Cereal Yield Forecasting with Satellite Drought-Based Indices, Weather Data and Regional Climate Indices Using Machine Learning in Morocco

El Houssaine Bouras, Lionel Jarlan, Salah Er-Raki, Riad Balaghi, Abdelhakim Amazirh, Bastien Richard, Saïd Khabba

remote sensing

MDPI

Article

Cereal Yield Forecasting with Satellite Drought-Based Indices, Weather Data and Regional Climate Indices Using Machine Learning in Morocco

El houssaine Bouras [1,2,*], Lionel Jarlan [2], Salah Er-Raki [1,3], Riad Balaghi [4], Abdelhakim Amazirh [3], Bastien Richard [5] and Saïd Khabba [3,6]

1 ProcEDE, Department of Applied Physique, Faculty of Sciences and Technologies, Cadi Ayyad University, Marrakech 40000, Morocco; s.erraki@uca.ma
2 CESBIO, University of Toulouse, IRD/CNRS/UPS/CNES, 31400 Toulouse, France; lionel.jarlan@ird.fr
3 Center for Remote Sensing Applications (CRSA), University Mohammed VI Polytechnic (UM6P), Benguerir 43150, Morocco; abdelhakim.amazirh@um6p.ma (A.A.); khabba@uca.ma (S.K.)
4 National Institute for Agronomic Research (INRA), Rabat 10000, Morocco; riad.balaghi@inra.ma
5 G-EAU, University Montpellier, AgroParisTech, CIRAD, IRD, INRAE, Institut Agro, 34000 Montpellier, France; bastien.richard@irstea.fr
6 LMFE, Department of Physics, Faculty of Sciences Semlalia, Cadi Ayyad University, Marrakech 40000, Morocco
* Correspondence: elhoussaine.bouras@ced.uca.ma; Tel.: +212-(0)-5-2443-3404 or +212-(0)-6-0136-8700

Abstract: Accurate seasonal forecasting of cereal yields is an important decision support tool for countries, such as Morocco, that are not self-sufficient in order to predict, as early as possible, importation needs. This study aims to develop an early forecasting model of cereal yields (soft wheat, barley and durum wheat) at the scale of the agricultural province considering the 15 most productive over 2000–2017 (i.e., 15 × 18 = 270 yields values). To this objective, we built on previous works that showed a tight linkage between cereal yields and various datasets including weather data (rainfall and air temperature), regional climate indices (North Atlantic Oscillation in particular), and drought indices derived from satellite observations in different wavelengths. The combination of the latter three data sets is assessed to predict cereal yields using linear (Multiple Linear Regression, MLR) and non-linear (Support Vector Machine, SVM; Random Forest, RF, and eXtreme Gradient Boost, XGBoost) machine learning algorithms. The calibration of the algorithmic parameters of the different approaches are carried out using a 5-fold cross validation technique and a leave-one-out method is implemented for model validation. The statistical metrics of the models are first analyzed as a function of the input datasets that are used, and as a function of the lead times, from 4 months to 2 months before harvest. The results show that combining data from multiple sources outperformed models based on one dataset only. In addition, the satellite drought indices are a major source of information for cereal prediction when the forecasting is carried out close to harvest (2 months before), while weather data and, to a lesser extent, climate indices, are key variables for earlier predictions. The best models can accurately predict yield in January (4 months before harvest) with an R^2 = 0.88 and RMSE around 0.22 t. ha^{-1}. The XGBoost method exhibited the best metrics. Finally, training a specific model separately for each group of provinces, instead of one global model, improved the prediction performance by reducing the RMSE by 10% to 35% depending on the provinces. In conclusion, the results of this study pointed out that combining remote sensing drought indices with climate and weather variables using a machine learning technique is a promising approach for cereal yield forecasting.

Keywords: crop yield forecasting; machine learning; remote sensing drought indices; climate indices; weather data; semiarid region

Citation: Bouras, E.h.; Jarlan, L.; Er-Raki, S.; Balaghi, R.; Amazirh, A.; Richard, B.; Khabba, S. Cereal Yield Forecasting with Satellite Drought-Based Indices, Weather Data and Regional Climate Indices Using Machine Learning in Morocco. *Remote Sens.* **2021**, *13*, 3101. https://doi.org/10.3390/rs13163101

Academic Editors: Bin Chen, Yufang Jin and Le Yu

Received: 4 June 2021
Accepted: 30 July 2021
Published: 6 August 2021

1. Introduction

Climate change will affect global crop production [1,2] and threaten food security in several regions of the globe including the Mediterranean areas that have long been identified as a hot spot for climate change [3,4]. It has been shown that a 1 °C increase in temperature would lead to a drop of 6% in global wheat production for instance [1]. Besides the expected change of the average characteristics of climate, including temperature and precipitation, extreme events can further reduce crop production. Indeed, drought, the frequency of which is expected to increase in the future [5], can be responsible for a 10% to 35% loss depending on its intensity, timing and duration [6]. The southern Mediterranean countries and Morocco, in particular, are characterized by a strong interannual variability in precipitation amounts and distribution and recurrent droughts that mainly affect rainfed crops among which wheat dominates with more than 90% of cultivated areas [7]. Within this context, achieving food security, one of the key points of the Sustainable Development Goals [8], relies on a reliable monitoring system of wheat production [2]. An early and reliable forecast of the pre-harvest cereal yield in large areas would assist decision-makers in order to anticipate important needs, especially in countries such as Morocco that are not always self-sufficient [9–12]. It would also help to identify yield gaps and to better understand the wheat response to local climatic and edaphic conditions [9,13,14].

Besides the agricultural statistics based on sample observations in the field, the monitoring and forecasting of wheat yields are mainly carried out using empirical regression-based models or crop growth models based on biophysiological processes [15]. The latter is able to describe crop growth and yield response to weather conditions, soil, and management practices [16] and can provide a good estimate of final crop yield when accurate values of input parameters and meteorological forcing variables are available; a strong drawback for southern countries considering the sparsity of the ground-based networks. Another limitation arises for seasonal forecasting in relation to the forcing meteorological data during the period between the forecast date and harvest time [17]. Seasonal weather forecasts either based on historical weather observation [18–20], on weather generators [21] or on climate model outputs [22] remain very uncertain. Given these limitations, the majority of the national agriculture departments use empirical regression-based models to forecast yield over large areas. These models rely on the use of some selected variables or indicators of environmental conditions (agrometeorological, and/or remotely sensed data) as independent variables to forecast crop yield [12,23–26]. In addition, as the quantity and the quality of observed data have increased in recent years, these models forecast crop yield with reasonable accuracy [24,27].

Weather data have long been used to explain crop yield variability [27–29]. In this context, Sierra and Brynsztein [30] have used temperature and precipitation as predictors to forecast wheat yield up to 3 months before the harvest in Argentina; Giri et al. [31] have used several meteorological variables to predict wheat yield at the district scale in India. The models developed in this study had an R^2 range between 0.6 and 0.92, depending on district's location. Nevertheless, the performance of the models was lower for some districts, which may be due to other variables influencing yields such as the soil type or the practical management. More recently, many research works have focused on establishing a relationship between remote sensing indices and observed crop yield [10,32–34]. The main advantage of using remote sensing observations in crop yield forecasting is that they allow the obtaining of information on a large scale, independent of territorial boundaries. The Normalized Difference Vegetation Index (NDVI) is one of the most widely used variables to forecast the final crop yield at a large scale [15]. For instance, the Moderate Resolution Imaging Spectroradiometer (MODIS) and Advanced Very High-Resolution Radiometer (AVHRR) derived NDVI has been used to develop linear regression models to predict maize, wheat and rice yields up to 2 to 3 months before harvest [24,35,36]. Besides, other studies have used remote sensing drought indices, such as the Vegetation Condition Index (VCI) and the Temperature Condition Index (TCI) from AVHRR data, to forecast wheat yield in the United States [34], and soybean yield in Brazil, respectively [37]. Interestingly

enough, it has been shown that the use of these drought indices to forecast crop yields in Spain outperformed models based on precipitation anomalies only [38]. This can be explained by the accurate detection of local drought conditions provided by these indices integrating information on climate and biophysical conditions when compared to indices based only on precipitation. In addition, the high spatial resolution of satellite products with regards to meteorological data provided by a coarse network of weather stations [38] may be an advantage, in particular for southern countries often characterized by sparse meteorological networks. Finally, several studies have also shown the impact of large-scale climate pseudo-oscillations on many components of the continental ecosystems including agricultural systems [39–41], and on the future monthly precipitation [41,42]. In light of this, large-scale climate indices and data, including El Niño Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO) and Sea Surface Temperature (SST) have been used as predictors of crop yield in different regions over the world [41,43–45]. In Morocco, in particular, wheat yields have been shown to be tightly linked to NAO value in December and to the leading mode of the SST in the tropical Atlantic [11]. In Australia, the large-scale climate indices related to ENSO have been incorporated into empirical models to predict wheat yield up to 3 months before the harvest [41]. Instead of using a single data source as predictors of crop yield, several studies have combined multi-source data to predict crop yield. For example, Cai and Sharma, [46] and Balaghi et al. [12] combined remote sensing data (NDVI) and weather data (rainfall and temperature) as predictors of rice and wheat yield in India and Morocco, respectively.

Most of the models developed in the previously cited studies are based on the classical Multiple Linear Regression (MLR) while the linkages between yields and potential predictors are likely to be non-linear. For this reason, non-linear machine learning algorithms have been employed to improve crop yield prediction [47]. Recently, several studies have examined the performance of machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF), eXtreme Gradient Boost (XGBoost), Artificial Neural Network (ANN) and Long-Short Term Memory (LSTM) for yield forecasting at county or province scales. They have used multi-source data as predictors, and they found that the non-linear machine learning methods showed a better performance for yield forecasting than the linear approach [48–54]. Schwalbert et al. [55] have used different machine learning algorithms (linear regression, RF and LSTM) to predict soybean yield at the municipality level in Brazil by using remote sensing data (NDVI, EVI, LST) and precipitation as predictors. They found that soybean yield can be forecasted with a mean absolute error of 0.42 t. ha^{-1} around 2 months before harvesting. Cai et al. [56] have also forecasted wheat yield in Australia by incorporating various predictors into SVM, RF and ANN algorithms. In their study, they used the enhanced vegetation index (EVI) from MODIS, solar-induced chlorophyll fluorescence from GOME-2 and several climate variables and they found that combining climate and satellite data achieved a high performance of wheat prediction (R^2 = 0.75). For Morocco, the existing literature on seasonal yield forecasting is limited. Balaghi et al. [12] proposed empirical linear regression models to forecast wheat yields up to 2 months before the harvest at a provincial and national scale. NDVI from AVHRR, rainfall sums and average monthly air temperatures were used. More recently, Lehmann et al. [45] have used SST and causal precursors from geopotential height anomalies at 500 hPa to forecast wheat yield anomalies at the country scale. Several studies have used the combination of multi-source datasets and machine learning algorithms to forecast crop yield including cereals [48–54]. However, the combination between remote sensing drought indices, weather data and climate indices as predictors of cereals yield has not been assessed yet.

In this context, the aim of this study is to investigate the potential of using machine learning for developing dynamic decision support systems for cereal production in Morocco, combining satellite-based drought indices, weather and climate data. Our specific objective is to develop empirical models that can forecast cereal yield early in the crop season (up to 4 months before harvest). More specifically, this study builds on previous work carried out in Morocco that highlighted biophysically sound linkages between wheat

yields and weather data and climate indices (Jarlan et al. [11]) and between wheat yields and drought indices (Bouras et al. [10]). It also aimed to go further than Balaghi et al. [12] by analyzing the potential of climate and drought indices information to forecast yields earlier in the season and at a finer spatial scale than Lehmann et al. [45].

2. Materials and Methods

In this study, the target variable is cereal yield. The potential predictors are the satellite drought indices, weather data (rainfall and temperature) and climate indices derived from atmospheric and oceanic variables. In order to limit the number of agricultural provinces, a threshold of 90% of the national production was set: the 15 selected provinces corresponding to the most productive are displayed in Figure 1. The forecasting models are then built using the extensively used multi-linear regression approach and three non-linear machine learning methods. An overview of the methodology is represented in the flowchart of Figure 2. Table 1 lists all the raw datasets with their sources. Table 2 displays the predictor variables derived from these raw datasets together with the time span of the year considered in the model based on biophysically sound linkages highlighted by previous studies.

Table 1. Summary of the raw characteristics of the data sets used for yields prediction as well as yields "observations" information. All the datasets used for yields prediction are then averaged at the agricultural province and the monthly time scales (see text).

Category	Product	Variable	Spatial Resolution	Temporal Resolution	Source of Data
Crop Yield		Crop yield	Province level	Yearly	Ministry of agriculture of Morocco
Remote sensing	MOD13A2	NDVI	1 km	16-Day	https://lpdaac.usgs.gov (accessed on 31 July 2021)
	MOD11A1	LST	1 km	Daily	
	ESA CCI SM COMBINED	SM	25 km	Daily	https://www.esa-soilmoisture-cci.org (accessed on 31 July 2021)
Weather	ERA5	Rainfall, Air temperature	30 km	Daily	https://www.ecmwf.int/en/forecasts/dataset/reanalysis-datasets/era5 (accessed on 31 July 2021)
Climate		NAO, SCA, SST		Monthly	https://psl.noaa.gov/data/climateindices/ (accessed on 31 July 2021)

Table 2. Input predictors for the forecasting model and time span of the year when these variables are considered based on previously highlighted biophysically sound linkages.

Predictor Variables	Raw Products	Time Span of the Year	Publication
VCI	NDVI	February–April	[10–12]
TCI	LST	January–February	[10]
SMCI	SM	October–November	[10]
Air temperature	ERA5 air temperature	December	[11,12]
Rainfall	ERA5 rainfall	October–November and January–March	[11,12]
NAO	Northern Hemispheric Teleconnection Patterns	December	[11]
SCA	Northern Hemispheric Teleconnection Patterns	January	[11]
Atlantic Tripole	SST	February	[11]
Atlantic Niño	SST	October	[11]

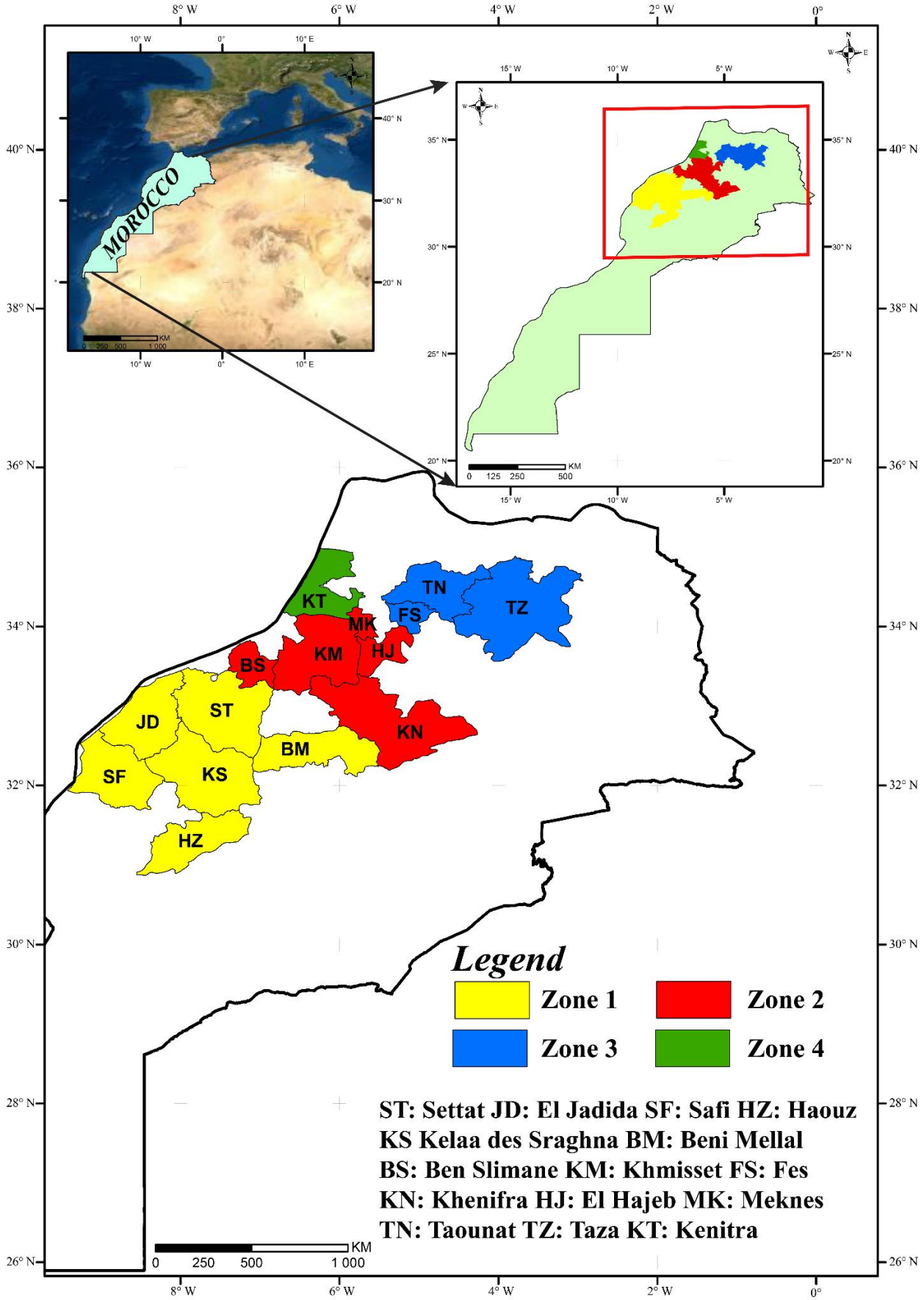

Figure 1. The study areas with the 15 provinces.

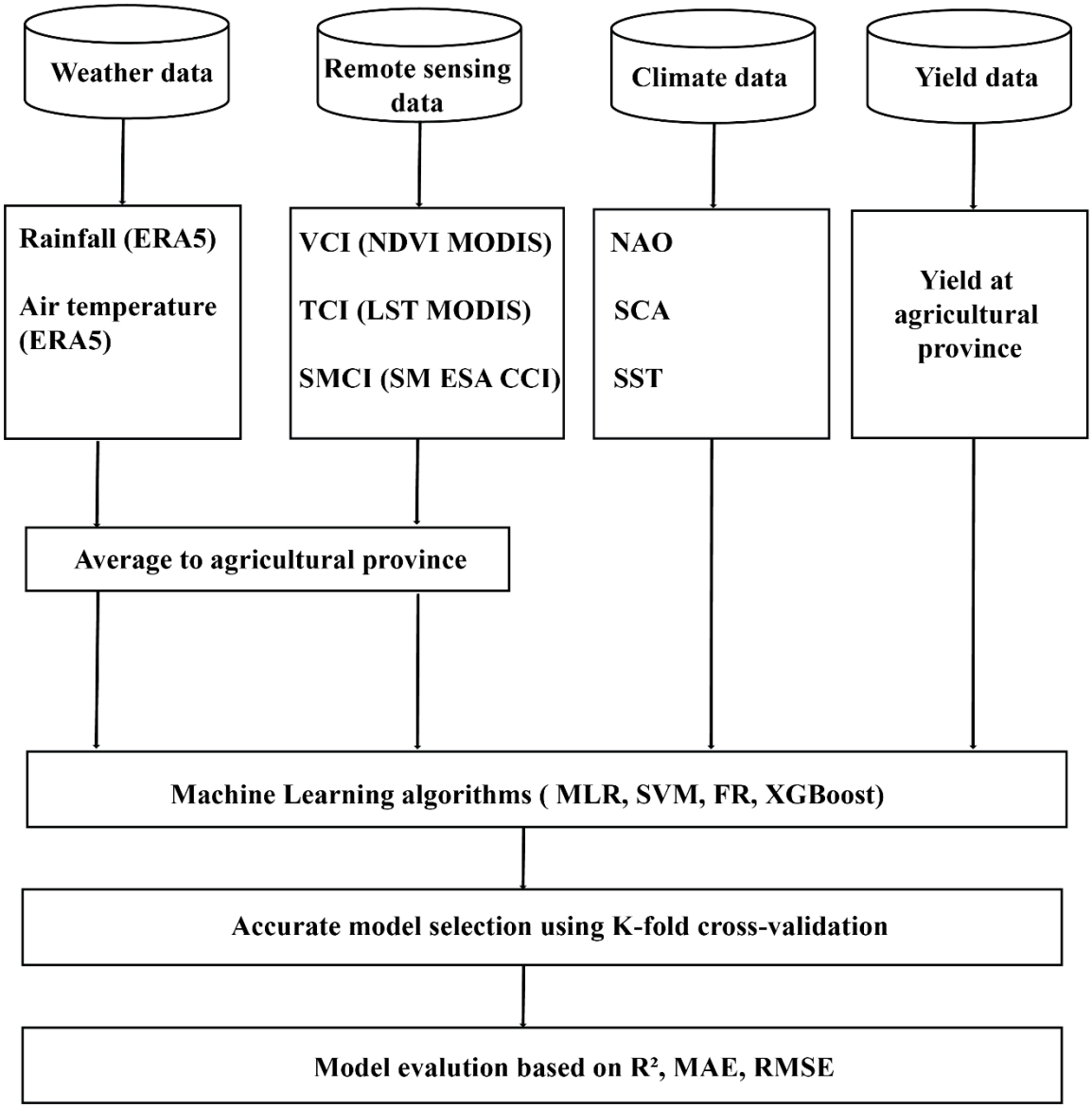

Figure 2. Schematic diagram presenting an overview of the main inputs data and the methodology proposed in this study.

2.1. Study Area

Morocco is a North African country (Figure 1) with a semi-arid climate influenced by the Atlantic Ocean, the Mediterranean Sea and the Sahara [42]. Precipitation in Morocco is characterized by its strong spatiotemporal variability and a rainfall season—extending through winter and spring from November to April, which coincides with the cereal growing season. The northern part of Morocco receives higher amounts of precipitation which can rise to 900 mm while the center of the country is marked by low amounts of precipitation below 350 mm. Similarly, the high spatial variability of the temperature is noted. The regions located at high elevation (the high Atlas Mountains) are marked by low temperatures when compared to other regions in the country [10,57]. Cereals are the main rainfed crop occupying up to 90% of the rainfed usable agricultural area in Morocco. The early sowing takes place in November if significant precipitation occurs at this time, while the sowing can be extended to January in the case of delays in precipitation. Late sowing usually leads to a lower production compared to early sowing, due to both a decrease of the cropped areas because a large part of farmers will not seed if precipitation arrives late in the season and because the last stage of

the season corresponds to periods of high temperature that may hamper yields [9]. Harvest takes place generally around the end of May.

2.2. Yield Data

Soft wheat, barley and durum wheat are the main types of cereals cropped in Morocco. Data on cereal crop productions and harvested areas over the study period 2000–2017 at the administrative provincial scale were gathered by the Economic Services of the Ministry of Agriculture in Morocco (https://www.agriculture.gov.ma/, accessed on 31 July 2021). The yearly cereal yield used in this study for each province was calculated as the ratio of the total crop production by the total harvest areas. The average cereal yield over the study period ranged from 0.7 t. ha^{-1} to 2.2 t. ha^{-1} depending on the province. The 15 selected provinces were classified into four groups with similar cereal yield interannual variability using a k-means algorithm based on the correlative distance [58]. More details about the cereal yield data and classification are provided in Bouras et al. [10].

2.3. Satellite-Based Drought Indices

Agricultural drought affects both vegetation and soil, the characteristics of which can be monitored by remote sensing observation [59]. We selected three extensively used satellite-based drought indices: the Vegetation Condition Index (VCI), the Temperature Condition Index (TCI) [60] and the Soil Moisture Condition Index (SMCI) [61]. The VCI, TCI and SMCI are the normalized anomalies of NDVI, Land Surface Temperature (LST) and soil moisture (SM), respectively. While the VCI is related to vegetation density and activity, the TCI is related to the thermal stress of vegetation and the SMCI describes soil moisture drought as it is based on soil moisture anomalies in the first centimeters. These indices were widely used in agricultural drought monitoring [10,38,62,63]. Bouras et al. [10] have analyzed the linkages between these indices and cereal yield at the provincial scale in Morocco. Their results have shown that the VCI in March and April during the heading stage of wheat is highly correlated to cereal yield. For TCI, the highest correlation with cereal yield was observed around the development stage in January–February. Finally, SMCI was found to be connected with cereal yield earlier at the beginning of crop season during the emergence stage (December–January).

The VCI was calculated with NDVI from MODIS (MOD13A2 collection 6). The VCI compares the currently observed value of NDVI to the minimum and maximum NDVI values observed during a study period. As such, the VCI lies between 0 and 100, with a low VCI value associated with below-normal vegetation development while above-normal vegetation development is indicated by a high VCI value. TCI was computed in a similar way to VCI but using the LST from MODIS LST (MOD11A1 collection 6). The high TCI values indicate low temperatures, then favorable climatic conditions, while lower values of TCI reflect unfavorable conditions with high temperatures. The SMCI is a normalization of soil moisture. SMCI lies between 0 and 100; the lower values indicate unfavorable soil moisture conditions (very dry), and the higher values indicate favorable conditions (very wet). In this study, we used the SM COMBINED version 4.2. product provided by the European Space Agency Climate Change Initiative (ESA CCI) [64].

2.4. Weather Data

The linkage between cereal yield and weather data, including rainfall and temperature over Morocco, has been analyzed in previous studies [10–12]. The main results of these studies are: for rainfall, a positive correlation with cereal yields was observed in November–December. When the rainfall is abundant during this period, it will speed up plant emergence and increase cropped areas as already highlighted. Concerning temperature, a positive correlation with cereal yields was observed in December and January during the early stage. Low temperatures during this period cause poor emergence and reduce the number of ears leading to lower yields. By contrast, a negative correlation with cereal yield was observed in March meaning that high temperatures should be avoided during the grain

filling stage occurring at this time of the year [65]. Weather data including air temperature at 2 m. above land surface and rainfall were extracted from the ERA5 re-analysis data set [66].

2.5. Climate Data

In a previous study, Jarlan et al. [11] have analyzed the relationship between provincial-scale wheat yields in Morocco and large-scale climate. Significant correlations have been found between yields and the NAO in December (negative sign), the Scandinavian Pattern (SCA) in January (positive sign) and the leading modes of SST on the northern hemisphere ("Atlantic tripole" mode) in February (negative sign) and on the equatorial Atlantic (the so-called "Atlantic Niño" mode) earlier in the season in October (positive sign). In this study, we evaluate the potentiality of introducing this climate information in addition to satellite drought indices and weather data to predict cereal yields. The NAO and the SCA are part of the Northern Hemisphere Teleconnection Patterns [67]. These indices are distributed with a monthly time scale by the Climate Prediction Center (https://www.cpc.ncep.noaa.gov/, accessed on 31 July 2021). In addition to these atmospheric indices, the monthly sea surface temperature (SST) leading modes are computed from monthly NOAA SST v2 at 0.25° resolution throughout the study on a North Atlantic window (20°N–70°N, 80°W–20°E) and an Equatorial Atlantic window (20°S–20°N, 80°W–20°E) for the "Atlantic tripole" and the "Atlantic Niño", respectively, following Jarlan et al. [11].

2.6. Machine Learning Methods for Cereal Yield Forecasting

In order to build the seasonal forecasting models, we relied on Multiple Linear Regression, and three non-linear machine learning algorithms extensively used for crop yield prediction [47], which are: Support Vector Machine (SVM), Random Forest (RF) and eXtreme Gradient Boost (XGBoost). The scikit-learn package in Python 3.7 [68] was used in this study.

2.6.1. Multiple Linear Regression

In Multiple Linear Regression (MLR) [69], the dependent variable y is linearly related to multiple independent variables $x_i = 1, \ldots, n$ as:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n, \quad (1)$$

where y in this study is the predicted yield, x_i $(i = 0, \ldots, n)$ are the satellite-based drought indices, the weather data and/or the climate indices and a_i $(i = 0, \ldots, n)$ are the regression coefficients.

2.6.2. Random Forest (RF)

The RF algorithm was introduced by Breiman (2001) and is used for both classification and regression. RF for regression is an ensemble of multiple decision trees regression model; each tree provides its prediction and the optimal prediction of RF obtained by averaging the prediction of all decision trees regression in RF [70]. The RF-based model follows three steps to provide the optimal predictions. In the first step, the dataset is split into data subdivisions. In the second step, each data subdivision is used to develop a single decision tree representing a sub-regression model that gives its prediction. In the last step, predictions of all decision trees are averaged to provide the final prediction. The hyper-parameters that need to be tuned in the RF algorithm are the number of trees or the number of regression trees, the number of features to consider when looking for the best split, and the maximum depth of the tree. These hyper-parameters were tuned with a grid search method, described in Section 2.7.

2.6.3. Support Vector Machine (SVM)

The Support Vector Machine (SVM) was originally developed to solve classification problems and it was extended to solve regression problems, namely support vector regression (SVR) [71]. The SVM algorithm uses kernels [72]. By relying not only on the

minimization of the distance to training data (the training error or empirical risk) but also by trying to limit the model "complexity" (i.e., to search a function as flat as possible: the structural risk), SVR may have, a priori, better generalization capacity (i.e., for data not contained in the training set) than MLR. The SVM regression-based model passes through two steps. In the first step, by using the kernel function, which can be linear or non-linear depending on the relationship between the independent (=Crop yield in our case) and dependent variables, the independent variables (remote sensing drought indices, weather and/or climate indices in our case) are transformed from the original space to a high-dimensional feature space. In the last step, a linear model is built by the new derived feature space to minimize the errors [73]. The SVM algorithm based on the Radial Basis Function (RBF) (the most popular choice in the literature) had two hyper-parameters: the penalty factor C aiming to find a trade-off between the fitting error and the model "complexity", and the kernel width gamma [74]. The SVM hyper-parameters were tuned with a grid search method.

2.6.4. eXtreme Gradient Boost (XGBoost)

eXtreme Gradient Boost (XGBoost) is a machine learning algorithm proposed by Chen and Guestrin (2016), derived from the Gradient Boosting Machines (GBM) [75,76]. The basic principle of the approach is to consider a set of weak learners (with high error) that are combined to develop a new stronger learner (with low error) through the introduction of training additive strategy. The main idea of boosting methods is to use the negative gradient direction of the model loss function, which was established previously, and then iteratively improves the accuracy of the model [77]. The hyper-parameters the number of gradients boosted trees, the maximum tree depth and the learning rate were tuned using the grid search method.

2.7. Model Evaluation

To select the best hyper-parameters of the ML algorithms, the comprehensive grid search (GS) was used in this study, to examine all possible combinations of the hyper-parameters, and cross-validation (CV) was used to assess the performance of the model [78]. In GS, a set of values was attributed for each hyper-parameter and a set of trials was formed by assembling every possible combination of values. The evaluation was performed using k-fold cross-validation. The CV is the most employed technique for algorithm selection and evaluation, due to its simplicity and ability to avoid over-fitting [79–82]. In k-fold cross-validation, the training data are randomly divided into k subsets and the hold-out method is repeated k times, such that each time, one of the k subsets is used as the validation set of the model constructed using $(k-1)$ subsets [82]. To evaluate the performance of the developed models, widely employed statistical metrics were used in this study. The coefficient of determination (R^2) reflects the degree of linear relationship between the observed and forecasted cereal yields. The mean absolute error (*MAE*) indicates the percentage of the average deviation of the forecasted yield from the observation. The Root Mean Square Error (*RMSE*) measures the discrepancy of forecasted yield around observations.

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(\left(O_i - \overline{O}\right)\left(F_i - \overline{F}\right)\right)\right)^2}{\sum_{i=1}^{n}\left(O_i - \overline{O}\right)^2 \sum_{i=1}^{n}\left(F_i - \overline{F}\right)^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|F_i - O_i| \quad (3)$$

$$RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(F_i - O_i)^2}, \quad (4)$$

where O_i is the observed yield, F_i is the forecasted yield by the machine learning algorithm, $\overline{O}$ and $\overline{F}$ are the averages of the observed and predicted yields and n is the number of samples used for the machine learning model.

2.8. Experiment Design

The identification of rainfed cereal areas is an important point in order to obtain values of the remote sensing drought indices representative of rainfed cereal conditions at the scale of the agricultural province. The identification of rainfed cereal areas was carried out based on the joint use of two land cover maps: ECOCLIMAP-II at a 1 km resolution [83] (https://opensource.umr-cnrm.fr/projects/ecoclimap/wiki, accessed on 31 July 2021) was used to determine the cereal areas and the land cover map provided by the Climate Change Initiative (CCI) land cover project of the ESA [84] (https://www.esa-landcover-cci.org/, accessed on 31 July 2021) at 300 m resolution was used to isolate the rainfed fields as described in Bouras et al. [10]. Remote sensing drought indices were then aggregated to the agricultural province by a simple average of these pixels identified as rainfed cereal areas. Weather data were also averaged at the scale of the agricultural province but without considering the rainfed cereal mask because of their coarse spatial resolution of about 25 km (Table 1).

Multicollinearity between the predictors variables is well known to increase the variance of the coefficients for MLR. This can limit the generalization capability of the MLR models as well as hamper the interpretation of the coefficients. In this study, no method to reduce data redundancy was applied because a pre-selection of the time span of the year considered for each predictor was carried out based on previous studies to limit collinearity. Nevertheless, the prediction metrics of the MLR models could probably be improved by applying methods to reduce collinearity such as Principal Component Analysis.

Four experiments were then designed. The first experiment was constructed to identify the best combinations of input datasets among the satellite-based drought indices, the weather data and the climate indices that will reach the high performance in forecasting final cereal yield in Morocco. For this reason, all machine learning algorithms were applied using the different combinations of available input data collected from October to April (see Table 2): (i) Satellite-based drought indices only; (ii) Satellite-based drought indices and weather data; and (iii) Satellite-based drought indices, weather and climate data. The second experiment was conducted in order to assess the performance of the models as a function of the lead time before harvest from 4 to 2 months. In this experiment, we used the best combination of inputs data, determined from the first experiment, and the input data were collected from October to January, October to February and from October to March, based on Table 2, to build the forecasting model in January, February and March, respectively. Then, the performance of machine learning models was evaluated in March, February and January which corresponds to 2, 3 and 4 months before the harvest. In addition, the importance of each input data point was computed using the best machine learning algorithm in order to assess the contribution of each input data point for each lead time of prediction. The third experiment was designed to test the practical performance of the developed models. For this reason, the predictions are performed using a "leave-one year-out" approach consisting in predicting the yield value of one year using all the other years data to train the model (for instance, yield in 2017 is predicted based on a model trained using the 2000–2016 database). Finally, the last experiment was designed to assess the performance of using specific models developed separately for each group of provinces with regards to one global model used in the previous experiments. In this experiment, the accuracy (RMSE) of the global model developed for all provinces was compared to the model developed at a regional level based on a leave-one province-out approach.

3. Results

Results are organized around three sections dedicated to: (1) the assessment of the best combination of predictor datasets; (2) the performance of the seasonal forecasting models as a function of the lead times before harvest; and (3) the evaluation of the added-value of developing a model separately for each group of provinces.

3.1. Choice of Input Data Sets

In order to identify the best combination of input data among the satellite-based drought indices, the weather data and the climate indices, the seasonal forecasting models of cereal yields were developed using the different combinations of input data within the season, about 1 month prior to harvest in April, by considering all available predictors from October to April (Table 2). All the provinces are considered to build a so-called global model. The statistical metrics for the different combinations of input datasets and for the different methods are reported in Table 3. The results presented in Table 3 show that the statistical metrics of the model improve with the increase of the number of datasets used for prediction. In addition, all statistical metrics are improved when adding a dataset and this is also true for all the tested methods. The results showed that the yield variability is reasonably explained with satellite-based drought indices only, with R^2 values ranging from 0.67 (for MLR) to 0.81 (for XGBoost) and RMSE from 0.66 t. ha^{-1} (for MLR) to 0.44 t. ha^{-1} (for XGBoost). By combining satellite-based drought indices and weather data, the performances of all models are improved by 2–7% for R^2 and by 25–30% for RMSE. The best statistical metrics are obtained by combining the three datasets with a further improvement of the statistical metrics by about 11–45% for RMSE and 4–10% for R^2 depending on the used method. This means that climate indices such as the Northern Hemisphere Teleconnection Patterns (NAO and SCA) and the main modes of SST variability in the Atlantic contributes to improving the model performances. In addition, when comparing the different methods, the non-linear machine learning approaches (RF, SVM and XGBoost), outperformed the linear approaches (MLR) as already shown by various authors when applied to seasonal predictions of yields [53] and streamflow [85]. This clearly reflects that most of the relationships between yield and the considered predictors are non-linear and that the non-linear methods can obviously better capture these relationships than the linear method. Finally, the best algorithm for yield forecasting in our study is XGBoost, which predicts the yield with R^2 = 0.95 and RMSE = 0.20 t. ha^{-1}. This finding was corroborated by several studies for seasonal yield forecasting that showed a better performance of XGBoost when compared to other non-linear machine learning approaches such as SVM and RF [52]. Interestingly enough, this model fits the forecasting error threshold usually accepted in European agro-statistics that is of about 0.20 t. ha^{-1} [86]. In the next section, the combination of satellite drought indices, weather data and climate indices are considered to predict cereal yield for several lead times before harvest.

Table 3. Statistical metrics of the forecasting models for several input data combination and for the 4 methods in April (1 month before harvest). All available input data from October to April were used (see Table 2). The metrics are computed for the 15 provinces.

Input Data	Models	RMSE (t. ha^{-1})	MAE (t. ha^{-1})	R^2
Satellite-based drought indices only	MLR	0.66	0.57	0.67
	SVM	0.54	0.43	0.78
	RF	0.46	0.35	0.80
	XGBoost	0.45	0.34	0.81
Satellite-based drought indices and weather data	MLR	0.46	0.39	0.72
	SVM	0.40	0.31	0.80
	RF	0.34	0.24	0.84
	XGBoost	0.37	0.25	0.86
Satellite-based drought indices, weather data and climate indices	MLR	0.41	0.31	0.75
	SVM	0.25	0.21	0.88
	RF	0.22	0.19	0.92
	XGBoost	0.20	0.16	0.95

3.2. Model Performance as a Function of Lead Time before Harvest

In this section, the performance of the forecasting models using the three datasets are evaluated as a function of the leading time prior to harvest from January to March (from 4 to 2 months before harvest). The RMSEs and R^2 of the models are plotted as a function of the lead time in Figure 3 to investigate the prediction accuracy. In addition, the relative importance of each dataset is reported in Figure 4 using the XGBoost algorithm as the method providing the best statistical metrics for all lead times.

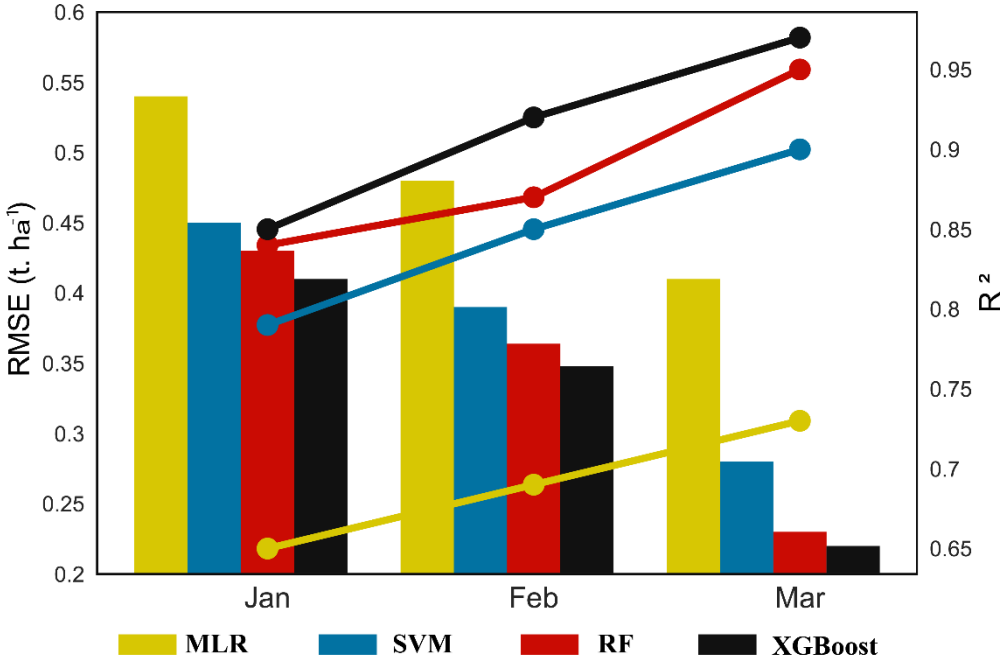

Figure 3. Model performance (R^2 -line- and RMSE -bar-) as a function of the lead time from 4 to 2 months before havers (from January to March) for the four methods (MLR, SVM, RF and XGBoost). All the available predictor variables at the time of prediction were used (see Table 2).

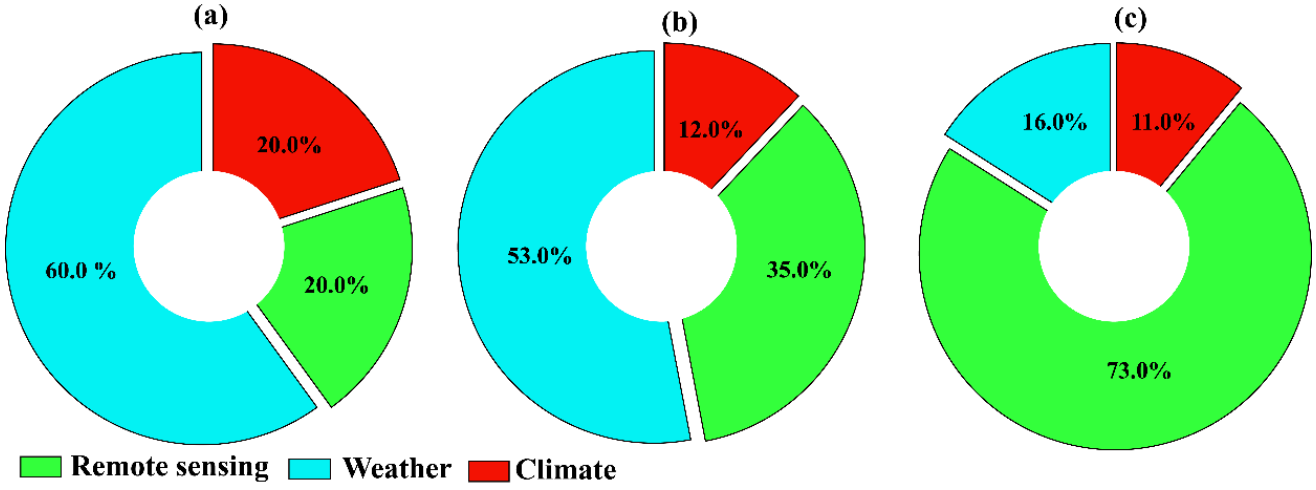

Figure 4. Importance of the different inputs datasets for yield prediction from January (**a**); February (**b**) and March (**c**). The considered predictors variables and their time span of the year for each model are reported in Table 2.

The closer to harvest the forecast is carried out, the better the performance metrics as shown by the increase of the correlation coefficient and the drop of RMSE when going from January to March at Figure 3. The best method whatever the lead time is XGBoost as

already shown, followed closely by RF based approaches. The models based on XGBoost explain 88%, 92% and 96% of yield variability (RMSE of 0.41, 0.34 and 0.22 t. ha^{-1}) for forecasting from January, February and March, respectively. By contrast, the poorest results are obtained with MLR with a strong gap of metrics with regards to the non-linear machine learning approaches (R^2 is below 0.75 for MLR while the correlations for the non-linear methods are above 0.90).

While a slight improvement of the model metrics is observed when going from January to February, considering predictors in March leads to a significant jump in the metrics with an RMSE close to the international standard of 0.20 t. ha^{-1} for the XGBoost method and, to a lesser extent, for the RF model. This is probably related to the very high correlation between NDVI around the crop development peak in March and wheat yields that were already shown by various authors [11,87] giving a large weight to VCI at this time. The dominating importance of the satellite drought indices in March for the model based on XGBoost supports this assumption (Figure 4).

Other striking comments can be made by analyzing the importance of the three datasets (Figure 4): (1) the weather data dominates largely in January and, to a lesser extent, in February, while a strong shift is observed in March when satellite drought indices take the lead over the two other datasets. This is in agreement with the already observed high correlation between yields and precipitation around emergence in October and November and between yields and temperature in December during the tillering stage [11]; (2) Likewise, the importance of climate indices decreases with the lead time and their contribution is the lowest of the three datasets apart from in January when it contributes to 20% like the satellite drought indices. Indeed, the highest correlation with yields was found in December and January for NAO and SCA, respectively, while the correlations with the SST leading modes peak in October and February for Atlantic Niño and Atlantic Tripole modes, respectively. In addition, linkages between climate indices are, in particular, based on SST, and yields occur through teleconnection, meaning that the relationships are very indirect. This means that when good quality precipitation and temperature data are available, they should be preferred to climate indices as they provide more direct information on growing conditions; (3) satellite drought indices play a dominating role in early prediction in March only when they contribute up to 73% to the prediction accuracy. Nevertheless, a significant contribution is observed in February (35%) and in January (20%). This is because VCI and TCI were found to be significantly correlated to final yields in January and February and because SMCI is significantly related to yields as early as October around the emergence stage [10]. Indeed, the high moisture in the upper soil layers at this time facilitates the emergence and significant rainfall event during October–December promotes the farmer to seed, leading to an increase in cereal production [10,12].

In order to assess the practical performance of the developed models to predict yield in Morocco, a "leave-one-year-out" experiment, mimicking the practical forecasting conditions of a manager who wants to predict yields for the season to come based on the historical dataset, is tested. Figure 5 presents the average of the observed and the predicted yields using the three non-linear methods (MLR was excluded with regards to its poorest performance) for a lead time from 4 to 2 months before harvest. As already highlighted, the statistical metrics improve when going from January to March but the models predict yields with reasonable accuracy as early as January. Beyond the average statistical metrics, the ability of the forecasting models to predict extreme values is another important feature of seasonal prediction. Within this context, the ability of the models to predict classified anomalies instead of absolute value is analyzed by partitioning the production in terms of below normal (average minus one standard deviation), normal and above normal (average plus one standard deviation) production. Like the statistical metrics, the extreme anomalies are better predicted when the lead time decreases, as one anomaly is correctly detected by the models for a prediction from January (2006–2007) while all significant anomalies are properly reproduced by the three methods with a slightly better ability of the SVM

approach at the expense of some false detection (such as in 2008–2009 when SVM predicts above normal production).

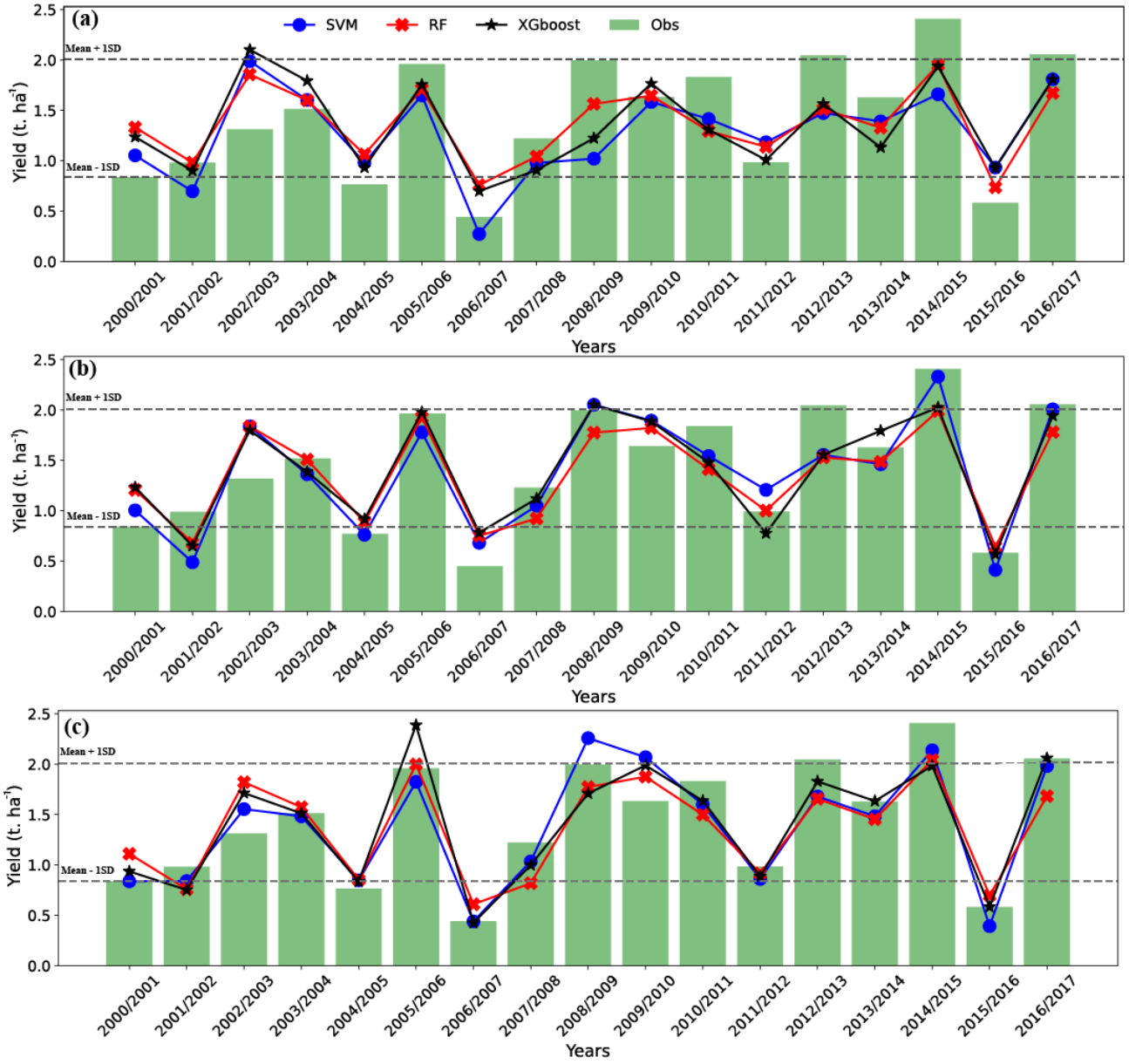

Figure 5. Average of observed and predicted yields using the "leave-one-year-out" technique at (**a**) January, (**b**) February and (**c**) March.

3.3. Model Performance at a Regional Scale

Cereal yields are dependent on many factors, such as weather conditions, local management and soil type, while the importance of each factor varies from one region to another [88]. Therefore, the high variability of crop yield from one season to another is also marked from province to province. To investigate the added value of developing a specific model for each group of provinces separately, Figure 6 compares the RMSE of the predicted yield using a global model and a local model with different lead times using a leave-one-out approach for each province. Only the XGBoost algorithm is retained as it exhibited the best metrics in the previous sections. The results illustrate that the performance of yield prediction improved when the lead time decreases, as already highlighted, and that the metrics show a high variability from one province to another. The use of a "regional"

model improved the RMSE whatever the provinces and the lead time: the RMSE values decrease in January by about 4% to 13% and by 12% to 32% in February and by 12% to 36% in March depending on the province. Interestingly enough, the better performances are obtained for some provinces that are known to be mostly covered by rainfed cereals, such as the ones located along the Atlantic coast (El Jadida JD, Settat ST and Khmisset KM), highlighting the problem of the scale mismatch between the typical size of the fields in the Mediterranean agriculture (<5 ha) and the coarse scale of the input predictor variables.

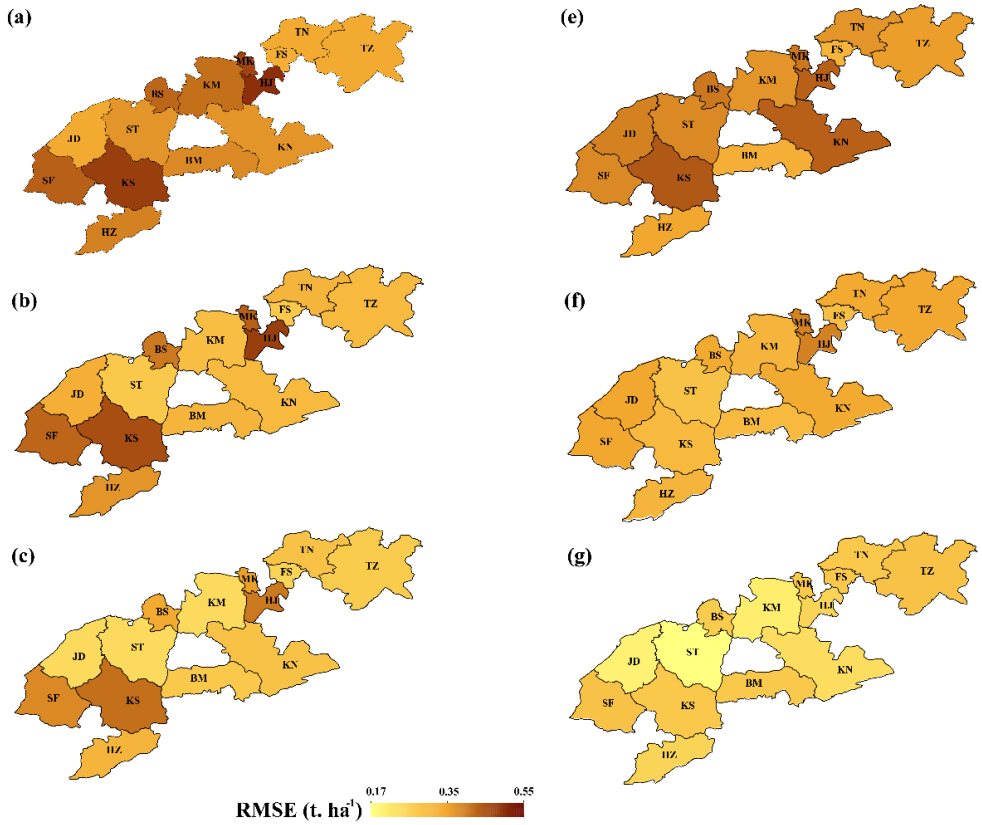

Figure 6. RMSE of global model at (**a**) January, (**b**) February, (**c**) March and regional models at (**e**) January, (**f**) February and (**g**) March.

4. Discussion

The proposed approach is based on a combination of three different datasets to forecast grain yields. From these datasets, the predictors were carefully selected based on statistically significant correlations with grain yields and biophysically sound mechanisms as explained in previous studies [10–12]. Atmospheric and oceanic indices are taken as proxies of local temperature and rainfall conditions. Interestingly enough, the impact of oceanic circulation on local weather can be not concomitant because of the remote nature of these phenomena occurring though teleconnections processes, such as for the Atlantic "El Niño" mode of SST variability. This explains why climate indices are important predictors for very early forecasting in January. Some authors have even based their modeling experiment on large scale climate information only to forecast yield, including Lehmann et al. [45] who showed that climate data (NAO, SST) could be used for the early forecasting (from December) of wheat yield anomaly at the country scale in Morocco. Nevertheless, the direct use of temperature and rainfall data should be preferred to these substitutes when gridded data of good quality exist, as shown by the dominating importance of weather variables for forecasting in January and February. Later in the season in March, when the crops are developed, drought indices providing a direct information on the cover density, health and hydric status obviously took the lead with regards to the other two data sets.

In brief, the satellite drought indices are a potential predictor of cereal yield when the forecasting is done close to harvest, while weather data and climate indices are the key variables for earlier forecasting of cereal yield. Other variables could also be considered to improve the models' skills. Soil type and management practices (water harvesting techniques, complementary irrigation, fertilizing inputs, planting dates etc.), for instance, are key factors for crop growth. While information on management practices is difficult to consider because of a strong variability from one farm to another (apart from the planting date, see below), large scale spatial patterns of soil type could be extracted from global soil maps such as soil grid [89]. For instance, considering information on the soil type could probably improve the performance metrics of the models on Safi, characterized by shallow and pebbly soils with a poor nutrient content, which are significantly lower than its surrounding coastal provinces, such as El Jadida and Settat (see for the global models Figure 4a–c).

The scale mismatch between the scale of the fields (lower than 5 ha) and the coarse resolution of the predictor variables (at best 1 km for the remote sensing drought index) is an important issue as already highlighted. The use of higher spatial and temporal resolution remote sensing data, such as Landsat and Sentinel, could thus improve the performance of the models developed in this study for those provinces with very heterogeneous land cover. For further studies at the field scale, higher resolution products, such as surface soil moisture derived from Sentinel-1 data [90,91] and Sentinel-2 NDVI, should be considered. In addition, cereal yields may be related to other factors that were not considered in our study, such as planting date, soil properties, local climate conditions and other management aspects [92]. In particular, the planting dates can shift the growing season with regards to the average growing period from November to May considered in this study. Local climate conditions can also shift the cereal season. For instance, the milder temperature conditions encountered in the Beni-Mellal province, located in the foothills of the Atlas, shift the cereal season by about one month with a harvest occurring in June on average while May is usually the harvest time for the provinces located in the plain (most of the provinces of our study area). This means that the considered time span of the year of the predictor variables (December for temperature, for instance, see Table 2) could not be optimal for all the provinces because of this time shift. A potential refinement of the models would thus be to consider the optimal time span of the predictor variables for each province or each group of provinces separately (for the last experiment considering a specific model for each group of provinces) instead of the same time period used in this study.

Finally, a last more general question arises about the model generalization to different crops and sites. In this study, the predictors were selected according to both the timing of the crop season and to the key phenological stages of wheat and companion cereals such as barley. As the timing of the crop season is similar for wheat that is usually cropped in winter in the whole north African area, it could be expected that the time span of local predictors, such as weather variables and drought indices, should be close for the other Maghreb countries. By contrast, the impact of oceanic and atmospheric indices on local climate may be different from one region to another. For instance, Tramblay et al. [93] found NAO to be related to rainfall in Morocco and Algeria while Tunisian rainfall was more correlated to the Mediterranean Oscillation (MO; [94]). Ouachani et al. [95] highlighted that ENSO could be a driving pattern of precipitation in Tunisia through teleconnections. This means that the use of other indices, proxies of climate pseudo-oscillations, should be considered for the development of forecasting models for other sites. Likewise, the forecasting of yields for other crops will require a different choice of predictors and their associated time spans according to their key phenological stages. For instance, maize is known to be relatively drought tolerant during the grain filling stages on the contrary to wheat [96]. By contrast, water deficit early in the season around seedling may hamper maize from complete recovery, even with full irrigation, during the vegetative growth stages [97]. This may have critical implications for the choice of the time span of the remote sensing drought indices.

5. Conclusions

Crop yield forecasting provides critical and timely information to enable farmers to make quick decisions to increase yields through improving agricultural practices during the growing season. In addition, it allows the modeling of global and local market prices [98]. The main objective of our study was to develop an approach to forecasting cereal yield in Morocco based on multi-source data and machine learning techniques. To this objective, this study presents a methodology based on different machine learning approaches (MLR, SVM, RF and XGBoost) to predict the cereal yield over Morocco for several lead times prior to harvest using freely available datasets including satellite-based drought indices, weather data and climate indices. Our results show that combining satellite-based drought indices, weather and climate data as predictors of cereal yield provided a better forecasting accuracy than using any single data source. In line with our results, several studies pointed out that the use of multi-source data increases the accuracy of the machine learning model for yield prediction [52,54,56]. XGBoost outperformed the other machine learning techniques by explaining 93% of yield variation at the scale of the country (RMSE = 0.23 t. ha^{-1}). Interestingly enough, the value of RMSE is close to the acceptance threshold of 0.2 t. ha^{-1} used in European agro-statistics [86]. It is also shown that the prediction accuracy increases as more observations along the growing season are added for all machine learning algorithms. Finally, the development of models at the regional level for each group of provinces improved the skills of yield prediction with regards to one "global" model applied to all provinces by decreasing the RMSE by about 4% to 36% depending on the province and the time of prediction, which is due to the high variability of cereal yield from one province to another.

The results presented in this study clearly showed that combining satellite-based drought indices, weather and climate data integrated into machine learning algorithms is a promising approach to forecasting cereal yields in Morocco. Moreover, the proposed approach provides a source of timely information needed for decision making during the growing season. In addition, this work could be used to map gain yield and yield gap at a provincial scale across Morocco. Then, the province with hotspots in terms of yield gap could be targeted for practice improvement and further research works.

Author Contributions: Conceptualization, E.h.B., L.J.; methodology, E.h.B., L.J.; software, E.h.B., L.J. and A.A.; Data curation, E.h.B., L.J. and R.B.; formal analysis, E.h.B. and L.J.; investigation, E.h.B. and L.J.; writing—original draft preparation, E.h.B.; writing—review and editing, L.J., S.E.-R., S.K., A.A., R.B. and B.R.; supervision, L.J. and S.E.-R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was carried out within the framework of the Joint International Laboratory TREMA (http://lmi-trema.ma, accessed on 31 July 2021). This work was funded by the ERANETMED03–62 CHAAMS project, the ACCWA project, grant agreement no: 823965 and by SAGESSE PPR/2015/48. E. Bouras was supported by a fellowship from the ARTS program from IRD, France. The H2020 PRIMA ALTOS project, MISTRALS/SICMED2, PHC Toubkal #39064WG/2018 and PRIMA-IDEWA project are also acknowledged for additional funding.

Acknowledgments: The authors acknowledge the Economic Services of the Ministry of Agriculture of Morocco for providing the crop production statistics. The authors are also grateful for the valuable and constructive comments from the anonymous reviewers.

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

References

1. Asseng, S.; Ewert, F.; Martre, P.; Rötter, R.; Lobell, D.; Cammarano, D.; Kimball, B.A.; Ottman, M.J.; Wall, G.W.; White, J.W.; et al. Rising temperatures reduce global wheat production. *Nat. Clim. Chang.* **2015**, *5*, 143–147. [CrossRef]
2. FAO. *Climate Change and Food Security: Risks and Responses*; Food and Agriculture Organization of The United Nations: Rome, Italy, 2016; ISBN 9789251089989.
3. Giorgi, F. Climate change hot-spots. *Geophys. Res. Lett.* **2006**, *33*, 101029. [CrossRef]

4. Lionello, P.; Scarascia, L. The relation between climate change in the Mediterranean region and global warming. *Reg. Environ. Chang.* **2018**, *18*, 1481–1493. [CrossRef]
5. Vicente-Serrano, S.M.; Quiring, S.M.; Peña-Gallardo, M.; Yuan, S.; Domínguez-Castro, F. A review of environmental droughts: Increased risk under global warming? *Earth Sci. Rev.* **2020**, *201*, 1–23. [CrossRef]
6. Kogan, F.; Guo, W.; Yang, W. Drought and food security prediction from NOAA new generation of operational satellites. *Geomat. Nat. Hazards Risk* **2019**, *10*, 651–666. [CrossRef]
7. Schilling, J.; Hertig, E.; Tramblay, Y.; Scheffran, J. Climate change vulnerability, water resources and social implications in North Africa. *Reg. Environ. Chang.* **2020**, *20*, 1–15. [CrossRef]
8. UN General Assembly. *Transforming Our World: The 2030 Agenda for Sustainable Development*; United Nations: New York, NY, USA, 2015.
9. Bouras, E.; Jarlan, L.; Khabba, S.; Er-Raki, S.; Dezetter, A.; Sghir, F.; Tramblay, Y. Assessing the impact of global climate changes on irrigated wheat yields and water requirements in a semi-arid environment of Morocco. *Sci. Rep.* **2019**, *9*, 1–14. [CrossRef]
10. Bouras, E.H.; Jarlan, L.; Er-Raki, S.; Albergel, C.; Richard, B.; Balaghi, R.; Khabba, S. Linkages between rainfed cereal production and agricultural drought through remote sensing indices and a land data assimilation system: A case study in Morocco. *Remote Sens.* **2020**, *12*, 4018. [CrossRef]
11. Jarlan, L.; Abaoui, J.; Duchemin, B.; Ouldbba, A.; Tourre, Y.M.; Khabba, S.; Le Page, M.; Balaghi, R.; Mokssit, A.; Chehbouni, G. Linkages between common wheat yields and climate in Morocco (1982–2008). *Int. J. Biometeorol.* **2014**, *58*, 1489–1502. [CrossRef]
12. Balaghi, R.; Tychon, B.; Eerens, H.; Jlibene, M. Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco. *Int. J. Appl. Earth Obs. Geoinf.* **2008**, *10*, 438–452. [CrossRef]
13. Lobell, D.B.; Burke, M.B.; Tebaldi, C.; Mastrandrea, M.D.; Falcon, W.P.; Naylor, R.L. Prioritizing climate change adaptation needs for food security in 2030. *Science* **2008**, *319*, 607–610. [CrossRef]
14. Sacks, W.J.; Kucharik, C.J. Crop management and phenology trends in the U.S. Corn Belt: Impacts on yields, evapotranspiration and energy balance. *Agric. For. Meteorol.* **2011**, *151*, 882–894. [CrossRef]
15. Basso, B.; Liu, L. Seasonal crop yield forecast: Methods, applications, and accuracies. *Adv. Agron.* **2019**, *54*, 201–255.
16. Jones, J.W.; Antle, J.M.; Basso, B.; Boote, K.J.; Conant, R.T.; Foster, I.; Godfray, H.C.J.; Herrero, M.; Howitt, R.E.; Janssen, S.; et al. Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. *Agric. Syst.* **2017**, *155*, 269–288. [CrossRef]
17. Lawless, C.; Semenov, M.A. Assessing lead-time for predicting wheat growth using a crop simulation model. *Agric. For. Meteorol.* **2005**, *135*, 302–313. [CrossRef]
18. Wang, X.; Zhao, C.; Li, C.; Liu, L.; Huang, W.; Wang, P. Use of Ceres-wheat model for wheat yield forecast in Beijing. In *Proceedings of the IFIP Advances in Information and Communication Technology*; Springer: Boston, MA, USA, 2009.
19. Li, Z.; Song, M.; Feng, H.; Zhao, Y. Within-season yield prediction with different nitrogen inputs under rain-fed condition using CERES-Wheat model in the northwest of China. *J. Sci. Food Agric.* **2016**, *96*, 2906–2916. [CrossRef]
20. Dumont, B.; Leemans, V.; Ferrandis, S.; Bodson, B.; Destain, J.P.; Destain, M.F. Assessing the potential of an algorithm based on mean climatic data to predict wheat yield. *Precis. Agric.* **2014**, *15*, 255–272. [CrossRef]
21. Hansen, J.W.; Indeje, M. Linking dynamic seasonal climate forecasts with crop simulation for maize yield prediction in semi-arid Kenya. *Agric. For. Meteorol.* **2004**, *125*, 143–157. [CrossRef]
22. Mishra, A.; Hansen, J.W.; Dingkuhn, M.; Baron, C.; Traoré, S.B.; Ndiaye, O.; Ward, M.N. Sorghum yield prediction from seasonal rainfall forecasts in Burkina Faso. *Agric. For. Meteorol.* **2008**, *148*, 1798–1814. [CrossRef]
23. Kogan, F.; Yang, B.; Guo, W.; Pei, Z.; Jiao, X. Modelling corn production in China using AVHRR-based vegetation health indices. *Int. J. Remote Sens.* **2005**, *26*, 2325–2336. [CrossRef]
24. Kogan, F.; Kussul, N.; Adamenko, T.; Skakun, S.; Kravchenko, O.; Kryvobok, O.; Shelestov, A.; Kolotii, A.; Kussul, O.; Lavrenyuk, A. Winter wheat yield forecasting in Ukraine based on Earth observation, meteorologicaldata and biophysical models. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *23*, 192–203. [CrossRef]
25. 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]
26. Meroni, M.; Fasbender, D.; Balaghi, R.; Dali, M.; Haffani, M.; Haythem, I.; Hooker, J.; Lahlou, M.; Lopez-Lozano, R.; Mahyou, H.; et al. Evaluating NDVI Data Continuity Between SPOT-VEGETATION and PROBA-V Missions for Operational Yield Forecasting in North African Countries. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 795–804. [CrossRef]
27. Mathieu, J.A.; Aires, F. Assessment of the agro-climatic indices to improve crop yield forecasting. *Agric. For. Meteorol.* **2018**, *253–254*, 15–30. [CrossRef]
28. Dumont, B.; Basso, B.; Leemans, V.; Bodson, B.; Destain, J.P.; Destain, M.F. A comparison of within-season yield prediction algorithms based on crop model behaviour analysis. *Agric. For. Meteorol.* **2015**, *204*, 10–21. [CrossRef]
29. Basso, B.; Cammarano, D.; Carfagna, E. Review of Crop Yield Forecasting Methods and Early Warning Systems. In Proceedings of the First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics, Rome, Italy, 18–19 July 2013.
30. Sierra, E.M.; Brynsztein, S.M. Wheat yield variability in the S.E. of the Province of Buenos Aires. *Agric. For. Meteorol.* **1990**, *49*, 281–290. [CrossRef]

31. Giri, A.K.; Bhan, M.; Agrawal, K.K. Districtwise wheat and rice yield predictions using meteorological variables in eastern Madhya Pradesh. *J. Agrometeorol.* **2017**, *9*, 366–368.
32. Rembold, F.; Atzberger, C.; Savin, I.; Rojas, O. Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sens.* **2013**, *1*, 5572–5573. [CrossRef]
33. Anderson, M.C.; Zolin, C.A.; Sentelhas, P.C.; Hain, C.R.; Semmens, K.; Tugrul Yilmaz, M.; Gao, F.; Otkin, J.A.; Tetrault, R. The Evaporative Stress Index as an indicator of agricultural drought in Brazil: An assessment based on crop yield impacts. *Remote Sens. Environ.* **2016**, *174*, 82–99. [CrossRef]
34. Salazar, L.; Kogan, F.; Roytman, L. Use of remote sensing data for estimation of winter wheat yield in the United States. *Int. J. Remote Sens.* **2007**, *28*, 3795–3811. [CrossRef]
35. Wang, M.; Tao, F.L.; Shi, W.J. Corn yield forecasting in northeast china using remotely sensed spectral indices and crop phenology metrics. *J. Integr. Agric.* **2014**, *13*, 1538–1545. [CrossRef]
36. 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]
37. Liu, W.T.; Kogan, F. Monitoring Brazilian soybean production using NOAA/AVHRR based vegetation condition indices. *Int. J. Remote Sens.* **2002**, *23*, 1161–1179. [CrossRef]
38. García-León, D.; Contreras, S.; Hunink, J. Comparison of meteorological and satellite-based drought indices as yield predictors of Spanish cereals. *Agric. Water Manag.* **2019**, *213*, 388–396. [CrossRef]
39. Nguyen-Huy, T.; Deo, R.C.; An-Vo, D.A.; Mushtaq, S.; Khan, S. Copula-statistical precipitation forecasting model in Australia's agro-ecological zones. *Agric. Water Manag.* **2017**, *191*, 153–172. [CrossRef]
40. Ceglar, A.; Turco, M.; Toreti, A.; Doblas-Reyes, F.J. Linking crop yield anomalies to large-scale atmospheric circulation in Europe. *Agric. For. Meteorol.* **2017**, *240–241*, 35–45. [CrossRef]
41. Wang, B.; Feng, P.; Waters, C.; Cleverly, J.; Liu, D.L.; Yu, Q. Quantifying the impacts of pre-occurred ENSO signals on wheat yield variation using machine learning in Australia. *Agric. For. Meteorol.* **2020**, *291*, 108043. [CrossRef]
42. Knippertz, P.; Christoph, M.; Speth, P. Long-term precipitation variability in Morocco and the link to the large-scale circulation in recent and future climates. *Meteorol. Atmos. Phys.* **2003**, *83*, 67–88. [CrossRef]
43. Podestá, G.; Letson, D.; Messina, C.; Royce, F.; Ferreyra, R.A.; Jones, J.; Hansen, J.; Llovet, I.; Grondona, M.; O'Brien, J.J. Use of ENSO-related climate information in agricultural decision making in Argentina: A pilot experience. *Agric. Syst.* **2002**, *74*, 371–392. [CrossRef]
44. Martinez, C.J.; Baigorria, G.A.; Jones, J.W. Use of climate indices to predict corn yields in southeast USA. *Int. J. Climatol.* **2009**, *29*, 1680–1691. [CrossRef]
45. Lehmann, J.; Kretschmer, M.; Schauberger, B.; Wechsung, F. Potential for Early Forecast of Moroccan Wheat Yields Based on Climatic Drivers. *Geophys. Res. Lett.* **2020**, *41*, 1–10. [CrossRef]
46. Cai, X.L.; Sharma, B.R. Integrating remote sensing, census and weather data for an assessment of rice yield, water consumption and water productivity in the Indo-Gangetic river basin. *Agric. Water Manag.* **2010**, *97*, 309–316. [CrossRef]
47. Van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
48. Abbas, F.; Afzaal, H.; Farooque, A.A.; Tang, S. Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy* **2020**, *10*, 1046. [CrossRef]
49. Cao, J.; Zhang, Z.; Tao, F.; Zhang, L.; Luo, Y.; Han, J.; Li, Z. Identifying the contributions of multi-source data for winter wheat yield prediction in China. *Remote Sens.* **2020**, *12*, 750. [CrossRef]
50. 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]
51. 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]
52. Kang, Y.; Ozdogan, M.; Zhu, X.; Ye, Z.; Hain, C.; Anderson, M. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environ. Res. Lett.* **2020**, *15*, 064005. [CrossRef]
53. Mateo-Sanchis, A.; Piles, M.; Muñoz-Marí, J.; Adsuara, J.E.; Pérez-Suay, A.; Camps-Valls, G. Synergistic integration of optical and microwave satellite data for crop yield estimation. *Remote Sens. Environ.* **2019**, *234*, 111460. [CrossRef]
54. 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]
55. Schwalbert, R.A.; Amado, T.; Corassa, G.; Pott, L.P.; Prasad, P.V.V.; Ciampitti, I.A. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agric. For. Meteorol.* **2020**, *284*, 107886. [CrossRef]
56. 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]
57. Driouech, F.; Déqué, M.; Mokssit, A. Numerical simulation of the probability distribution function of precipitation over Morocco. *Clim. Dyn.* **2009**, *2*, 1055–1063. [CrossRef]

58. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An. Introduction to Cluster Analysis*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 1990; ISBN 0471735787.
59. West, H.; Quinn, N.; Horswell, M. Remote sensing for drought monitoring & impact assessment: Progress, past challenges and future opportunities. *Remote Sens. Environ.* **2019**, *232*, 111291. [CrossRef]
60. Kogan, F.N. Application of vegetation index and brightness temperature for drought detection. *Adv. Space Res.* **1995**, *15*, 91–100. [CrossRef]
61. Zhang, A.; Jia, G. Monitoring meteorological drought in semiarid regions using multi-sensor microwave remote sensing data. *Remote Sens. Environ.* **2013**, *134*, 12–23. [CrossRef]
62. Jiao, W.; Tian, C.; Chang, Q.; Novick, K.A.; Wang, L. A new multi-sensor integrated index for drought monitoring. *Agric. For. Meteorol.* **2019**, *268*, 74–85. [CrossRef]
63. Bento, V.A.; Trigo, I.F.; Gouveia, C.M.; DaCamara, C.C. Contribution of Land Surface Temperature (TCI) to Vegetation Health Index: A comparative study using clear sky and all-weather climate data records. *Remote Sens.* **2018**, *10*, 1324. [CrossRef]
64. Dorigo, W.; Wagner, W.; Albergel, C.; Albrecht, F.; Balsamo, G.; Brocca, L.; Chung, D.; Ertl, M.; Forkel, M.; Gruber, A.; et al. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sens. Environ.* **2017**, *203*, 185–215. [CrossRef]
65. Heng, L.K.; Asseng, S.; Mejahed, K.; Rusan, M. Optimizing wheat productivity in two rain-fed environments of the West Asia-North Africa region using a simulation model. *Eur. J. Agron.* **2007**, *26*, 121–129. [CrossRef]
66. Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **2020**, *146*, 1999–2049. [CrossRef]
67. Barnston, A.G.; Livezey, R.E. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Mon. Weather Rev.* **1987**, *115*, 1083–1126. [CrossRef]
68. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
69. Henry, N.W.; Cohen, J.; Cohen, P. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. *Contemp. Sociol.* **1977**, *6*, 320. [CrossRef]
70. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
71. Sain, S.R.; Vapnik, V.N. The Nature of Statistical Learning Theory. *Technometrics* **1996**, *38*, 409. [CrossRef]
72. Gunn, S. Support Vector Machines for classification and regression. *Analyst* **1998**, *135*, 230–267. [CrossRef]
73. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
74. Kuter, S. Completing the machine learning saga in fractional snow cover estimation from MODIS Terra reflectance data: Random forests versus support vector regression. *Remote Sens. Environ.* **2021**, *255*, 112294. [CrossRef]
75. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
76. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
77. Song, Y.; Liu, X.; Zhang, L.; Jiao, X.; Qiang, Y.; Qiao, Y.; Liu, Z. Prediction of double-high biochemical indicators based on lightGBM and XGBoost. In Proceedings of the ACM International Conference Proceeding Series, Wuhan, China, 12–13 July 2019.
78. Kaneko, H.; Funatsu, K. Fast optimization of hyperparameters for support vector regression models with highly predictive ability. *Chemom. Intell. Lab. Syst.* **2015**, *142*, 64–69. [CrossRef]
79. Picard, R.R.; Cook, R.D. Cross-validation of regression models. *J. Am. Stat. Assoc.* **1984**, *79*, 575–583. [CrossRef]
80. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Int. Jt. Conf. Artif. Intell.* **1995**, *2*, 1137–1143.
81. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
82. Baumann, K. Cross-validation as the objective function for variable-selection techniques. *TrAC Trends Anal. Chem.* **2003**, *22*, 395–406. [CrossRef]
83. Faroux, S.; Kaptué Tchuenté, A.T.; Roujean, J.-L.; Masson, V.; Martin, E.; Le Moigne, P. ECOCLIMAP-II/Europe: A twofold database of ecosystems and surface parameters at 1 km resolution based on satellite information for use in land surface, meteorological and climate models. *Geosci. Model. Dev.* **2013**, *6*, 563–582. [CrossRef]
84. ESA. *Land Cover CCI Product User Guide Version 2.0*; ESA: Paris, France, 2017. Available online: http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf (accessed on 31 July 2021).
85. Rasouli, K.; Hsieh, W.W.; Cannon, A.J. Daily streamflow forecasting by machine learning methods with weather and climate inputs. *J. Hydrol.* **2012**, *414–415*, 284–293. [CrossRef]
86. Genovese, G.P.; Fritz, S.; Bettio, M. A comparison and evaluation of performances among crop yield forecasting models based on remote sensing: Results from the geoland observatory of food monitoring. *Int. Arch. Photogramm. Remote Sens. Spacial Inf. Sci.* **2006**, *36*, 71–77.
87. Belaqziz, S.; Khabba, S.; Er-Raki, S.; Jarlan, L.; Le Page, M.; Kharrou, M.H.; Adnani, M.E.; Chehbouni, A. A new irrigation priority index based on remote sensing data for assessing the networks irrigation scheduling. *Agric. Water Manag.* **2013**, *119*, 1–9. [CrossRef]
88. Satir, O.; Berberoglu, S. Crop yield prediction under soil salinity using satellite derived vegetation indices. *Field Crop. Res.* **2016**, *192*, 134–143. [CrossRef]

89. Hengl, T.; De Jesus, J.M.; Heuvelink, G.B.M.; Gonzalez, M.R.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef]
90. El Hajj, M.; Baghdadi, N.; Zribi, M.; Belaud, G.; Cheviron, B.; Courault, D.; Charron, F. Soil moisture retrieval over irrigated grassland using X-band SAR data. *Remote Sens. Environ.* **2016**, *176*, 202–218. [CrossRef]
91. Ouaadi, N.; Jarlan, L.; Ezzahar, J.; Zribi, M.; Khabba, S.; Bouras, E.; Bousbih, S.; Frison, P.-L. Monitoring of wheat crops using the backscattering coefficient and the interferometric coherence derived from Sentinel-1 in semi-arid areas. *Remote Sens. Environ.* **2020**, *251*, 112050. [CrossRef]
92. Zhang, Z.; Jin, Y.; Chen, B.; Brown, P. California almond yield prediction at the orchard level with a machine learning approach. *Front. Plant. Sci.* **2019**, *10*, 809. [CrossRef] [PubMed]
93. Tramblay, Y.; El Adlouni, S.; Servat, E. Trends and variability in extreme precipitation indices over maghreb countries. *Nat. Hazards Earth Syst. Sci.* **2013**, *13*, 3235–3248. [CrossRef]
94. Conte, M.; Giuffrida, A.; Tedesco, S. *The Mediterranean Oscillation. Impact on Precipitation and Hydrology in Italy Climate Water*; Academy of Finland: Helsinki, Fenland, 1989.
95. Ouachani, R.; Bargaoui, Z.; Ouarda, T. Power of teleconnection patterns on precipitation and streamflow variability of upper Medjerda Basin. *Int. J. Climatol.* **2013**, *33*, 58–76. [CrossRef]
96. Kang, S.; Shi, W.; Zhang, J. An improved water-use efficiency for maize grown under regulated deficit irrigation. *Field Crop. Res.* **2000**, *67*, 207–214. [CrossRef]
97. Song, L.; Jin, J.; He, J. Effects of severe water stress on maize growth processes in the field. *Sustainability* **2019**, *11*, 5086. [CrossRef]
98. Peng, Y.H.; Hsu, C.S.; Huang, P.C. Developing crop price forecasting service using open data from Taiwan markets. In Proceedings of the TAAI 2015—2015 Conference on Technologies and Applications of Artificial Intelligence, Tainan, Taiwan, 20–22 November 2015.