2022.

I Supit and E van der Goot. *Updated system description of the WOFOST crop growth simulation model as implemented in the crop growth monitoring system applied by the European Commission.* Treemail Pubilshers, Heelsum, 2003.

I Supit, AA Hooijer, and CA Van Diepen. System description of the WOFOST 6.0 crop simulation model implemented in CGMS. vol. 1. theory and algorithms. In *EUR Publication No. 15959 EN*, pp. 146. Office for Official Publications of the European Communties, Luxembourg, 1994.

Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021. doi:10.1145/3411764.3445088.

The Galaxy Community. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, 50(W1):W345–W351, 2022. doi:10.1093/nar/gkac247.

M.M. Thornton, Y. Wei, P.E. Thornton, R. Shrestha, S. Kao, and B.E. Wilson. Daymet: Station-Level Inputs and Cross-Validation Result for North America, Version 4, 2020. URL `https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1850`.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi:10.1111/j.2517-6161.1996.tb02080.x.

David Tilman, Christian Balzer, Jason Hill, and Belinda L Befort. Global food demand and the sustainable intensification of agriculture. In *Proceedings of the National Academy of Sciences*, volume 108(50), pp. 20260–20264. National Academy of Sciences of the US, 2011. doi:10.1073/pnas.1116437108.

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

USDA-NASS. Statistics by subject - crops, 2022. `https://www.nass.usda.gov/Statistics_by_Subject/index.php?sector=CROPS`, Last accessed: August 10, 2022.

USGS-EROS. USGS EROS Archive - Digital Elevation - Global 30 Arc-Second Elevation (GTOPO30), 2021. `https://www.usgs.gov/centers/eros/data`, Last accessed: May 11, 2021.

Simone van der Burg, Leanne Wiseman, and Jovana Krkeljas. Trust in farm data sharing: Reflections on the EU code of conduct for agricultural data sharing. *Ethics and Information Technology*, 23:185–198, 2021. doi:10.1007/s10676-020-09543-1.

M van der Velde and L Nisini. 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. *Agricultural Systems*, 168:203–212, 2019. doi:10.1016/j.agsy.2018.06.009.

M van der Velde, B Baruth, A Bussay, A Ceglar, S Garcia Condado, S Karetsos, R Lecerf, R Lopez, A Maiorano, L Nisini, et al. In-season performance of European Union wheat forecasts during extreme impacts. *Scientific Reports*, 8(1):1–10, 2018. doi:10.1038/s41598-018-33688-1.

Marijn van der Velde, Gunter Wriedt, and Fayçal Bouraoui. Estimating irrigation use and effects on maize yield during the 2003 heat wave in France. *Agriculture, Ecosystems & Environment*, 135(1-2):90–97, 2010. doi:10.1016/j.agee.2009.08.017.

Marijn van der Velde, Irene Biavetti, Mohamed El-Aydam, Stefan Niemeyer, Fabien Santini, and Maurits van den Berg. Use and relevance of European Union crop monitoring and yield forecasts. *Agricultural systems*, 168:224–230, 2019.

CA van Diepen, J Wolf, H Van Keulen, and C Rappoldt. WOFOST: a simulation model of crop production. *Soil Use and Management*, 5(1):16–24, 1989. doi:10.1111/j.1475-2743.1989.tb00755.x.

Thomas Van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020. doi:10.1016/j.compag.2020.105709.

Justin van Wart, Patricio Grassini, Haishun Yang, Lieven Claessens, Andrew Jarvis, and Kenneth G Cassman. Creating long-term weather data from thin air for crop simulation modeling. *Agricultural and Forest Meteorology*, 209:49–58, 2015. doi:10.1016/j.agrformet.2015.02.020.

Luan Peroni Venancio, Everardo Chartuni Mantovani, Cibele Hummel do Amaral, Christopher Michael Usher Neale, Ivo Zution Gonçalves, Roberto Filgueiras, and Isidro Campos. Forecasting corn yield at the farm level in Brazil based on the FAO-66 approach and soil-adjusted vegetation index (SAVI). *Agricultural Water Management*, 225:105779, 2019. doi:10.1016/j.agwat.2019.105779.

C von der Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100, 1973. doi:10.1007/BF00288907.

Toby Walsh. The ai revolution. *NSW Department of Education Education: Future Frontiers*, 2017. `https://www.saeon.com.au/toniedoc/ai-revolution.pdf`, Last accessed: July 27, 2022.

Achim Walter, Robert Finger, Robert Huber, and Nina Buchmann. Smart farming is key to developing sustainable agriculture. *Proceedings of the National Academy of Sciences*, 114 (24):6148–6150, 2017. doi:10.1073/pnas.1707462114.

Sherrie Wang, François Waldner, and David B. Lobell. Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision. *arXiv preprint arXiv:2201.04771*, 2022. doi:10.48550/arXiv.2201.04771.

Xinlei Wang, Jianxi Huang, Quanlong Feng, and Dongqin Yin. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of china with deep learning approaches. *Remote Sensing*, 12(11):1744, 2020. doi:10.3390/rs12111744.

Zhiguang Wang and Jianbo Yang. Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation. In *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018. `https://www.aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16668/0`, Last accessed: June 29, 2022.

CL Wiegand, AJ Richardson, DE Escobar, and AH Gerbermann. Vegetation indices in crop assessments. *Remote sensing of Environment*, 35(2-3):105–119, 1991. doi:10.1016/0034-4257(91)90004-P.

Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 55(4):1–37, 2022. doi:10.1145/3514228.

Simon Willcock, Javier Martínez-López, Danny AP Hooftman, Kenneth J Bagstad, Stefano Balbi, Alessia Marzo, Carlo Prato, Saverio Sciandrello, Giovanni Signorello, Brian Voigt, et al. Machine learning for ecosystem services. *Ecosystem services*, 33:165–174, 2018. doi:10.1016/j.ecoser.2018.04.004.

Aleksandra Wolanin, Gonzalo Mateo-García, Gustau Camps-Valls, Luis Gómez-Chova, Michele Meroni, Gregory Duveiller, You Liangzhi, and Luis Guanter. Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environmental Research Letters*, 15(2):024019, 2020. doi:10.1088/1748-9326/ab68ac.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1-3):37–52, 1987. doi:10.1016/0169-7439(87)80084-9.

Bingfang Wu, Jihua Meng, Qiangzi Li, Nana Yan, Xin Du, and Miao Zhang. Remote sensing-based global crop monitoring: experiences with china's cropwatch system. *International Journal of Digital Earth*, 7(2):113–137, 2014. doi:10.1080/17538947.2013.821185.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*, pp. 563–574. Springer, 2019. doi:10.1007/978-3-030-32236-6_51.

Jiaxuan You, Xiaocheng Li, Melvin Low, David Lobell, and Stefano Ermon. Deep Gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. `https://www-cs.stanford.edu/~ermon/papers/cropyield_AAAI17.pdf`, Last accessed: July 25, 2022.

Liangzhi You, Stanley Wood, Ulrike Wood-Sichra, and Wenbin Wu. Generating global crop distribution maps: From census to grid. *Agricultural Systems*, 127:53–60, 2014. doi:10.1016/j.agsy.2014.01.002.

Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59 (11):56–65, 2016. doi:10.1145/2934664.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014. doi:10.1007/978-3-319-10590-1_53.

Yan Zhao, Andries B Potgieter, Miao Zhang, Bingfang Wu, and Graeme L Hammer. Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling. *Remote Sensing*, 12(6):1024, 2020. doi:10.3390/rs12061024.

Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018. doi:10.1093/nsr/nwx106.

Acknowledgements

Funding: The research in this thesis was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825355 (CYBELE).

Data: I 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, I 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.

Ideas: I acknowledge D. Tuia and Hiske Overweg (previously from Geo-Information and Remote Sensing Lab of Wageningen University and Research) for contributing their ideas. I 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 their feedback on *Chapter 1*. I am grateful to Prof Baskar Ganapathysubramanian and Mojdeh Saadati from Iowa State University, Ames, IA for their feedback on the thesis outline.

Interpretability assessment: I would like to express my sincere thanks to the experts who helped analyze interpretability of deep learning models: Allard de Wit and Hendrik Boogaard from Wageningen Environmental Research; Majid Ali Magham from Plant Productions Group of Wageningen University and Research; Wassim Ben Aoun and Martin Claverie from the European Commission's Joint Research Centre (JRC) in Ispra, Italy.

Co-authors: I would like to acknowledge the contribution, guidance and support of all co-authors: Allard de Wit, Hendrik Boogaard, Diego Marcos, Christos Pylianidis, Marijn van der Velde, Luigi Nisini and Martin Claverie.

Supervisors: I would like to sincerely appreciate the guidance and supervision provided by Dr Sjoukje Osinga, Dr Sander Janssen and Prof Ioannis N. Athanasiadis during my PhD. They not only supervised my research, but also provided support on personal matters.

Departments: I would like to thank Prof Bedir Tekinerdogan and other colleagues from Information Technology department for their support. I would also like to acknowledge guidance received from Claudia Ravestein, Marieke Moller, Jeannette Lubbers-Poortvliet, Natasja Ariesen, Laura Simon. From Geo-Information and Remote Sensing department, I would like to thank Truus van de Hoef and Patty van Beek.

Wageningen School of Social Sciences: I would acknowledge Heleen Danen, Marcella Haan and Fennie van Straalen from WASS and Roelfina Mihalj-Ijken from HR.

Summary

Accurate crop yield forecasts are valuable to many stakeholders, such as farmers, commodity traders, policymakers and food aid programs, to adjust food security policies, farm management practices and marketing plans for farm products. To support informed decision-making, all stakeholders need access to timely, consistent, unbiased and reliable yield forecasting models. Model reliability depends on accuracy of predictions as well as human understanding of the reasons behind those predictions. Operational large-scale yield forecasting systems commonly build linear statistical models due to their simplicity and interpretability. These models do not always capture complex relationships between predictors and yield. Machine learning can combine existing domain knowledge with data-driven learning and model complex relationships. Machine learning also provides increased automation to produce crop yield forecasts at large scale, both in terms of multiple crops and countries and high resolution. Some machine learning methods, such as linear regression and decision trees, are inherently interpretable. For other methods and neural networks, post-hoc feature attribution methods can be used to learn explanations based on predictions of models. Overall, machine learning and deep learning methods provide a set tools that have not been explored fully in large-scale crop yield forecasting.

This thesis investigated the benefits and challenges of using machine learning for large-scale crop yield forecasting. It serves as a roadmap for large-scale crop yield forecasting by identifying key requirements (the "what") and designing and implementing workflows to address them (the "how"). The requirements include automated workflows to provide a consistent and reproducible method of forecasting yields across multiple spatial levels (e.g. grids, regions, countries). The workflows have to account for data sparsity and produce crop yield forecasts backed by explanations understandable to human stakeholders. The four sub-objectives addressed these requirements by defining the forecasting setup, designing workflows suitable for large scale, handling missing data and evaluating interpretability of model forecasts.

In *Chapter 2*, we designed a generic explainable, modular and reusable machine learning workflow to forecast crop yields for multiple crops and countries. The workflow is called a machine learning baseline for large-scale crop yield forecasting. To combine the strengths of existing methods with data-driven learning, the input data included crop model outputs and remote sensing indicators. Static features helped to capture some regional variation in the mean yield. Yield trend features accounted for improvements in technology, genetics and farm management. Crop modeling and agronomy experts selected the important indicators in different stages of crop growth and development to design seasonal features. The workflow

is modular to enable updates and improvements over time. Indicators selected to design features can be updated based on new knowledge or context. Similarly, we can experiment with new machine learning algorithms or neural network architectures. The workflow can be reused by selecting configuration options, such as crop and country. The emphasis on reusability helped shift focus from point solutions to specific case studies to generic and scalable workflows for many crops and locations. The baseline can be extended to other crops and countries in Europe. Applications to other continents are possible when equivalent crop productivity indicators (e.g. dry-weight yield biomass, leaf area, development stage) are available.

In *Chapter 3*, the machine learning baseline was improved and evaluated on six crops and nine major crop-growing countries of Europe at both regional level and national level. In Europe, the European Commission's Directorate General for Agriculture and Rural Development (DG-AGRI) shares national forecasts from the MARS Crop Yield Forecasting System (MCYFS) with policymakers (in the European Parliament and Member States), stakeholders monitoring crop production and trade and the Agricultural Market Information System (AMIS). The forecasts help stakeholders anticipate fluctuations in expected yields. Our main finding was that regional differences cancel out at national level and national forecasts do not provide information about those differences. Hence our work underscored the need for timely publication of regional crop yield forecasts in Europe. Supporting country-level JRC forecasts with consistent and reliable regional forecasts would provide useful information to stakeholders in the food production chain. We found that machine learning workflows provide a consistent and reproducible method to forecast both regional and national yields. Therefore, MCYFS analysts could use our workflows for cases where machine learning performs well (based on validation set performance) and continue using expertise-based methods for others.

In *Chapter 4*, the expert-driven feature design used in previous steps was replaced by automatic feature learning. Features learned automatically by neural networks were evaluated for predictive power and interpretability. For interpretability, human experts in agronomy, crop modeling and operational yield forecasting in Europe looked at feature attributions extracted from post-hoc analysis of model predictions and provided feedback on how well the relative magnitude and direction (positive or negative) of feature influence on yield matched their knowledge. Deep learning extracted features that performed similar to expert-designed ones, and model forecasts were based on plausible relationships between features and crop yield. Feature influence matched expert knowledge particularly well for cases with high forecasting accuracy (and low errors). Previous studies have looked at feature importance or activation maps for interpretability. However, they have not validated whether their analysis is understandable to human stakeholders. We developed an approach to assess model interpretability with feedback from human experts that went beyond feature importance plots, which are often not understandable on their own.

In *Chapter 5*, we designed a weakly supervised deep learning framework to produce high resolution crop yield forecasts using high resolution predictor data and low resolution yields and crop areas. Strong supervised models can be built only at the administrative levels where yield statistics are published (e.g. NUTS2 or NUTS3 statistical regions in Europe and counties in the US). We showed that weak supervision can produce yield forecasts at higher

resolutions than these. Weak supervised models produced reliable high resolution forecasts for two different settings (Europe and the United States), both in terms of agro-environmental factors and spatial resolution. The weakly supervised framework is useful to researchers working on similar problems, where labels may be missing for various reasons. Weakly supervised models can work as intermediate solutions in places where high resolution labels can be collected over time. Strongly supervised models can be built when the number of labeled instances becomes large.

Overall, our work showed that machine learning can incorporate domain knowledge and complement expert-driven approaches to provide a more automated, consistent and reproducible approach to crop yield forecasting across multiple spatial levels. When looking at the bigger picture, there are gaps and challenges related to data availability and quality, the expertise required to select the forecasting setup, a shortage of customized tools and benchmarks, concerns about interpretability and insufficient attention to stakeholder requirements. The convergence of knowledge-based methods, remote sensing and machine learning and the emergence of explainable AI or interpretable machine learning provides opportunities to address many of these challenges. Significant progress has been made in areas where researchers can work on their own; more effort is required in areas that involve collaborating with experts from another domain or interacting with stakeholders.

List of publications

Peer-reviewed journal publications

Dilli Paudel, Hendrik Boogaard, Allard de Wit, Sander Janssen, Sjoukje Osinga, Christos Pylianidis, and Ioannis N Athanasiadis. Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, 187:103016, 2021. doi:10.1016/j.agsy.2020.103016

Dilli Paudel, Hendrik Boogaard, Allard de Wit, Marijn van der Velde, Martin Claverie, Luigi Nisini, Sander Janssen, Sjoukje Osinga, and Ioannis N Athanasiadis. Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*, 276:108377, 2022a. doi:10.1016/j.fcr.2021.108377

Sjoukje A Osinga, Dilli Paudel, Spiros A Mouzakitis, and Ioannis N Athanasiadis. Big data in agriculture: Between opportunity and solution. *Agricultural Systems*, 195:103298, 2022. doi:10.1016/j.agsy.2021.103298

Dilli Paudel, Allard de Wit, Hendrik Boogaard, Diego Marcos, Sjoukje Osinga, and Ioannis N Athanasiadis. Interpretability of deep learning models for crop yield forecasting. *Computers and Electronics in Agriculture*, 206:107663, 2023a. doi:10.1016/j.compag.2023.107663

Journal articles under review

Dilli R. Paudel, Diego Marcos, Allard de Wit, Hendrik Boogaard, and Ioannis N. Athanasiadis. A weakly supervised framework for high resolution crop yield forecasts. *Environmental Research Letters* (Under Review), 2023b

Deborah V. Gaso, Dilli R. Paudel, Allard de Wit, Laila Puntel, Adugna Mullissa, and Lammert Kooistra. Beyond assimilation of leaf area index: leveraging additional spectral information usingmachine learning for site-specific soybean yield prediction. *International Journal of Applied Earth Observation and Geoinformation* (Under Review), 2023

Conference publications

Dilli R. Paudel, Diego M. Gonzalez, Allard de Wit, Hendrik Boogaard, and Ioannis N. Athanasiadis. A weakly supervised framework for high-resolution crop yield forecasts. In *AI for Earth and Space Science Workshop at ICLR 2022*. International Conference on Learning Representations, 2022b. doi:10.48550/ARXIV.2205.09016

About the author

Dilli Raj Paudel was born on June 23, 1984 in a village called Binamare in Baglung, Nepal. He attended Budhanilkantha School, a residential school in Kathmandu, from grades 4 to 12. After completing high school, he taught at the same school for 5 years. The education and teaching experience at Budhanilkantha School shaped the rest of his life in many ways. In particular, his experience as a student and as a teacher helped him get to Stanford University in California, USA. He holds BS and MS degrees in computer science from Stanford. His coursework focused on software development, systems design, data management and machine learning. While at Stanford, he interned at several software companies, including Yahoo!, Intuit and Twitter.

After completing MS, Dilli worked at Oracle Corporation of network security, specifically firewalls, and a distributed key vault for storing passwords, certificates and encryption keys of databases. In 2019, he started a PhD at the Information Technology department of Wageningen University and Research.

His current research focuses on application of machine learning and deep learning to large-scale crop yield forecasting. His main interest is to combine existing domain knowledge, from experts or process-based models, with remote sensing and machine learning for crop yield forecasting and other environmental applications, such as landslide hazard assessment.

Dilli R. Paudel
Wageningen School of Social Sciences (WASS)
Completed Training and Supervision Plan

Name of the learning activity	Department/Institute	Year	ECTS*
A) Project related competences			
A1 Managing a research project			
WASS Introduction	WASS	2019	1
Writing research proposal	WUR	2019	6
Course on Scientific Writing	WGS	2019	1.8
"Machine learning for large-scale crop yield forecasting"	WASS PhD-day	2020	0.5
Review papers	Environmental Modelling and Software; Algal Research	2020-2021	2
A2 Integrating research in the corresponding discipline			
Deep Learning (GRS 34806)	PE&RC	2021	6
Deep Learn 2019 Summer School	IRDTA	2019	2
Art of Modelling	PE&RC	2019	3
Machine learning for spatial data	PE&RC	2019	1.5
B) General research related competences			
B1 Placing research in a broader scientific context			
Advances in Intercropping	PE&RC	2021	1.5
Academic Publication and Presentation in Social Sciences	WASS	2020	4
Ethics for Social Sciences Research	WGS	2019	0.5
B2 Placing research in a societal context			
Organized CYBELE Demonstrators' Workshop (2-days)	-	2020	2
Organized Dragon Project Workshop: "Machine learning for yield prediction" (1-day)	-	2021	1
C) Career related competences/personal development			
C1 Employing transferable skills in different domains/careers			
Supervision of MSc thesis	WUR	2021-2022	1
Teaching Assistant : Big Data (INF 33806)	WUR	2019-2020	2
Supervising BSc/MSc students	WGS	2021	0.64
Total			**36.44**

*One credit according to ECTS is on average equivalent to 28 hours of study load

The research in this thesis was partially supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 825355 (CYBELE).

Financial support received from Wageningen University for printing this thesis is gratefully acknowledged.

Cover Design by Emma Tseng

Printed by www.proefschriftmaken.nl

ISBN: 978-9-46447-599-9

https://doi.org/10.18174/588095