remote sensing

MDPI

Article

Dynamic Maize Yield Predictions Using Machine Learning on Multi-Source Data

Michele Croci [1,2,*], Giorgio Impollonia [1,2], Michele Meroni [3] and Stefano Amaducci [1,2]

[1] Department of Sustainable Crop Production, Università Cattolica del Sacro Cuore, 29122 Piacenza, Italy
[2] Remote Sensing and Spatial Analysis Research Center (CRAST), Università Cattolica del Sacro Cuore, 29122 Piacenza, Italy
[3] European Commission, Joint Research Centre (JRC), Via E. Fermi 2749, 21027 Ispra, Italy
* Correspondence: michele.croci@unicatt.it

Abstract: Timely yield prediction is crucial for the agri-food supply chain as a whole. However, different stakeholders in the agri-food sector require different levels of accuracy and lead times in which a yield prediction should be available. For the producers, predictions during the growing season are essential to ensure that information is available early enough for the timely implementation of agronomic decisions, while industries can wait until later in the season to optimize their production process and increase their production traceability. In this study, we used machine learning algorithms, dynamic and static predictors, and a phenology approach to determine the time for issuing the yield prediction. In addition, the effect of data reduction was evaluated by comparing results obtained with and without principal component analysis (PCA). Gaussian process regression (GPR) was the best for predicting maize yield. Its best performance (nRMSE of 13.31%) was obtained late in the season and with the full set of predictors (vegetation indices, meteorological and soil predictors). In contrast, neural network (NNET) and support vector machines linear basis function (SVMl) achieved their best accuracy with only vegetation indices and at the tasseling phenological stage. Only slight differences in performance were observed between the algorithms considered, highlighting that the main factors influencing performance are the timing of the yield prediction and the predictors with which the machine learning algorithms are fed. Interestingly, PCA was instrumental in increasing the performances of NNET after this stage. An additional benefit of the application of PCA was the overall reduction between 12 and 30.20% in the standard deviation of the maize yield prediction performance from the leave one-year outer-loop cross-validation, depending on the feature set.

Keywords: Sentinel-2; yield prediction; phenology; machine learning; multi-source data; dimensionality reduction

check for updates

Citation: Croci, M.; Impollonia, G.; Meroni, M.; Amaducci, S. Dynamic Maize Yield Predictions Using Machine Learning on Multi-Source Data. *Remote Sens.* **2023**, *15*, 100. https://doi.org/10.3390/rs15010100

Academic Editor: Raffaele Casa

Received: 16 November 2022
Revised: 14 December 2022
Accepted: 21 December 2022
Published: 24 December 2022

1. Introduction

The projections of demographic increase in the next decades [1], climate change previsions [2], and the changes in economic systems are expected to create strong pressure on agri-food supply chains in the near future [3,4]. A number of approaches have been proposed to face these global challenges [5]. Still, irrespective of the solution, it is important to develop systems that ensure constant monitoring of the entire agri-food supply chain. Such systems are critical for reducing the risks associated with the driving uncertainty factors (e.g., extreme weather events) that negatively impact the economic, environmental, and social aspects of the agri-food supply chain.

Monitoring crop health, growth, and productivity can help address the global food challenges by enabling long-term sustainable development strategies to be planned. In recent decades, many researchers have focused their attention on developing monitoring techniques to predict crop yield [6]. Early and accurate crop yield forecasting is essential for the entire agricultural food production chain, from individual producers to processing industries and regional authorities [7]. In particular, timely and reliable crop forecasts play an important role in supporting national and international food security policies for stabilizing markets and planning interventions in food-insecure countries [8–10]. For producers, a timely forecast of yield can support the optimization of agronomic management [11,12]. At the same time, the agri-food industry is interested in yield forecasts to optimize food processing, storage, transport, and marketing [7,13–16].

Among the most popular technologies for yield prediction, remote sensing plays a pivotal role. Indeed, the diversity of sensors and the multiplicity of satellite missions enable the monitoring of the productivity of crops throughout the growing season at different spatial resolutions. Previously, satellite remote sensing could only provide yield predictions at a regional scale, owing to the low spatial resolution of instruments available with the frequent revisit time required (e.g., MODIS, SPOT-VGT). The arrival of a new generation of satellite constellations equipped with high spatial-resolution sensors and a frequent revisit time (e.g., Sentinel-1 and -2) has recently enabled within-field yield variability to be estimated [17–19]. Crop yield prediction methods can be classified into two groups: methods based on crop growth models and data-driven statistical methods. Growth model-based methods require many hard-to-obtain calibration parameters for large-scale applications and thus have a limited capacity in estimating actual yield at a regional scale [20,21], though they produce satisfactory results at the field scale when agronomic practices are known [22,23].

Statistical methods for yield predictions are typically based on linear regressions between vegetation indices (VIs) (e.g., NDVI; [24]) or meteorological data (or both) and yield data. As input for establishing the linear regressions, these methods can use values of VIs obtained at specific dates or with specific characteristics, such as the maximum value over the growing season of the considered VI [25,26]. In addition, the cumulative values of VIs during the growing season [27,28] can also be used as proxies for green biomass estimation [29,30]. Statistical methods based on satellite data are widely used in large-scale yield prediction. However, they are specific to the crops, phenological stages, and geographic regions in which they were calibrated [31,32]. A drawback of the methods based on linear regressions is that they fail to capture the complex interactions between environmental conditions and yield [33]. ML algorithms are interesting tools for facing this issue, as they have demonstrated reliable performance in estimating yield for a range of crops and environments [17,34–38]. A large set of ML algorithms have been used to perform regression for yield predictions. Among them, frequently used algorithms are the random forest, support vector machines, and neural networks [37,39]. ML algorithms can use heterogeneous information (i.e., derived from different sources- remote sensing, meteorological and soil data) to find non-linear relations with crop yield [33]. Additionally, the ML approach has been reported to be advantageous due to its ability to extract valuable information from a large number of predictors and because it does not require assumptions on data distributions [40,41].

However, in most cases, the statistical approaches were applied in retrospective mode (i.e., seeking to explain past yield realization with RS or meteorological predictors) and cannot be used to perform yield predictions. This is because the required predictors can only be calculated using data measured over the growing season [34,42,43], thus restricting predictions to end-of-season predictions. In only a few cases, they were explicitly designed to perform yield predictions without requiring data from the whole growing season, thus, being able to predict yield within the season [44]. Even fewer studies have focused on determining the optimal time window for yield prediction. Several ML approaches predict crop yield based on the aggregation of available data (remotely sensed predictors, meteorological data, etc.) within a fixed time scale, such as a month, a decade, or a week [38,45–49]. These approaches, however, do not consider the spatial variability among different phenological crop stages (e.g., the variability arising from fields sown at different dates) nor the temporal inter-annual variability of phenological development (e.g., anticipations and delays of the development due to different temperatures in different years), potentially reducing the performances of yield prediction. In this regard, recent experiments have shown that the dynamic extraction of phenological information based on raw data (e.g., from VIs time series) can reduce the problem of predictor heterogeneity and improve yield predictions [50,51]. For example, the correlation of LAI to maize yield in the USA and China was higher when LAI was calculated over a specific phenological phase than over a fixed period of time [52]. This suggests that the combination of time series of VIs and phenological information could improve yield prediction accuracy [53–55]. Bai et al. [56] pointed out that the duration of the phenological stage could be combined with the growth rate of the VIs or with the value of the indices themselves to improve crop yield prediction [57].

Typically, the ML workflow includes model definition and data reduction, hyperparameter optimization, and model testing [28,33]. Setting up the modeling framework of ML models poses practical difficulties. For example, the use of a large set of predictors can worsen dimensionality problems and lead to the introduction of autocorrelated and noninformative predictors within the algorithm, reducing the performances of the ML algorithms [33,58].

The main objective of this study was to develop an ML workflow to predict maize (*Zea mays* L.) yield at the field scale, using static soil predictors, VIs, and meteorological predictors dynamically aggregated by phenological stages. Specific objectives were: (i) to compare and to identify the most suitable ML configuration (i.e., combination of input feature set, data reduction option, and ML algorithm) to predict maize yield at field scale; (ii) to determine the effect of the phenological stage on the accuracy of maize yield prediction, (iii) to evaluate the effect of data reduction on each ML configuration; and (iv) to identify the most important features in the yield prediction process and evaluate the relative contribution of meteorological and soil data.

2. Materials

2.1. Study Area

This study was performed to support the supply chain management of maize in northern Italy. The fields of interest are in northwest Italy (Figure 1; 6.41–10.41°E, 44.13–46.05°N), where the majority of the Italian maize acreage is concentrated.

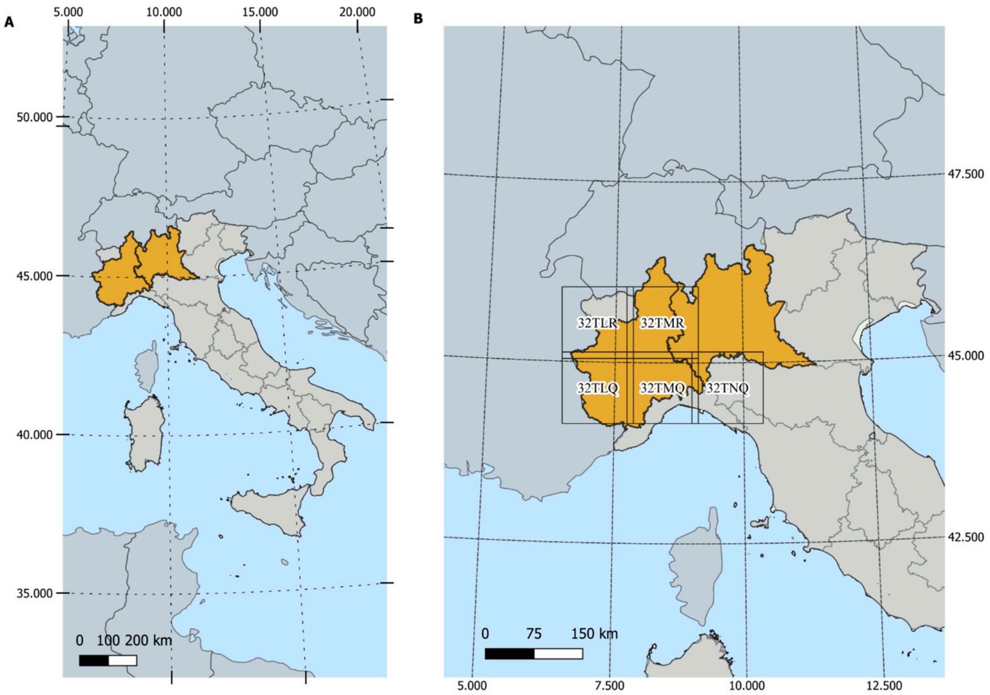

Figure 1. (**A**) Geographic location of the study area; (**B**) Extension of the five Sentinel-2 tiles used (32TLR, 32TMR, 32TLQ, 32TNQ).

Maize is usually sown from mid-March to the end of April, while harvesting is carried out at the end of summer (August to September). The area has a temperate subcontinental climate (Kottek et al., 2006). During the study period (2017–2020), the average annual rainfall was 817 mm, with 55% of the yearly precipitation occurring between March and September. In the March to July period, the average daily temperature was 14.5 °C, with temperatures ranging from 9.6 °C (minimum) to 24.7 °C (maximum) in March and July, respectively. The study area is characterized by high maize yields due to high technological inputs and irrigation without limiting factors.

2.2. Data Sources

We collected maize grain yield, meteorological data, soil information, and vegetation indices (VIs) calculated from remote sensing imagery (Table 1).

Table 1. Summary of the datasets used.

Category	Input	Acronym	Unit	Spatial Resolution	Temporal Resolution	Source
Crop	Yield		Mg ha^{-1}	Field level	Yearly	Producers
	Sowing	SD	date			
Meteorological	Air temperature (min, mean, max)	T_{min} T_{max} and T_{avg}	°C	25 × 25 km	Daily	AGRI4CAST
	Vapor pressure	VPD	hPa			
	Total global radiation	RAD	KJ m^{-2} d^{-1}			
	Sum of precipitation	cumPrec	mm d^{-1}			
	Potential evapotranspiration from a crop canopy	ET0	mm d^{-1}			
	Mean daily wind speed at 10 m heigh	WindSpeed	m s^{-1}			
Soil	Nitrogen (N)	N_mean	g kg^{-1}	500 × 500 m	Static	LUCAS
	Phosphorus (P)	P_mean	mg kg^{-1}			
	Potassium (K)	K_mean	mg kg^{-1}			
	Soil cation exchange capacity (CEC)	CEC_mean	mS m^{-1}			
	Carbon:nitrogen ratio (CN)	CN_mean				
	Calcium carbonate (caco$_3$)	CaCO$_3$_mean	g kg^{-1}			
	Soil texture USDA	Tess_mean	class			
Satellite	Normalized Difference Vegetation Index	NDVI		10 × 10 m	Approx. 2–3 days	ESA Copernicus
	Normalized Difference Red-Edge	NDRE				
	Normalized Difference Water Index	NDWI				
	Green Normalized Difference Vegetation Indices	GNDVI				

2.2.1. Yield and Crop Data

Maize yield data was collected over 340 fields based on producers' declarations. The average area of the considered fields was 3.35 ha, the smallest having an area of 0.5 ha while the largest having an area of 25 ha. The maize yield used in this study for each field was calculated as the ratio of the total crop production to the total harvest areas. Over the four years of the study (2017–2020), maize yields ranged from 7 to 18 Mg ha^{-1} (Figure S1). The hybrid planted belonged to five different FAO earliness classes, here aggregated into three classes: (i) "early" for FAO classes 300 and 400; (ii) "medium" for FAO class 500; and "late" for FAO classes 600 and 700. For each earliness class, a specific phenological table (Table S1) based on the accumulated growing degree days (AGDD) was applied to derive phenological stages. The different thresholds in AGDD adopted to determine the succession of the different phenological phases were those used in the IRRINET service [59]. More details of the procedure used for the phenological stage estimation were reported in Section 3.1. Field boundaries, sowing dates, and yield data are provided within the framework of a collaboration between the university and a private company.

2.2.2. Meteorological Data

Meteorological data at the field sites from 2017 to 2020 were obtained from gridded data provided by the EU Joint Research Centre, MARS-AGRI4CAST project [60]. Specifically, daily mean temperature (°C), daily maximum temperature (°C), daily minimum temperature (°C), precipitation sum (mm d^{-1}), vapor pressure (hPa), and total global radiation (KJ m^{-2} d^{-1}) for the entire growing season and at a resolution of 25 km^2 were used in this study.

2.2.3. Soil Data

Six soil properties, nitrogen (N), phosphorus (P), potassium (K), pH, soil cation exchange capacity (CEC), and calcium carbonate ($CaCO_3$) content, were extracted for each field from the LUCAS project maps [61]. Soil properties and nutrient content (nitrogen, phosphorus, and potassium) varied greatly across the studied area (Figure S2). Data are available at https://esdac.jrc.ec.europa.eu/content/lucas2015-topsoil-data (accessed on 15 July 2022).

2.2.4. Sentinel-2/MSI Datasets

Remote sensing data were provided by the Copernicus mission satellites Sentinel-2 A and B, launched in 2015 and 2017, respectively. Sentinel-2A/B MultiSpectral Imager (MSI) instruments capture images of the Earth's surface in 13 spectral bands at a spatial resolution of 10, 20, and 60 m [62], ranging from visible and near-infrared (VNIR) to shortwave infrared (SWIR). Sentinel-2 provides data every 10 days at the equator with one satellite and 5 days with 2 satellites, resulting in 2–3 days at mid-latitudes.

Sentinel-2 imagery was processed by applying the cloud-masking function (maskS2clouds) from the Google Earth Engine (GEE) [63]. Field boundaries were checked with a Sentinel-2 image in the middle of the growing season, and then a negative 10-m buffer was applied to avoid the edge effect. The average VI value of each field was then extracted using GEE. Four vegetation indices were used for the study: Normalized Difference Vegetation Index (NDVI [24], Green Normalized Difference Vegetation Index (GNDVI, [64]), Normalized Difference Red-Edge (NDRE, [65]), and Normalized Difference Water Index (NDWI, [66]). The NDRE, GNDVI, and NDWI were included in the analysis due to their ability to bring additional information on biomass growth dynamics and photosynthetic potential [67] and their reduced sensitivity to saturation issues at over-dense canopy as compared to NDVI, the most used vegetation index [6].

3. Methods

The proposed yield prediction method works at the pace of phenological timings. The growing season of each maize field was divided into six time periods by phenological stage using a threshold method based on AGDD (see Section 3.1). The forecasting events were then triggered at the end of each growth phase by adding the predictors' values during the last crop stage. Therefore, the number of predictors increased during the growing season when more and more stages were completed.

3.1. Phenological Stage Estimation

The main phenological stages were estimated using the accumulated growing degree days (AGDD) thresholds method for all the monitored fields. AGDD was calculated by summing the growing degree days (GDD) obtained each day starting from the sowing date. The GDD were calculated using Equation (1):

$$GDD = \left[\frac{(T_{max} - T_{min})}{2}\right] - B \quad (1)$$

where T_{max} and T_{min} represent the daily maximum and minimum temperatures, respectively, and B represents a base temperature value of 10 °C.

In this calculation, the following adjustments were made: (i) $T_{max} - T_{min}$ below the base temperature (10 °C) set GDD at 0, and (ii) mean temperature above 30 °C was set at 30 °C. This upper limit is normally applied in maize [68]. This study considered six phenological stages: three vegetative growth stages and three reproductive stages. The vegetative growth stages were emergence (V1), six leaf collars (V6), and tassel (VT), while the reproductive stages were silk (R1), dough (R4), and maturity (R6). The AGDD thresholds used for the estimation of the six phenological stages based on FAO classes (300–400, 500, and 600–700) are reported in Table S1. This method enables near-real-time (NRT) operational yield predicting, unlike remote sensing methods for phenological stage estimation, which typically require the full seasonal VI temporal trajectory to estimate key phenological timings. Upon reaching a given AGDD in a field, yield prediction at that time is triggered. That is, yield predictions are not issued at fixed times of the year but upon passing specific phenological stages that may happen at different times of the year in different fields and/or in different years. Figure 2 displays the trend and variability of the four vegetation indices for each phenological stage and earliness class.

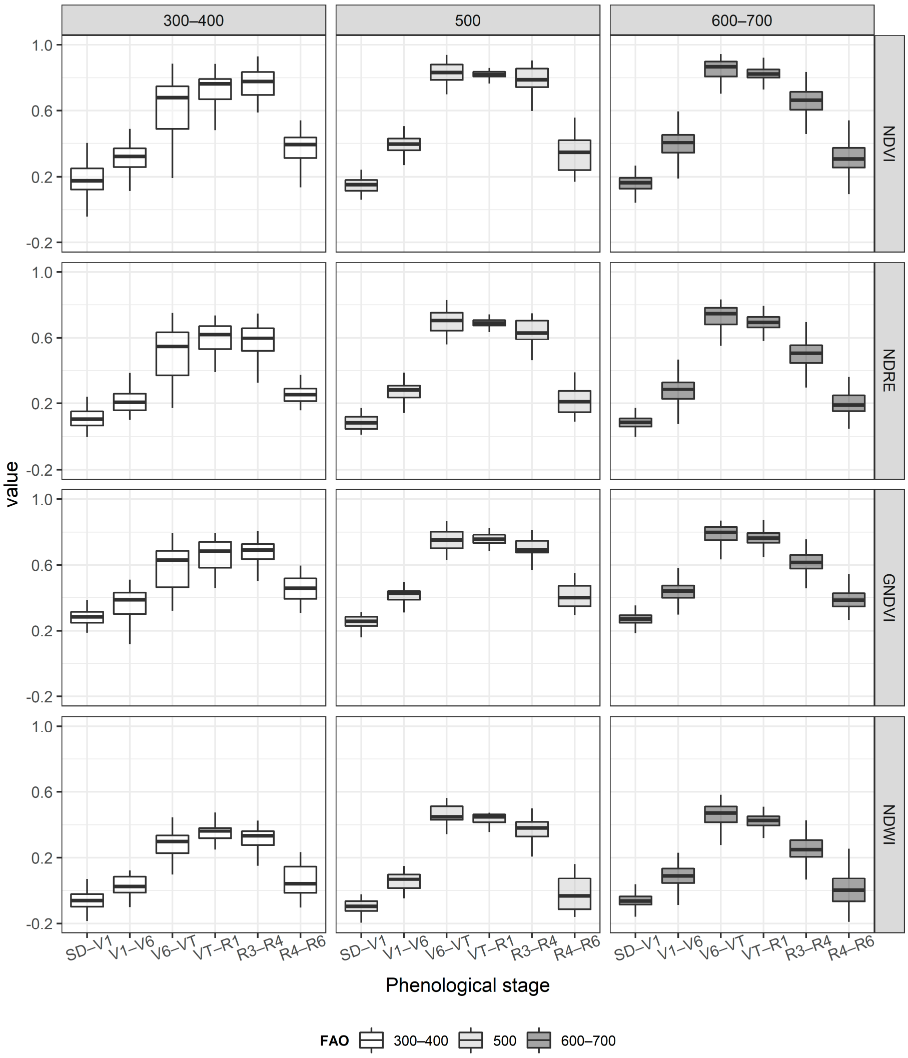

Figure 2. Distribution of average VIs (NDVI, NDRE, GNDVI, and NDWI) value by phenological phases for the three investigated earliness varieties (FAO classes reported). Stage abbreviation is as follows. SD: sowing; V1: emergence; V6: jointing-six leaf collars VT: tasselling; R1: Silk R4: dough; R6: maturity.

3.2. Feature Extraction by Phenological Stage

At each phenological stage, the time series of the four VIs calculated for each field were temporarily smoothed using a generalized additive model (GAM) to remove the outliers and regularize the time series and enabling fitting in a flexible way the trajectory of the vegetation index [69–71]. At the end of each phenological stage, the fitted values of the previous stage are updated. The GAM was fitted using the "mgcv" package in R [72]. The generic formula of a one-dimensional GAM is displayed in Equation (2):

$$\mathrm{VI}(t) = \beta_0 + s(t) + \varepsilon_t \tag{2}$$

where VI is the observed vegetation index value at time t, B_0 is the intercept of the function at t = 0, s represents a smoothing function of covariate t (in this case DOY), and ε_t is a random error term with $\varepsilon_t \sim N\ (0, \sigma^2)$. The smoothing function is defined by Equation (3):

$$s(t) = \sum_{k=1}^{k} \beta_k b_k(t) \tag{3}$$

where the final smoothing function s(t) is the sum of all K basis functions $b_k(t)$ multiplied by its corresponding weight β_k. For each phenological stage, two types of predictors were extracted and used as input predictors to predict observed maize yields: (i) the rate of growth calculated as the mean of the first derivative of each phenological stage (fd_, e.g., fd.NDVI_VT-R1), and (ii) the mean values of each selected vegetation index (i.e., NDVI, GNDVI, NDRE, and NDWI) in a specific interval of the phenological stage (e.g., NDVI_VT-R1). Together, these two types of predictors enabled consideration of both biomass level and growth rate in the period between two phenological stages. As with satellite data, meteorological data were aggregated according to the phenological stage. Rainfall data were cumulated for each phenological stage (e.g., cumPrec_VT-R1), while the average values for all other meteorological predictors were calculated for each phenological stage (e.g., Tmax_VT-R1, Tmin_VT-R1, TAvg_VT-R1, VPD_VT-R1, ET0_VT-R1).

3.3. Model Configurations

Pre-processing is an important step in machine learning workflows. It can improve prediction accuracy by discarding irrelevant features and speeding up model training [33]. Two pre-processing techniques were tested in the modeling framework: either manually defining input feature sets or automatically reducing the dimensionality. First, from the full set of available satellite, meteorological and soil predictors (named Vis + S + M), some relevant subsets were defined: satellite data only (VIs), satellite data and meteorological data (Vis + M), and satellite data and soil data (VIs + S). Second, an additional step was carried out via principal component analysis (PCA), retaining the principal component (PCs) up to 95% explained variance of the entire set of predictors. PCA reduces the noise effect caused by data redundancy and multicollinearity. In addition, compared to feature selection techniques, PCA is not limited to reducing collinearity/multicollinearity and redundant information, as it can also extract new predictors [73]. In summary, at each phenological stage, all ML algorithms are trained on four input feature sets, with and without PCA (the predictors are always centered and scaled), for a total of 8 different combinations for each phenological stage.

3.4. Cross-Validation Strategy

In total, three independent data sets were required to determine both hyper-parameters and model performance: the training set used to train the model, the validation set used to optimize the hyper-parameter set, and the test set used to estimate model performance. Similarly to [33], a nested leave-one-year-out cross-validation was adopted. This cross-validation strategy required splitting the data into an outer cross-validation loop and an inner cross-validation loop. In the outer cross-validation loop, data from one year of the n available year (n = 4) were held out. The remaining 3 years were used in an inner cross-validation loop. In the inner cross-validation loop, data from one year of the 3 available are held out one at a time for validation. During each inner cross-validation loop, model hyper-parameters were selected using the remaining 2 years. After identifying the optimal hyper-parameters, the best model obtained was used to predict the group's performance in the outer loop by comparing the predicted and actual values (i.e., the model test). This procedure is repeated for all 4 years to evaluate the model's prediction performance. Following this approach, four versions of the model (i.e., same architecture but different hyper-parameters and coefficients at each iteration possible) were trained: one for each of the n outer loops. Thus, different hyper-parameters and coefficients of the model can be selected in each iteration of the outer loop. The mean value and the standard deviation obtained from the different outer loops for each feature set, ML algorithm, and phenological stage were calculated for each evaluation metric.

3.5. Machine Learning Algorithms

In this study, seven machine learning algorithms were selected and compared: random forest (RF, [74]), cubist (CUB, Quinlan, 1992), single-layer perceptron feedforward neural networks (NNET, [75]), support vector regression with linear and radial basis function kernels (SVMl and SVMr, [76]), Gaussian process regression (GPR, [77]) and k-nearest neighbors (kNN, [78]). Algorithms were implemented using the R package "caret" [79]. Each model has a set of hyper-parameters to be optimized. Here we used systematic grid search optimization, i.e., testing a grid of empirically chosen candidates. A cross-validation strategy was used to determine the best hyper-parameters and evaluate each ML configuration properly, avoiding over-fitting issues [80,81]. A brief description of each algorithm and its hyper-parameters is presented in Supplementary Table S2.

3.6. Predictor Variables of Maize Yield

Predictor importance was assessed by repeatedly permuting the values of a predictor and examining how model performance changed (i.e., permutational importance of predictors) [82]. The more the model's performance decreases, the more important the predictor. Specifically, variable importance was calculated by evaluating the increase in Root Mean Square Error (RMSE) of prediction when shuffling the predictors within a dataset. The predictor importance was calculated using the DALEX package. To demonstrate the uncertainty of the predictor's importance estimation, the predictor importance was calculated for 10 permutations [83].

3.7. Model Evaluation

To evaluate the performance of each ML configuration, the RMSE, the coefficient of determination (R^2), the normalized RMSE (nRMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) were calculated using the reported Equations (4)–(8):

$$\mathrm{R}^2 = \frac{\left(\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\overline{\mathrm{y}}_\mathrm{i})\left(\mathrm{f_i}-\bar{\mathrm{f}}_\mathrm{i}\right)\right)^2}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\overline{\mathrm{y}}_\mathrm{i})^2\ \sum_{\mathrm{i}=1}^{\mathrm{n}}\left(\mathrm{f_i}-\bar{\mathrm{f}}_\mathrm{i}\right)^2} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\mathrm{f_i})^2} \tag{5}$$

$$\mathrm{nRMSE}\ (\%) = \frac{\mathrm{RMSE}}{\overline{\mathrm{y}}_\mathrm{i}}\cdot 100 \tag{6}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|\mathrm{y_i}-\mathrm{f_i}|}{n} \tag{7}$$

$$\mathrm{MAPE} = \frac{\sum_{i=1}^{n}\frac{|\mathrm{y_i}-\mathrm{f_i}|}{\mathrm{y_i}}\cdot 100}{n} \tag{8}$$

where n (i = 1, 2, . . . , n) is the number of samples used to test the ML model, y_i is the observed yield, $\overline{y}_i$ is the corresponding mean value, f_i is the predicted yield and $\overline{f}_i$ is the corresponding mean value. The closer the R^2 is to 1, the higher the model's prediction performance. Small nRMSE (%) and RMSE values indicate less discrepancy within the observed and predicted yield.

4. Results

In total, seven machine learning models were trained with eight ML configurations. The prediction performance was evaluated using the R^2, RMSE, and nRMSE calculated from a test dataset's predicted and observed yield. The results of the leave-one-year-out cross-validation were summarized for each ML configuration (Table S2 and Figure 3).

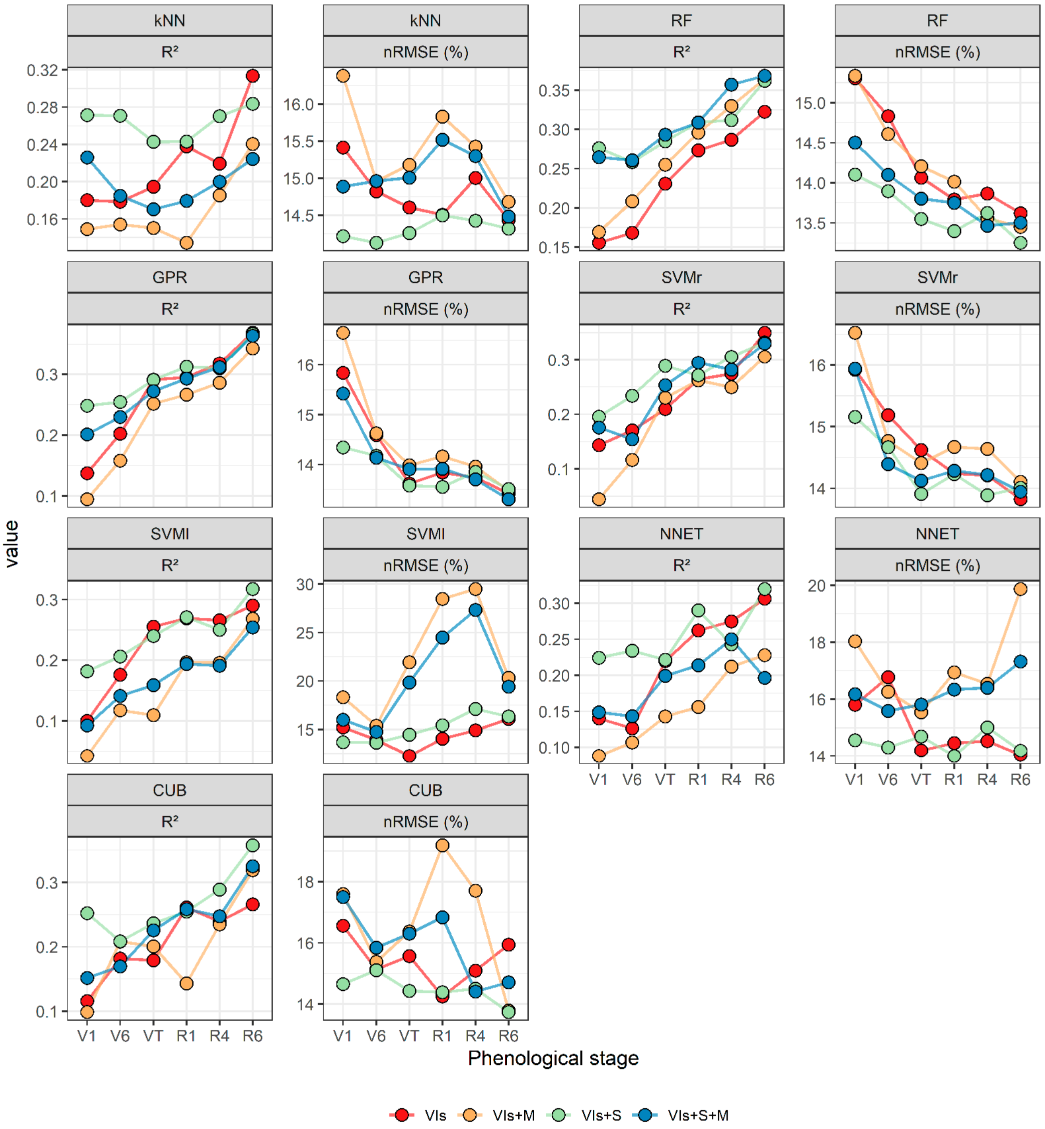

Figure 3. Model performance at six phenological stages for four feature sets (VIs, VIs + M, VIs + S, VIs + S + M) using the seven ML algorithms (kNN, SVMr, RF, CUB, GPR, SVMl, and NNET) and without principal component analysis (PCA) data reduction. The filled point represents the mean values of R^2 and nRMSE (%).

4.1. Performance of Yield Predictions

Figure 3 displays the temporal evolution of the cross-validated performance of the various model configurations (i.e., best-performing configuration per ML algorithm and by predictor set) as more and more information is made available through the growing season. In general, the performance improved as the prediction period approached the end of the growing season (i.e., R^2 gradually increased with time, while nRMSE and RMSE decreased). Focusing on the configurations based on only vegetation indices (VIs) without the application of a PCA (Figure 4), it was observed that the best model at the maturity stage (R6) was the GPR model with an RMSE of 1.80 Mg ha^{-1}, an R^2 of 0.37 and an nRMSE of 13.42% (Table S3). RF and SVMr follow the GPR model, respectively, with an RMSE of 1.83 and 1.85 Mg ha^{-1} and an nRMSE of 13.62% and 13.83%. The performance of the GPR model greatly increased during the initial phenological stages (from V1 to VT) with an R^2 from 0.14 (V1) to 0.29 (VT), and nRMSE decreased from 15.84% (V1) to 13.62% (VT).

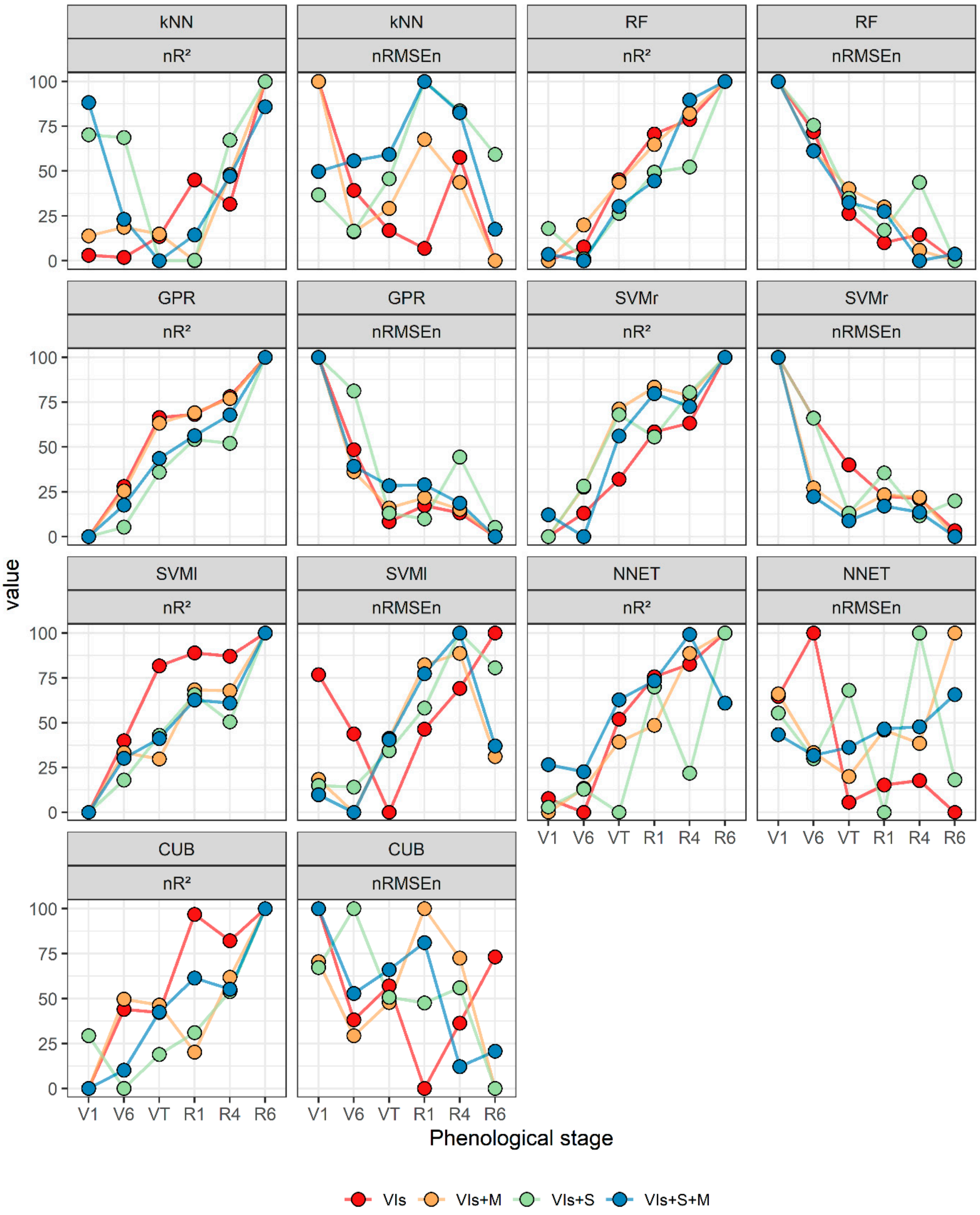

Figure 4. Normalized values of R^2 and nRMSE (nR^2 and nRMSEn) at six phenological stages for four feature sets (VIs, VIs + M, VIs + S, VIs + S + M) using the seven ML algorithms (kNN, SVMr, RF, CUB, GPR, SVMl, and NNET).

After this initial large increase, the GPR performance kept increasing at a lower rate until the introduction of the variables related to R1 (milk stage). Subsequently, different behaviors were observed depending on the ML algorithm considered; for SVMl and NNET, performance dropped in terms of both RMSE and R^2. In particular, SVMl achieved its best performance at VT (RMSE of 1.90 Mg ha^{-1} and nRMSE of 14.26%) and its worst at R6 (RMSE of 2.41 Mg ha^{-1} and an nRMSE of 18.11%). The lowest RMSE of SVMl models at VT was comparable to that of CUB at the R1 stage and to that of RF, GPR, and SVMr at R4 (Table S3). Overall, for the best models (i.e., RF, GPR, and SVMr), the inclusion of meteorological variables (VIs + M), compared to the use of VIs only, did not improve the prediction performance of the ML algorithms in terms of RMSE and nRMSE (%) but in terms of R^2. With this feature set, GPR and RF remain the two best ML algorithms at R6, with an RMSE of 1.80 Mg ha^{-1} for both and an nRMSE of 13.48% and 13.45%, respectively. The addition of soil data (VIs + S) to the VIs increased R^2 at the V1 (Figure 3). With this feature set (VIs + S), the best ML algorithms at the R6 were the RF and GPR (an RMSE of 1.78 Mg ha^{-1} and 1.80 Mg ha^{-1}, an nRMSE of 13.25% and 13.51%, respectively). At the R6, GPR, RF, and SVMr were the best ML algorithms (nRMSE of 13.31%, 13.50%, and 13.95%, respectively). The addition of both meteorological and soil data to the VIs improved the performance of the maize yield prediction, particularly early in the season (e.g., V1 and V6). In contrast, after VT, neural networks and SVMl decreased their performance during the season when meteorological predictors were considered.

4.2. Influence of Lead Time on Yield Prediction Performance

In order to reveal the magnitude of the changes observed for each model and disregard the different levels of performances of different models, performance from V1 to R6 (maturity stage), the R^2 and nRMSE were normalized from 0% to 100% using a min-max approach (expressed as nR^2 and nRMSEn) [51]. In general, the greatest increase in model performances occurred during V1 to VT (Figure 4). At the same time, after VT, there was only a slight improvement (SVMr, RF, and GPR) or even a reduction in performance (SVMl and NNET). In particular, using the full feature set (Vis + M + S), analyzing the nRMSEn, the optimal phenological stage for maize yield predictions was V6 for NNET and SVMl, and R4 for other ML algorithms.

4.3. Effect of Data Reduction (PCA)

To study the effect of dimensionality reduction, PCA was applied to each ML configuration (combinations of ML algorithms, feature set, and phenological stage) (Figure 5 and Table S4). The application of PCA showed contradictory results: in some cases, performance decreased, while in other cases, it increased. For example, when meteorological data were included in the predictors, the application of PCA improved performance, especially at the beginning of the growing season. When PCA was applied to SVMr, performance decreased as the season progressed after reaching the R1 stage. In contrast, SVMl and NNET showed a gradual increase in performance during the growing season, but only when PCA was applied, especially for the feature set with meteorological data (Vis + M). Overall, without PCA, the performance of SVMl and NNET decreased from the V6 stage to the maturity stage (R6). At the maturity stage, when PCA was applied, SVMl achieved the best overall performance in terms of RMSE 1.77 Mg Ha^{-1}, followed by RF and NNET at the same phenological stage.

An additional benefit of the application of PCA was the overall reduction in the standard deviation of the RMSE (on average by 30.20% (VIs + M), 12.12% (VIs + S), and 29.62% (VIs + S + M) calculated from the outer-loop cross-validation at the maturity stage (R6). The most significant reduction in the standard deviation was observed in the NNET and SVMl with the full feature set (VIs + S + M) at the dough stage (R4) and at the maturity stage (R6), respectively (84.72% and 72.30%) (Figure 6 and Table S4).

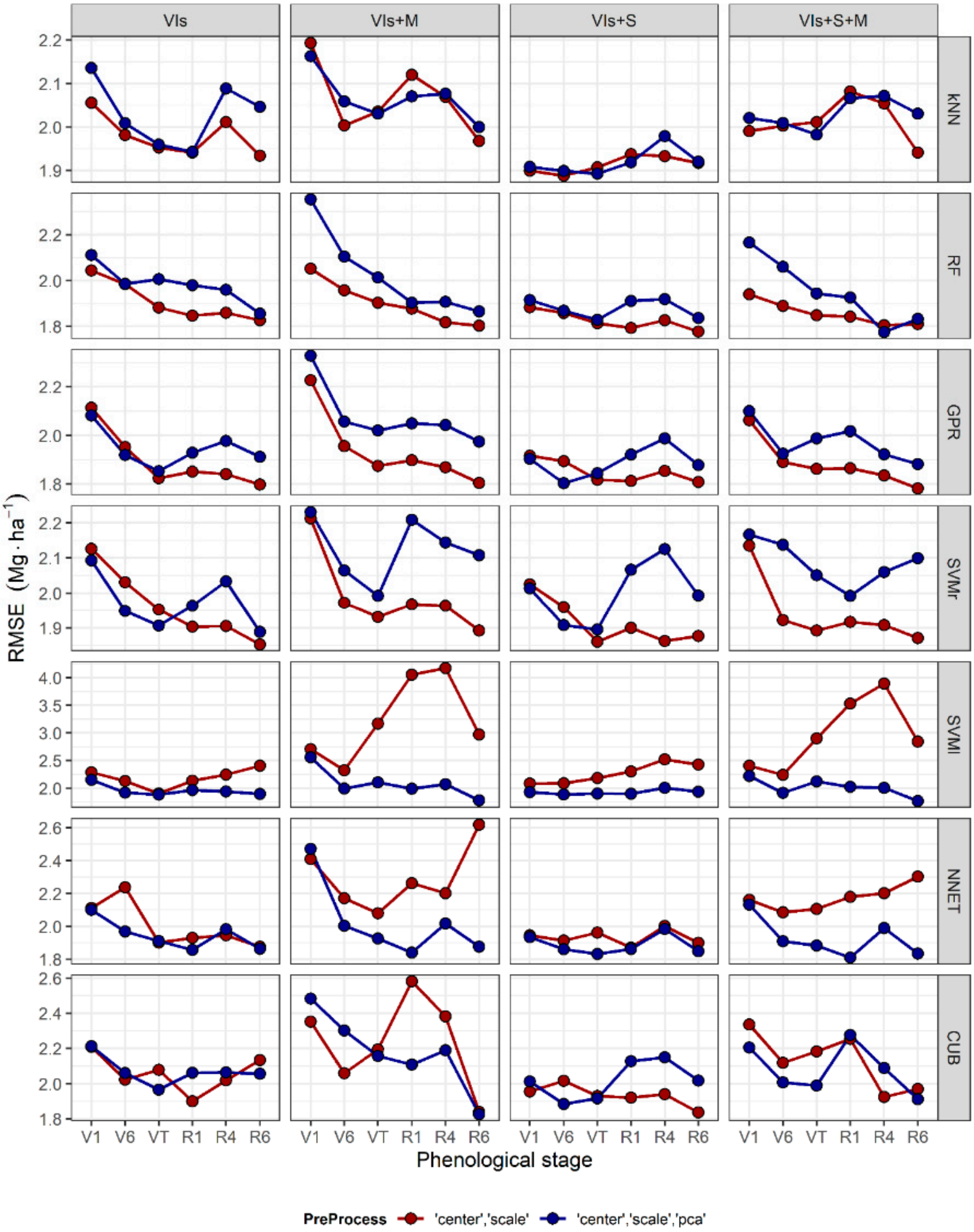

Figure 5. Impact of feature sets and data reduction on yield prediction performance (RMSE) at six phenological stages for four feature sets (VIs, VIs + M, VIs + S, VIs + S + M) and data reduction PCA) using the seven ML algorithms (kNN, SVMr, RF, CUB, GPR, SVMl, and NNET).

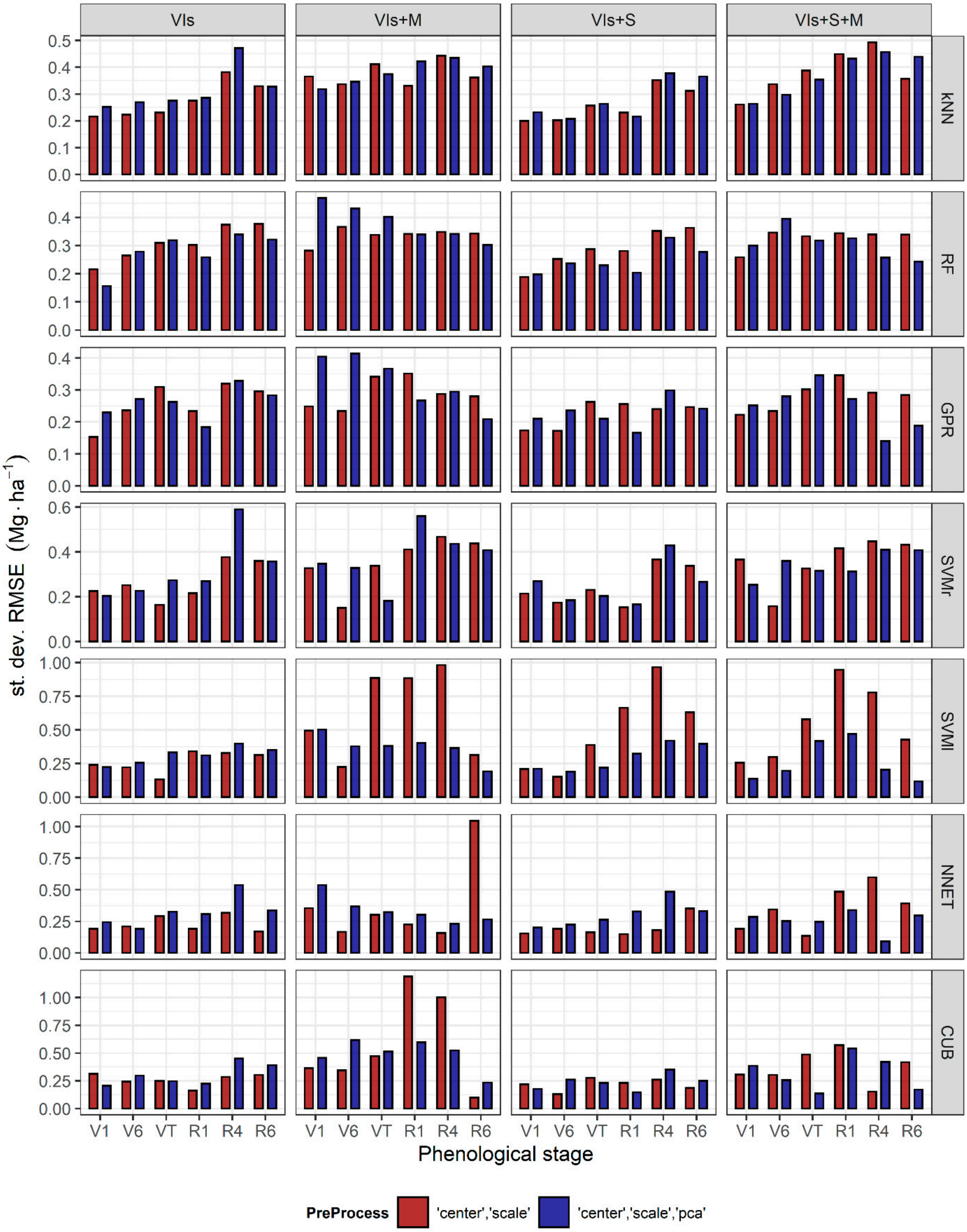

Figure 6. Summary of the standard deviation outer loop cross-validation (nested cross-validation) after the application of PCA.

4.4. Informative Predictor Variables

In order to study the importance of the predictors, the full feature set configuration (VIs + M + S) at the maturity stage was analyzed without PCA dimensionality reduction. Overall, the VIs-based predictors were often among the ten most important predictors. For

CUB, GPR, RF, and SVMr, at least six predictors out of ten were based on VIs, and several of these predictors were calculated using NDWI (Figure 7). Only NNET had the majority of the predictors based on meteorological data. CUB, RF, and SVMr had more than 50% of the predictors collected before the R1 stage.

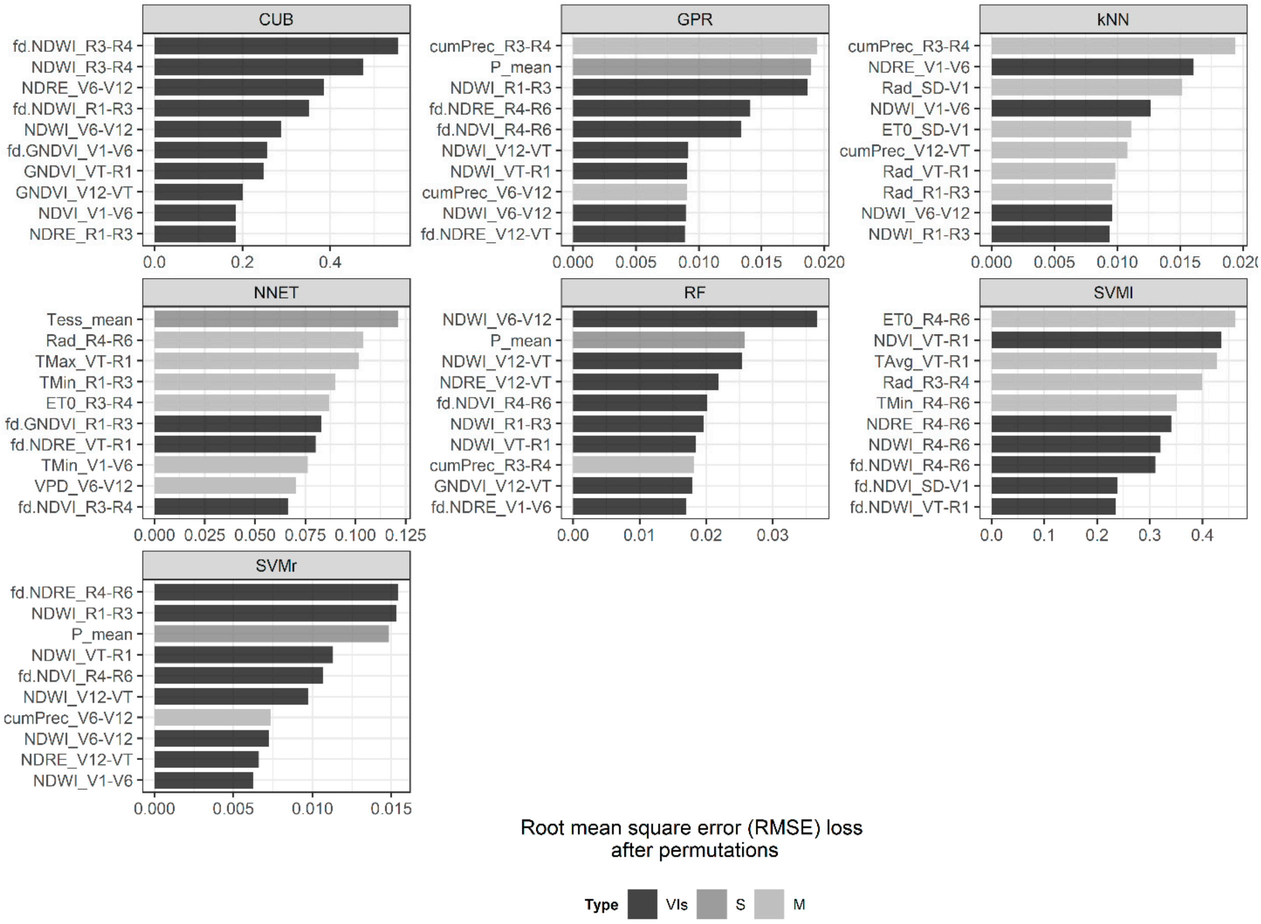

Figure 7. The variable importance of each ML algorithm (with the full feature set, VIs + S + M) for maize yield prediction at R6 (maturity stage) was expressed by the drop-out loss of model performance for each predictor related to the drop-out loss of the full model. The acronym "fd._" indicates the growth rate of a specific vegetation index in a specific phenological stage.

5. Discussion

This study proposed to identify an accurate modeling framework for maize yield prediction using satellite, meteorological, and soil data. The accuracies of yield prediction of seven ML algorithms (CUB, GPR, kNN, NNET, RF, SVMl, and SVMr) were evaluated at six phenological stages (V1, V6, VT, R1, R4, and R6) testing four feature sets (VIs, VIs + M, VIs + S, and VIs + M + S) and comparing the performances achieved with and without data reduction using PCA. GPR, RF, and SVMr were the best ML algorithms at R6 with the full feature set (VIs + S + M), showing high performance in terms of nRMSE (13.31%, 13.50%, and 13.95%), while SVMl and NNET consistently ranked last at this stage. The good performance obtained with SVMr was also reported in a study on wheat yield prediction in Australia by Kamir et al. [35]. The authors explained that SVMr is superior to other algorithms because it provides maximum generalization based on risk minimization. This minimizes an upper bound of the generalization error rather than minimizing the training error, thus achieving better generalization and transferability performance. In fact, SVMr minimizes an upper bound on error generalization rather than minimizing the training error. However, in this study, GPR and RF outperformed SVMr. This result could be explained by the fact that both the GPR and RF have their own aptitude for selecting and weighing the most important variables, making them robust in cases where there are several autocorrelated or noninformative variables [84,85]. Only slight differences were observed across the compared algorithms. This highlights that the main factors that influence the predictive performance of the algorithms are the predictors with which the ML algorithm is fed and the phenological stage at which the yield prediction is carried out.

Depending on the purpose of the yield prediction, different levels of accuracy and timing at which prediction should be available are required. In precision agriculture, for example, it is essential to make predictions during the growing season to ensure that information is available early enough for the timely implementation of agronomic decisions, such as fertilization, irrigation, and phytosanitary treatments [11,12]. Previous studies identified one to two months before harvest as the optimal prediction time for maize yield prediction [86,87]. In this study, the predicting time providing the largest accuracy varied depending on the ML algorithms used. With SVMl and NNET algorithms, the optimal prediction time was at V6 and VT, whereas, for the other algorithms (SVMr, RF, GPR, CUB, kNN), the optimal phenological stage was R4. The V6 and VT were also found to be the optimal phenological stages for maize yield prediction in other environments [19,88,89]. In agreement with Li et al. [86], the results presented in this work show that the maize yield prediction performance significantly increases until VT. Then it remains stable or has only a slight increase during the rest of the season, most likely because, at VT, the canopy is fully developed (maximum LAI) and has reached its maximum capacity to intercept radiation [90].

More than half of the ten most important predictors (Figure 7) are derived from satellite information. Although the RMSE loss after permutations of the individual variables for the best algorithm (GPR) is relatively small, the absence of predictors before the V6 stage reflects the high uncertainty in predicting final yields before that phenological stage.

The slight increase in performance that occurred after VT (Figure 4) seems to be in contradiction with the analysis of the importance of the predictors. In fact, among the most important predictors were several meteorological predictors after VT within the ten most important predictors, such as the cumulative rainfall in the period between R3 and R4 (cumPrec_R3-R4). This could mean that the best ML algorithm (GPR) correctly interpreted the importance of rainfall at the end of the growing season in predicting actual yield despite performance not improving after VT. Regarding SVMl and NNET, they reached their best performance at V6 and VT, and their performances declined after these phenological stages. This trend was likely due to data dimensionality and collinearity, depending on the feature set and the phenological stage considered. The decline in performance following the introduction of phenological stage-related variables at R4 becomes more evident as the feature set size increases. The NNET displayed the greatest decrease in performance, possibly due to the excessive number of parameters added to the model [58]. Other studies apply dimensionality reduction techniques, such as principal component analysis (PCA), to enhance model performance. In this study, the most significant impact of data reduction using PCA was mainly observed for the full feature set and with the SVMl and NNET algorithms. On the contrary, SVMr, CUB, RF, and GPR were negatively affected by PCA, highlighting how these algorithms are more resistant to redundant variables and can extract relevant features more efficiently than PCA, confirming the observations of Meroni et al. [33]. In addition, the growth rate of NDWI between the R4–R6 stages was among the ten most important variables, which seems to confirm that the senescence rate is correlated with yield, as reported by Ji et al. [44]. The relevance of the NDWI and NDRE measured from V6 to VT confirms the importance of these vegetative growth stages, when stems and leaves grow vigorously and continuously accumulate nitrogen [91]. Among the meteorological variables, cumulative precipitation from the V6 stage through the end of the season (V6-VT, R4-R6) was among the most important predictors. This is consistent with other studies that demonstrated a high correlation between precipitation and yield [92]. In fact, many studies have found that during the tasseling stage, maize is highly drought-sensitive [93]. The inclusion of soil predictors, albeit at coarse resolution, resulted in improved predictive capabilities, indicating how they provide complementary information to the time series of vegetation indices, especially early in the growing season. However, it is important to note that the importance of predictors may vary by analyzing earlier phenological stages.

6. Conclusions

This study developed an ML modeling framework for maize yield prediction using multi-source data (vegetation indices, soil, and meteorological data, VIs + S + M). This ML modeling framework was used to identify the best-performing ML configuration and optimal lead time to provide different agri-food stakeholders (i.e., producers and processing industries) with yield predictions. The main factors that influence the predictive performance of the ML algorithms are the predictors with which the ML algorithm is fed and the phenological stage at which the yield prediction is carried out. Overall, GPR and RF were the best algorithms for predicting maize yield, and their best performance was achieved late-season (R6). The NNET achieved similar accuracy to GPR and RF but earlier in the season (VT-R1). However, after VT, it was observed that the performance of NNET decreased if data reduction (principal component analysis, PCA) was not applied, possibly due to the introduction of noninformative and autocorrelated predictors and an increase in the excessive number of parameters added to the neural network. This decrease in performance as the season progressed was not observed when applying PCA. Furthermore, the application of PCA reduces the standard deviation of maize yield prediction performance from outer-loop cross-validation. Therefore, based on the results presented in this study, two different approaches can be recommended based on the stakeholder. The first, applicable when the main stakeholder is represented by the producer itself, who needs to predict yield with perhaps lower accuracy but earlier in the growing season, includes neural networks at the VT stage. The second, applicable when the main stakeholder is represented by the industry, which wants greater accuracy and can wait later in the growing season than the manufacturer, involves the application of RF at the R6 stage.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/rs15010100/s1, Figure S1: Frequency distribution of measured maize yield during 2017–2020; Figure S2: Frequency distribution of the macro-element of the soil during 2017–2020; Table S1: Accumulated growing degree days (AGDD) thresholds used for the estimation of the six phenological stages based on FAO classes (300–400, 500 and 600–700); Table S2: Algorithm hyper-parameter tuned, and the range of values tested for each of them. N is the total number of hyper-parameter combinations tested for each algorithm; Table S3: Summary of the machine learning algorithms' performances (R^2, RMSE, and nRMSE); Table S4: Summary of the percentage change of the standard deviation outer loop cross-validation (nested cross-validation) after the application of PCA.

Author Contributions: M.C.: conceptualization, data curation, methodology, investigation, formal analysis, writing—original draft; G.I.: review and editing; M.M.: methodology, review, and editing; S.A.: conceptualization, review, and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this work are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Population Prospects—Population Division—United Nations. Available online: https://www.un.org/development/desa/pd/ (accessed on 20 July 2022).
2. IPCC. 2021Global Warming of 1.5 °C. In *Special Report Intergovernmental Panel on Climate Change*; IPCC: Geneva, Switzerland, 2021.

3. Rounsevell, M.D.A.; Ewert, F.; Reginster, I.; Leemans, R.; Carter, T.R. Future Scenarios of European Agricultural Land Use: {II}. Projecting Changes in Cropland and Grassland. *Agric. Ecosyst. Environ* **2005**, *107*, 117–135. [CrossRef]
4. Godfray, H.C.J.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Pretty, J.; Robinson, S.; Thomas, S.M.; Toulmin, C. Food Security: The Challenge of Feeding 9 Billion People. *Science* **2010**, *327*, 812–818. [CrossRef]
5. Lezoche, M.; Hernandez, J.E.; del Alemany Díaz, M.M.E.; Panetto, H.; Kacprzyk, J. Agri-Food 4.0: A Survey of the Supply Chains and Technologies for the Future Agriculture. *Comput. Ind.* **2020**, *117*, 103187. [CrossRef]
6. Schauberger, B.; Jägermeyr, J.; Gornott, C. A Systematic Review of Local to Regional Yield Forecasting Approaches and Frequently Used Data Resources. *Eur. J. Agron.* **2020**, *120*, 126153. [CrossRef]
7. Liu, J.; Shang, J.; Qian, B.; Huffman, T.; Zhang, Y.; Dong, T.; Jing, Q.; Martin, T. Crop Yield Estimation Using Time-Series {MODIS} Data and the Effects of Cropland Masks in Ontario, Canada. *Remote Sens.* **2019**, *11*, 2419. [CrossRef]
8. Basso, B.; Liu, L. Seasonal Crop Yield Forecast: Methods, Applications, and Accuracies. In *Advances in Agronomy*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 201–255.
9. Ben-Ari, T.; Boé, J.; Ciais, P.; Lecerf, R.; der Velde, M.; Makowski, D. Causes and Implications of the Unforeseen 2016 Extreme Yield Loss in the Breadbasket of France. *Nat. Commun.* **2018**, *9*, 1627. [CrossRef]
10. Funk, C.; Shukla, S.; Thiaw, W.M.; Rowland, J.; Hoell, A.; McNally, A.; Husak, G.; Novella, N.; Budde, M.; Peters-Lidard, C.; et al. Recognizing the Famine Early Warning Systems Network: Over 30 Years of Drought Early Warning Science Advances and Partnerships Promoting Global Food Security. *Bull. Am. Meteorol. Soc.* **2019**, *100*, 1011–1027. [CrossRef]
11. Rejeb, A.; Rejeb, K.; Zailani, S. Big Data for Sustainable Agri-food Supply Chains: A Review and Future Research Perspectives. *J. Data Inf. Manag.* **2021**, *3*, 167–182. [CrossRef]
12. Wolfert, S.; Ge, L.; Verdouw, C.; Bogaardt, M.-J. Big Data in Smart Farming—A Review. *Agric. Syst.* **2017**, *153*, 69–80. [CrossRef]
13. Mundi, I.; Alemany, M.M.E.; Poler, R.; Fuertes-Miquel, V.S. Review of Mathematical Models for Production Planning under Uncertainty Due to Lack of Homogeneity: Proposal of a Conceptual Model. *Int. J. Prod. Res.* **2019**, *57*, 5239–5283. [CrossRef]
14. Esteso, M.M.E.; Alemany, A. Deterministic and Uncertain Methods and Models for Managing Agri-Food Supply Chain. *Dir. Organ.* **2017**, *62*, 41–46.
15. Esteso, A.; Alemany, M.M.E.; Ortiz, A. Conceptual Framework for Designing Agri-Food Supply Chains under Uncertainty by Mathematical Programming Models. *Int. J. Prod. Res.* **2018**, *56*, 4418–4446. [CrossRef]
16. Mondino, P.; Gonzalez-Andujar, J.L. Evaluation of a Decision Support System for Crop Protection in Apple Orchards. *Comput. Ind.* **2019**, *107*, 99–103. [CrossRef]
17. Shahhosseini, M.; Hu, G.; Archontoulis, S.V. Forecasting Corn Yield With Machine Learning Ensembles. *Front. Plant Sci.* **2020**, *11*, 1120. [CrossRef] [PubMed]
18. Khaki, S.; Pham, H.; Wang, L. Simultaneous Corn and Soybean Yield Prediction from Remote Sensing Data Using Deep Transfer Learning. *Sci. Rep.* **2021**, *11*, 11132. [CrossRef]
19. Kayad, A.; Sozzi, M.; Gatto, S.; Marinello, F.; Pirotti, F. Monitoring Within-Field Variability of Corn Yield Using Sentinel-2 and Machine Learning Techniques. *Remote Sens.* **2019**, *11*, 2873. [CrossRef]
20. Bazgeer, S.; Kamali, G.; Mortazavi, A. *Wheat Yield Prediction through Agrometeorological Indices for Hamedan, Iran*; Desert (Biabian): Riyadh, Saudi Arabia, 2007.
21. Palosuo, T.; Kersebaum, K.C.; Angulo, C.; Hlavinka, P.; Moriondo, M.; Olesen, J.E.; Patil, R.H.; Ruget, F.; Rumbaur, C.; Takáč, J.; et al. Simulation of Winter Wheat Yield and Its Variability in Different Climates of Europe: A Comparison of Eight Crop Growth Models. *Eur. J. Agron.* **2011**, *35*, 103–114. [CrossRef]
22. Elavarasan, D.; Vincent, D.R.; Sharma, V.; Zomaya, A.Y.; Srinivasan, K. Forecasting Yield by Integrating Agrarian Factors and Machine Learning Models: A Survey. *Comput. Electron. Agric.* **2018**, *155*, 257–282. [CrossRef]
23. Hochman, Z.; van Rees, H.; Carberry, P.S.; Hunt, J.R.; McCown, R.L.; Gartmann, A.; Holzworth, D.; van Rees, S.; Dalgliesh, N.P.; Long, W.; et al. Re-Inventing Model-Based Decision Support with Australian Dryland Farmers. 4. Yield Prophet® helps Farmers Monitor and Manage Crops in a Variable Climate. *Crop Pasture Sci.* **2009**, *60*, 1057. [CrossRef]
24. Rouse, J.W.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring Vegetation Systems in the Great Plains with ERTS. In Proceedings of the Third ERTS Symposium, NASA SP-351, Washington DC, USA, 10–14 December 1973; pp. 309–317.
25. Becker-Reshef, I.; Vermote, E.; Lindeman, M.; Justice, C. A Generalized Regression-Based Model for Forecasting Winter Wheat Yields in Kansas and Ukraine Using {MODIS} Data. *Remote Sens. Environ.* **2010**, *114*, 1312–1323. [CrossRef]
26. Franch, B.; Vermote, E.F.; Becker-Reshef, I.; Claverie, M.; Huang, J.; Zhang, J.; Justice, C.; Sobrino, J.A. Improving the Timeliness of Winter Wheat Production Forecast in the United States of America, Ukraine and China Using {MODIS} Data and {NCAR} Growing Degree Day Information. *Remote Sens. Environ.* **2015**, *161*, 131–148. [CrossRef]
27. López-Lozano, R.; Duveiller, G.; Seguini, L.; Meroni, M.; García-Condado, S.; Hooker, J.; Leo, O.; Baruth, B. Towards Regional Grain Yield Forecasting with 1km-Resolution EO Biophysical Products: Strengths and Limitations at Pan-European Level. *Agric. For. Meteorol.* **2015**, *206*, 12–32. [CrossRef]
28. Meroni, M.; Marinho, E.; Sghaier, N.; Verstrate, M.; Leo, O. Remote Sensing Based Yield Estimation in a Stochastic Framework—Case Study of Durum Wheat in Tunisia. *Remote Sens.* **2013**, *5*, 539–557. [CrossRef]
29. Battude, M.; al Bitar, A.; Morin, D.; Cros, J.; Huc, M.; Marais Sicre, C.; le Dantec, V.; Demarez, V. Estimating Maize Biomass and Yield over Large Areas Using High Spatial and Temporal Resolution Sentinel-2 like Remote Sensing Data. *Remote Sens. Environ.* **2016**, *184*, 668–681. [CrossRef]

30. Waldner, F.; Horan, H.; Chen, Y.; Hochman, Z. High Temporal Resolution of Leaf Area Data Improves Empirical Estimation of Grain Yield. *Sci. Rep.* **2019**, *9*, 15714. [CrossRef]
31. Doraiswamy, P. Crop Condition and Yield Simulations Using Landsat and MODIS. *Remote Sens. Environ.* **2004**, *92*, 548–559. [CrossRef]
32. Fang, H.; Liang, S.; Hoogenboom, G. Integration of MODIS LAI and Vegetation Index Products with the CSM–CERES–Maize Model for Corn Yield Estimation. *Int. J. Remote Sens.* **2011**, *32*, 1039–1065. [CrossRef]
33. Meroni, M.; Waldner, F.; Seguini, L.; Kerdiles, H.; Rembold, F. Yield Forecasting with Machine Learning and Small Data: What Gains for Grains? *Agric. For. Meteorol.* **2021**, *308–309*, 108555. [CrossRef]
34. Cai, Y.; Guan, K.; Lobell, D.; Potgieter, A.B.; Wang, S.; Peng, J.; Xu, T.; Asseng, S.; Zhang, Y.; You, L.; et al. Integrating Satellite and Climate Data to Predict Wheat Yield in Australia Using Machine Learning Approaches. *Agric. For. Meteorol.* **2019**, *274*, 144–159. [CrossRef]
35. Kamir, E.; Waldner, F.; Hochman, Z. Estimating Wheat Yields in Australia Using Climate Records, Satellite Image Time Series and Machine Learning Methods. *ISPRS J. Photogramm. Remote Sens.* **2020**, *160*, 124–135. [CrossRef]
36. Wolanin, A.; Mateo-García, G.; Camps-Valls, G.; Gómez-Chova, L.; Meroni, M.; Duveiller, G.; Liangzhi, Y.; Guanter, L. Estimating and Understanding Crop Yields with Explainable Deep Learning in the Indian Wheat Belt. *Environ. Res. Lett.* **2020**, *15*, 24019. [CrossRef]
37. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop Yield Prediction Using Machine Learning: A Systematic Literature Review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
38. Han, J.; Zhang, Z.; Cao, J.; Luo, Y.; Zhang, L.; Li, Z.; Zhang, J. Prediction of Winter Wheat Yield Based on Multi-Source Data and Machine Learning in China. *Remote Sens.* **2020**, *12*, 236. [CrossRef]
39. Benos, L.; Tagarakis, A.C.; Dolias, G.; Berruto, R.; Kateris, D.; Bochtis, D. Machine Learning in Agriculture: A Comprehensive Updated Review. *Sensors* **2021**, *21*, 2758. [CrossRef]
40. Jiang Correspond, D.; Yang, X.; Clinton, N.; Wang, N. An Artificial Neural Network Model for Estimating Crop Yields Using Remotely Sensed Information. *Int. J. Remote Sens.* **2004**, *25*, 1723–1732. [CrossRef]
41. Uno, Y.; Prasher, S.O.; Lacroix, R.; Goel, P.K.; Karimi, Y.; Viau, A.; Patel, R.M. Artificial Neural Networks to Predict Corn Yield from Compact Airborne Spectrographic Imager Data. *Comput. Electron. Agric.* **2005**, *47*, 149–161. [CrossRef]
42. Iqbal, M.A.; Shen, Y.; Stricevic, R.; Pei, H.; Sun, H.; Amiri, E.; Penas, A.; del Rio, S.; Shen, Y.; Stricevic, R.; et al. Evaluation of the {FAO} {AquaCrop} Model for Winter Wheat on the North China Plain under Deficit Irrigation from Field Experiment to Regional Yield Simulation. *Agric. Water Manag.* **2014**, *135*, 61–72. [CrossRef]
43. Chen, Y.; Zhang, Z.; Tao, F. Improving Regional Winter Wheat Yield Estimation through Assimilation of Phenology and Leaf Area Index from Remote Sensing Data. *Eur. J. Agron.* **2018**, *101*, 163–173. [CrossRef]
44. Ji, Z.; Pan, Y.; Zhu, X.; Wang, J.; Li, Q. Prediction of Crop Yield Using Phenological Information Extracted from Remote Sensing Vegetation Index. *Sensors* **2021**, *21*, 1406. [CrossRef]
45. Li, Y.; Guan, K.; Yu, A.; Peng, B.; Zhao, L.; Li, B.; Peng, J. Toward Building a Transparent Statistical Model for Improving Crop Yield Prediction: Modeling Rainfed Corn in the U.S. *Field Crops Res.* **2019**, *234*, 55–65. [CrossRef]
46. Kern, A.; Barcza, Z.; Marjanović, H.; Árendás, T.; Fodor, N.; Bónis, P.; Bognár, P.; Lichtenberger, J. Statistical Modelling of Crop Yield in Central Europe Using Climate Data and Remote Sensing Vegetation Indices. *Agric. For. Meteorol.* **2018**, *260–261*, 300–320. [CrossRef]
47. Lobell, D.B.; Thau, D.; Seifert, C.; Engle, E.; Little, B. A Scalable Satellite-Based Crop Yield Mapper. *Remote Sens. Environ.* **2015**, *164*, 324–333. [CrossRef]
48. Johnson, D.M. An Assessment of Pre-and within-Season Remotely Sensed Variables for Forecasting Corn and Soybean Yields in the United States. *Remote Sens. Environ.* **2014**, *141*, 116–128. [CrossRef]
49. Cao, J.; Zhang, Z.; Tao, F.; Zhang, L.; Luo, Y.; Zhang, J.; Han, J.; Xie, J. Integrating Multi-Source Data for Rice Yield Prediction across China Using Machine Learning and Deep Learning Approaches. *Agric. For. Meteorol.* **2021**, *297*, 108275. [CrossRef]
50. Jiang, H.; Hu, H.; Zhong, R.; Xu, J.; Xu, J.; Huang, J.; Wang, S.; Ying, Y.; Lin, T. A Deep Learning Approach to Conflating Heterogeneous Geospatial Data for Corn Yield Estimation: A Case Study of the {US} Corn Belt at the County Level. *Glob. Chang. Biol.* **2020**, *26*, 1754–1766. [CrossRef] [PubMed]
51. Feng, P.; Wang, B.; Liu, D.L.; Waters, C.; Xiao, D.; Shi, L.; Yu, Q. Dynamic Wheat Yield Forecasts Are Improved by a Hybrid Approach Using a Biophysical Model and Machine Learning Technique. *Agric. For. Meteorol.* **2020**, *285–286*, 107922. [CrossRef]
52. Ban, H.-Y.; Kim, K.; Park, N.-W.; Lee, B.-W. Using {MODIS} Data to Predict Regional Corn Yields. *Remote Sens.* **2016**, *9*, 16. [CrossRef]
53. Bolton, D.K.; Friedl, M.A. Forecasting Crop Yield Using Remotely Sensed Vegetation Indices and Crop Phenology Metrics. *Agric. For. Meteorol.* **2013**, *173*, 74–84. [CrossRef]
54. Sakamoto, T.; Gitelson, A.A.; Arkebauer, T.J. {MODIS-Based} Corn Grain Yield Estimation Model Incorporating Crop Phenology Information. *Remote Sens. Environ.* **2013**, *131*, 215–231. [CrossRef]
55. Peng, Y.; Zhu, T.; Li, Y.; Dai, C.; Fang, S.; Gong, Y.; Wu, X.; Zhu, R.; Liu, K. Remote Prediction of Yield Based on {LAI} Estimation in Oilseed Rape under Different Planting Methods and Nitrogen Fertilizer Applications. *Agric. For. Meteorol.* **2019**, *271*, 116–125. [CrossRef]

56. Bai, T.; Zhang, N.; Mercatoris, B.; Chen, Y. Jujube Yield Prediction Method Combining Landsat 8 Vegetation Index and the Phenological Length. *Comput. Electron. Agric.* **2019**, *162*, 1011–1027. [CrossRef]
57. Magney, T.S.; Eitel, J.U.H.; Huggins, D.R.; Vierling, L.A. Proximal {NDVI} Derived Phenology Improves In-Season Predictions of Wheat Quantity and Quality. *Agric. For. Meteorol.* **2016**, *217*, 46–60. [CrossRef]
58. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*, 1st ed.; Springer: New York, NY, USA, 2013.
59. Giannerini, G.; Genovesi, R. Irrinet: IT Services for Farm Water Management, a Large Scale Implementation in Italy. In Proceedings of the EFITA 2011 Conference Proceedings, Prague, Czech Republic, 8–10 June 2011; pp. 11–14.
60. Biavetti, I.; Karetsos, S.; Ceglar, A.; Toreti, A.; Panagos, P. European Meteorological Data: Contribution to Research, Development, and Policy Support. In Proceedings of the Second International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2014), Paphos, Cyprus, 12 August 2014; Hadjimitsis, D.G., Themistocleous, K., Michaelides, S., Papadavid, G., Eds.; SPIE: Bellingham, WA, USA, 2014.
61. Orgiazzi, A.; Ballabio, C.; Panagos, P.; Jones, A.; Fernández-Ugalde, O. LUCAS Soil, the Largest Expandable Soil Dataset for Europe: A Review: {LUCAS} Soil, Pan-European Open-Access Soil Dataset. *Eur. J. Soil Sci.* **2018**, *69*, 140–153. [CrossRef]
62. Drusch, M.; del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: {ESA's} Optical High-Resolution Mission for {GMES} Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [CrossRef]
63. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
64. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the Radiometric and Biophysical Performance of the {MODIS} Vegetation Indices. *Remote Sens. Environ.* **2002**, *83*, 195–213. [CrossRef]
65. Gitelson, A.; Merzlyak, M.N. Quantitative Estimation of Chlorophyll-a Using Reflectance Spectra: Experiments with Autumn Chestnut and Maple Leaves. *J. Photochem. Photobiol. B* **1994**, *22*, 247–252. [CrossRef]
66. Gao, B. NDWI—A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]
67. Johnson, D.M. A Comprehensive Assessment of the Correlations between Field Crop Yields and Commonly Used MODIS Products. *ITC J.* **2016**, *52*, 65–81. [CrossRef]
68. Viña, A.; Gitelson, A.A.; Rundquist, D.C.; Keydan, G.; Leavitt, B.; Schepers, J. Monitoring Maize (*Zea Mays* L.) Phenology with Remote Sensing. *Agron. J.* **2004**, *96*, 1139–1147. [CrossRef]
69. Kowalski, K.; Senf, C.; Hostert, P.; Pflugmacher, D. Characterizing Spring Phenology of Temperate Broadleaf Forests Using Landsat and Sentinel-2 Time Series. *ITC J.* **2020**, *92*, 102172. [CrossRef]
70. Antonucci, G.; Croci, M.; Miras-Moreno, B.; Fracasso, A.; Amaducci, S. Integration of Gas Exchange With Metabolomics: High-Throughput Phenotyping Methods for Screening Biostimulant-Elicited Beneficial Responses to Short-Term Water Deficit. *Front. Plant Sci.* **2021**, *12*, 1002. [CrossRef] [PubMed]
71. Impollonia, G.; Croci, M.; Martani, E.; Ferrarini, A.; Kam, J.; Trindade, L.M.; Clifton-Brown, J.; Amaducci, S. Moisture Content Estimation and Senescence Phenotyping of Novel Miscanthus Hybrids Combining UAV-based Remote Sensing and Machine Learning. *GCB Bioenergy* **2022**, *14*, 639–656. [CrossRef]
72. Wood, S.N. Thin Plate Regression Splines: Thin Plate Regression Splines. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2003**, *65*, 95–114. [CrossRef]
73. Jolliffe, I.T.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [CrossRef]
74. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
75. Murtagh, F. Multilayer Perceptrons for Classification and Regression. *Neurocomputing* **1991**, *2*, 183–197. [CrossRef]
76. Vapnik, V. The Support Vector Method of Function Estimation. In *Nonlinear Modeling*; Springer: Boston, MA, USA, 1998; pp. 55–85.
77. Williams, C.K.I. Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond. In *Learning in Graphical Models*; Springer: Dordrecht, The Netherlands, 1998; pp. 599–621.
78. Györfi, L.; Kohler, M.; Krzyżak, A.; Walk, H. *A Distribution-Free Theory of Nonparametric Regression*; Springer: New York, NY, USA, 2002.
79. Kuhn, M. Classification and Regression Training [R Package Caret Version 6.0-90]. 2021. Available online: https://cran.r-project.org/web/packages/caret/caret.pdf (accessed on 20 July 2022).
80. Arlot, S.; Celisse, A. A Survey of Cross-Validation Procedures for Model Selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
81. Picard, R.R.; Cook, R.D. Cross-Validation of Regression Models. *J. Am. Stat. Assoc.* **1984**, *79*, 575. [CrossRef]
82. Biecek, P. {DALEX}: Explainers for Complex Predictive Models in R. *J. Mach. Learn. Res.* **2018**, *19*, 3245–3249.
83. Impollonia, G.; Croci, M.; Ferrarini, A.; Brook, J.; Martani, E.; Blandinières, H.; Marcone, A.; Awty-Carroll, D.; Ashman, C.; Kam, J.; et al. UAV Remote Sensing for High-Throughput Phenotyping and for Yield Prediction of Miscanthus by Machine Learning Techniques. *Remote Sens.* **2022**, *14*, 2927. [CrossRef]
84. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, MA, USA, 2005; ISBN 9780262256834.

85. Verrelst, J.; Muñoz, J.; Alonso, L.; Delegido, J.; Rivera, J.P.; Camps-Valls, G.; Moreno, J. Machine Learning Regression Algorithms for Biophysical Parameter Retrieval: Opportunities for Sentinel-2 and -3. *Remote Sens. Environ.* **2012**, *118*, 127–139. [CrossRef]
86. Li, L.; Wang, B.; Feng, P.; Wang, H.; He, Q.; Wang, Y.; Liu, D.L.; Li, Y.; He, J.; Feng, H.; et al. Crop Yield Forecasting and Associated Optimum Lead Time Analysis Based on Multi-Source Environmental Data across China. *Agric. For. Meteorol.* **2021**, *308–309*, 108558. [CrossRef]
87. Meng, W.; Tao, F.; Shi, W.; Meng, W.; Tao, F.; Shi, W. Corn Yield Forecasting in Northeast China Using Remotely Sensed Spectral Indices and Crop Phenology Metrics. *J. Integr. Agric.* **2014**, *13*, 1538–1545.
88. Shanahan, J.F.; Schepers, J.S.; Francis, D.D.; Varvel, G.E.; Wilhelm, W.W.; Tringe, J.M.; Schlemmer, M.R.; Major, D.J. Use of Remote-sensing Imagery to Estimate Corn Grain Yield. *Agron. J.* **2001**, *93*, 583–589. [CrossRef]
89. Maestrini, B.; Basso, B. Predicting Spatial Patterns of Within-Field Crop Yield Variability. *Field Crops Res.* **2018**, *219*, 106–112. [CrossRef]
90. Chen, S.; Jiang, T.; Ma, H.; He, C.; Xu, F.; Malone, R.W.; Feng, H.; Yu, Q.; Siddique, K.H.M.; Dong, Q.; et al. Dynamic Within-Season Irrigation Scheduling for Maize Production in Northwest China: A Method Based on Weather Data Fusion and Yield Prediction by {DSSAT}. *Agric. For. Meteorol.* **2020**, *285–286*, 107928. [CrossRef]
91. Peoples, M.B.; Beilharz, V.C.; Waters, S.P.; Simpson, R.J.; Dalling, M.J. Nitrogen Redistribution during Grain Growth in Wheat (Triticum Aestivum L.): II. Chloroplast Senescence and the Degradation of Ribulose-1,5-Bisphosphate Carboxylase. *Planta* **1980**, *149*, 241–251. [CrossRef]
92. French, R.J.; Schultz, J.E. Water Use Efficiency of Wheat in a Mediterranean-Type Environment. I. The Relation between Yield, Water Use and Climate. *Aust. J. Agric. Res.* **1984**, *35*, 743. [CrossRef]
93. Atteya, A.M. Alteration of Water Relations and Yield of Corn Genotypes in Response to Drought Stress. *Bulg. J. Plant Physiol.* **2003**, *29*, 63–76.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.