Received: 9 July 2021 | Revised: 16 November 2022 | Accepted: 21 November 2022 | Published on: 20 December 2022

DOI: 10.1002/joc.7947

RESEARCH ARTICLE

International Journal of Climatology RMetS

Predicting crop yields in Senegal using machine learning methods

Alioune Badara Sarr[1,2] | **Benjamin Sultan**[2]

[1]Laboratoire d'Océanographie, des Sciences de l'Environnement et du Climat (LOSEC), UFR Sciences et Technologies, Université A. SECK de Ziguinchor, Ziguinchor, Senegal

[2]IRD-ESPACE-DEV, Maison de la Télédétection, Montpellier Cedex, France

Correspondence
Alioune Badara Sarr, Laboratoire d'Océanographie, des Sciences de l'Environnement et du Climat (LOSEC), UFR Sciences et Technologies, Université A. SECK de Ziguinchor, Ziguinchor, Senegal.
Email: aliounebadara.sarr@ird.fr

Funding information
Make Our Planet Great Again

Abstract

Agriculture plays an important role in Senegalese economy and annual early warning predictions of crop yields are highly relevant in the context of climate change. In this study, we used three main machine learning methods (support vector machine, random forest, neural network) and one multiple linear regression method, namely Least Absolute Shrinkage and Selection Operator (LASSO), to predict yields of the main food staple crops (peanut, maize, millet and sorghum) in 24 departments of Senegal. Three combination of predictors (climate data, vegetation data or a combination of both) are used to compare the respective contribution of statistical methods and inputs in the predictive skill. Our results showed that the combination of climate and vegetation with the machine learning methods gives the best performance. The best prediction skill is obtained for peanut yield likely due to its high sensitivity to interannual climate variability. Although more research is needed to integrate the results of this study into an operational framework, this paper provides evidence of the promising performance machine learning methods. The development and operationalization of such prediction and their integration into operational early warning systems could increase resilience of Senegal to climate change and contribute to food security.

KEYWORDS

climate change scenario, crop yield prediction, machine learning, Senegal

1 | INTRODUCTION

Senegalese agriculture occupies 12% of the national territory and constitutes the economic base of the country (Rapport National sur le Développement Humain au Sénégal, 2009). The agriculture sector is one of the pillars of the long-term economic development strategy called the Emerging Senegal Plan (Plan Sénégal Émergent) whose objectives are aligned with climate policies (e.g., NDC, sectoral National Determined Contributions). However, despite an increase of crop yields over the last decades, year-to-year production remains heavily dependent on climate variability. Indeed, the severity of climate impacts in this part of the world is particularly strong since rainfed agriculture is the main source of food and income and since the means to control the crop environment (irrigation, mechanization, fertilizers and other off-farm inputs) are largely unavailable to small-scale farmers

(Ingram *et al.*, 2002; Sultan *et al.*, 2010). The anthropic global warming has the potential to aggravate the severity of these impacts with a warmer and drier climate expected in the country under climate change scenarios (Sultan *et al.*, 2014). Indeed, numerous studies have found that such climate change leads to crop production losses of cereals (millet, sorghum, maize) in West Africa and more variable yields which indicate a greater risk of crop failures under a warmer climate (Jones and Thornton, 2003; Schlenker and Lobell, 2010; Sultan *et al.*, 2013; 2014; Sultan and Gaetani, 2016).

In this context, the development of operational, timely and accurate early warning systems to assist decision-making are needed more than ever to build resilience and save lives by allowing communities to prepare and, if necessary, accelerates international food aids or to take advantage of favourable climate conditions. Numerous national and international early warning systems (Funk *et al.*, 2019) support the early identification of emerging droughts and food crises: Famine Early Warning Systems Network (FEWS NET), World Food Programme (WFP) Food Security Analysis, European Commission Monitoring Agriculture Resources (MARS) among others. In the Sahel, the unprecedented droughts of the early 1970s and their severe impacts on population (thousands to millions of inhabitants affected) initiated the emergence of several early warning systems in the region (Traoré *et al.*, 2010). Among them, the CILSS/AGRHYMET (French acronym for Permanent Inter-state Committee for Drought Control in the Sahel/Regional Center for training and applications in operational AGRicultural METeorology and HYdrology), which uses agrometeorological information to warn stakeholders on the possibility of food crops failure (Traoré *et al.*, 2010). These early warning systems draw from many sources of data and model outputs: remote sensing, climate observations, agroclimatic monitoring and hydrologic measurements. Crop models can translate these data into crop yields using either process-based crop models or empirical crop models (see Di Paola *et al.*, 2016 for a review) and can even use climate forecasts as inputs to provide information early enough to adjust critical agricultural decisions and increase agricultural efficiency (Hansen, 2002). Over the past 30 years, satellite and other data, along with the improvement of weather and climate predictions, have strongly improved the accuracy, timeliness and affordability of such early warning systems (WASP, 2021).

However, the ever-increasing amounts of environmental data call for the use of advanced methods such as machine learning methods that can make the best use of various data from satellite to climate models outputs. Along with the improvement of weather and climate predictions, machine learning techniques have also advanced considerably over the past several decades. Machine learning algorithms use large historical data as input to predict new output values. There are a lot of widely used machine learning techniques such as multiple regressions, neural networks or decision trees used. These techniques are distinct from classical statistic approaches by being much more focused on prediction of outcomes rather than taking into account causal relationships or mechanistic processes generating those outcomes (Crane-Droesch, 2018). Several recent studies demonstrated the potential of machine learning on crop yield prediction and climate change impact assessment in agriculture. Crane-Droesch (2018) used a variant of a deep neural network and showed that this approach outperforms classical statistical methods in predicting corn yields in US Midwest. Paudel *et al.* (2021) combined principles of crop modelling with machine learning for crop yield forecasting in Europe. Cai *et al.* (2019) used the LASSO regression method and three mainstream machine learning methods (support vector machine, random forest and neural network) to predict crop yields in Australia. They combined climate data (precipitation, minimum and maximum temperatures and solar radiation) with satellite data such as vegetation index data (normalized difference vegetation index [NDVI] or enhanced vegetation index [EVI]) (Chang *et al.*, 2007; Wardlow *et al.*, 2008; Holzman *et al.*, 2014) and solar-induced chlorophyll fluorescence (SIF) data (Li *et al.*, 2018; Wei *et al.*, 2019; Yao *et al.*, 2021) as inputs for their advanced statistical approaches. Cai *et al.* (2019) found that the best results are obtained by combining both climate data and remote sensing data can capture plant growth using multiple spectral bands (Guan *et al.*, 2017; Cai *et al.*, 2019). Regression methods used in machine learning seem to be well suited for crop predictions (Jeong *et al.*, 2016; Nigam *et al.*, 2019) as they allow for understanding and resolving the complex relationships between crop data and climate and satellite variables (Kim and Lee, 2016). In the Sahel, promising predicting skills of early and end-of-season maize yields in Burkina Faso were found by Leroux *et al.* (2020) by applying a multiple linear model and a random forest model.

Here, this study investigates the predictability of yields of the main staple food crops in Senegal (peanut, maize, millet and sorghum) using four advanced regression methods (three commonly used machine learning methods, that is, support vector machine, random forest, neural network and one multiple linear regression method namely least absolute shrinkage and selection operator) with climate and vegetation data as inputs. We closely follow the approach of Cai *et al.* (2019) which found

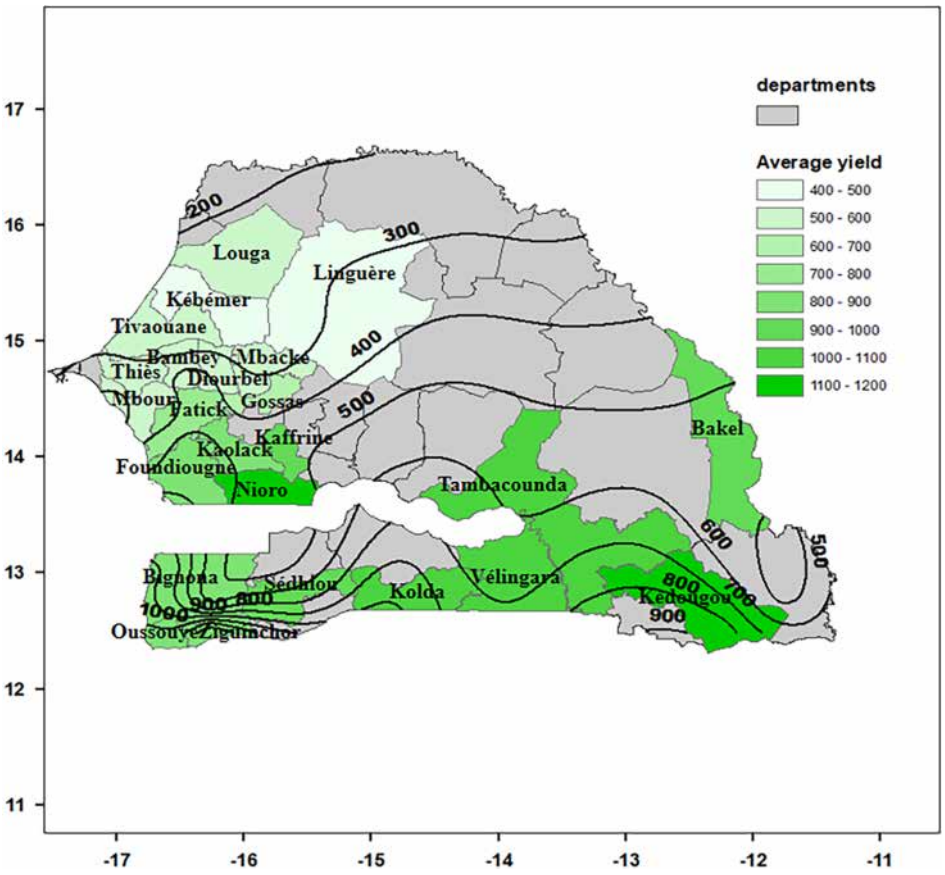

FIGURE 1 Map of studied area. Selected departments with crop yield data (in green). The department in grey were not used in this study (no crop yield data). Shaded values represent the 2000–2013 mean of crop yields of maize, millet, sorghum and peanut (kg·ha^{-1}). Contour lines represent the annual rainfall (mm·year^{-1}) based on the rain gauge stations network of the National Agency for Civil Aviation and Meteorology during summer (June–September) from 1991 to 2005 [Colour figure can be viewed at wileyonlinelibrary.com]

promising results from the use of machine learning approaches to predict yields but in a different geographic location. The skill of each method is discussed in regards to other published studies but also in regards to the climate change scenarios in the region. Indeed, since regional climate in Senegal is expected to change under global warming scenarios, one can expect changes not only in future crop yields but also in crop yields predictability.

2 | MATERIALS AND METHODS

2.1 | The studied area

This study is focused on Senegal which is located in the westernmost part of the African continent. It extends in latitude between 12.8°N and 16.41°N and in longitude between 11.21°W and 17.32°W with a relief generally flat. Like most Sahelian countries, the climate in Senegal is characterized by an unimodal rainfall regime with a single rainy season from June to September called summer period and a long dry season from October to May. Annual rainfall amounts follow a north–south gradient with a semi-arid climate in the north and a tropical climate in the south (Sultan and Janicot, 2003). The annual cumulative rainfall ranges between 200 and 400 mm·year^{-1} in the north, between 400 and 800 mm·year^{-1} in the centre and between 800 and 1,200 mm·year^{-1} in the south of the country (Figure 1). The maximum average yield in peanut, maize, millet and sorghum is generally localized in the western centre and in the south of the country (Figure 1). Agriculture in Senegal accounts for 17.5% of gross domestic product (GDP), 36% of overall exports and is about 95% rainfed (https://agriculture.gouv.fr/senegal) (Touré *et al.*, 2020). Furthermore, the agriculture sectors plays a significant role in the livelihood and economy of the country, especially the cultivation of peanuts, maize, millet and sorghum (Touré *et al.*, 2020). It contributes to feeding the rural environment and the city by providing 60% of the foodstuffs (Ngom, 2014; Ndiaye, 2018). The soils type of the studies areas are generally ferruginous soils weakly leached in the centre and ferruginous and leached in the south (Gavaud, 1988).

2.2 | Data

To carry out this study, we used agronomic data, climate observations and simulations and vegetation data from satellite retrievals (Table 1). All these data have been interpolated to 0.5° spatial resolution using a bilinear

TABLE 1 Description of data used for crop yield prediction

Category	Variables	Spatial resolution	Temporal resolution	Time coverage	Sources/description
Crop yield and area for crop yield	Crop yield (peanut, maize, millet, sorghum) Crop area	Department scale 0.5°	Yearly	1981–2013 2000	DAPSA (Touré *et al.*, 2020) MIRCA2000 (Portmann *et al.*, 2010)
Satellite data for vegetation dynamics	NDVI	0.05°	16-day	Feb 2000–Dec 2014	MODIS MODI3C1 (Heck *et al.*, 2019)
Climate data	Precipitation, temperature and VPD related variables, including cloud cover percentage (Cld), potential evapotranspiration (Pet) and wet day frequency (Wet), VPD is calculated based on *T* and Vap Direct/diffuse flux of shortwave radiation (SDr/SDf), direct/diffuse flux of surface PAR (PDr/PDf)	0.5° 1°	Monthly Monthly	1901–2018 Mar 2000–Dec 2014	CRU (Harris *et al.*, 2014) CERES SYN1deg (Wielicki *et al.*, 1996)

interpolation method (Kim *et al.*, 2019) and convert at month interval.

2.2.1 | Agronomic data

The agronomic data used in this study come from the Direction de l'Analyse, de la Prévision et des Statistiques Agricoles (DAPSA). They contain annual data of surface, production and yield of peanut, maize, millet, sorghum for 24 departments of Senegal (Figure 1). They were selected because of their high use and availability in several departments of Senegal. According to DAPSA, from 2013 to 2017, rice (mainly grown on the river valley in the north) is the most important cereal production (about 771,682 tons), followed by millet (about 640,170 tons), maize (about 293,065 tons) and sorghum (about 155,274 tons). As for legume production, peanut is the most cultivated (about 985,695 tons) over the period 2013–2017 followed by cowpea (about 78,836 tons). These data cover the period 1981–2013 with a percentage of missing values over the period extracted (2000–2013) of 0% for peanut, 27.38% for maize, 12.50% for millet and 14.28% for sorghum. Only yield data was used as a predictand for machine learning methods.

The localization of cultivated areas in each department was not available in the DAPSA dataset. We thus used the MIRCA2000 dataset downloaded from http://www.geo.uni-frankfurt.de/ipg/ag/dl/forschung/MIRCA/index.html (Portmann *et al.*, 2010) which gives maps of the cultivated areas at a resolution of 30 arc min. The MIRCA2000 is a global dataset of monthly irrigated and rainfed crop areas developed around the year 2000. It is based on the assumption that the location of the different crops remains constant from year to year because we only have 1 year of data. It has been used in Africa on rice crops (van Oort and Zwart, 2018), in West Africa on cereal crops such as maize, millet and sorghum (Egbebiyi *et al.*, 2019). We used maps of cultivated areas of peanut, millet, sorghum and maize as a mask (one mask per crop) to extract climate and satellite data. Finally, we used an administrative shapefile to aggregate climate and satellite data at the department level so that all datasets, predictands and predictors (climate, satellite data and yield data) are at the same spatial scale.

2.2.2 | Vegetation data

This study used the vegetation data from MODIS MOD13C1. These are normalized difference vegetation index (NDVI) data that have 16-days temporal resolution and a 0.05° spatial resolution (Heck *et al.*, 2019). Data were extracted from February 2000 to December 2013 at the following link: https://modis.gsfc.nasa.gov/data/dataprod/mod12.php. It is an important vegetation index because it can capture plant growth using multiple spectral bands (Guan *et al.*, 2017). It has been widely used in Africa to predict crop yield (Petersen, 2018; Gcayi *et al.*, 2019).

2.2.3 | Climate data

To study the links between crop and climate, we selected nine climate variables from the Climatic Research Unit (CRU; Jones and Harris, 2013): precipitation, minimum and maximum temperature, mean temperature, cloud cover percentage, potential evapotranspiration and wet day frequency, vapour pressure and vapour pressure deficit (VPD). All these variables, except VPD, are available for free (https://sites.uea.ac.uk/cru/data/) at the monthly timescale in average over a 0.5° grid. They were extracted from 1901–2018 (Table 1).

VPD was calculated from vapour pressure (Vap) and mean air temperature (T) according the following formula:

$$\begin{cases} \text{VPD} = e_{\text{sat}} - \text{Vap} \\ e_{\text{sat}} = 6.108 \times e^{\left(\frac{17.7 \times T}{273.3+T}\right)} \end{cases}, \quad (1)$$

where e_{sat} is the saturated water vapour pressure (in hPa) calculated from the Claussius–Clapeyron equation.

In addition, we used four radiation related variables from SYN1deg (Wielicki *et al.*, 1996): direct/diffuse flux of shortwave radiation (SDr/SDf), direct/diffuse flux of surface of photosynthetic active radiation (PAR) (PDr/PDf). These data which come from the satellite sensor of Clouds and the Earth's Radiant Energy System (CERES) is widely used to estimate crop yield (Awad, 2019; Cai *et al.*, 2019). Their horizontal resolution is 1° and they were extracted from March 2000 to December 2013.

2.2.4 | Climate simulations

In order to assess the predictability of crop yield in the future, we used outputs of 18 climate models from the sixth Coupled Model Intercomparison Project (CMIP6; Eyring *et al.*, 2016). The description of these data is presented in Table 2. Data were first bias-corrected using CDF-t method following the protocol described by Famien *et al.* (2018), then re-scaled by month. This bias correction method is largely used in Africa and worldwide both as a statistical downscaling model and as a bias correction method (Vigaud *et al.*, 2013; Vrac and Ayar, 2016; Lanzante *et al.*, 2019). These data are available in netcdf format in the CICLAD platform (https://mesocentre.ipsl.fr/). The horizontal resolution of these data is 0.5°. They cover the period 1951–2014 for the historical and 2015–2100 for the future. The global climate models are forced for the future by two Shared Socioeconomic Pathways SSP2-4.5 and SSP5-8.5 which correspond respectively to the medium and pessimistic scenarios (Eyring *et al.*, 2016). These data were extracted at the departments scale using nearest neighbour method.

TABLE 2 CMIP6 models used in the study

Model	Institute	References
ACCES-CM2	CSIRO-ARCCSS	Bi *et al.* (2013)
ACCESS-ESM1-5	CSIRO	Law *et al.* (2017)
CanESM5	CCCma	Swart *et al.* (2019)
CNRM-CM6-1	CNRM-CERFACS	Voldoire (2019a)
CNRM-CM6-1-HR	CNRM-CERFACS	Voldoire (2019b)
CNRM-ESM2	CNRM-CERFACS	Séférian *et al.* (2019)
FGOALS-g3	CAS	Li *et al.* (2020)
GFDL-CM4	GFDL	Held *et al.* (2019)
GFDL-ESM4	GFDL	Dunne *et al.* (2020)
INM-CM4-8	INM	Volodin *et al.* (2018)
INM-CM5-0	INM	Volodin *et al.* (2017)
IPSL-CM6A-LR	IPSL	Dufresne *et al.* (2013)
KACE-1-0-G	NIMS-KMA	Lee *et al.* (2019)
MIROC6	MIROC	Tatebe *et al.* (2019)
MIROC-ES2L	MIROC	Hajima *et al.* (2020)
MPI-ESM1-2-HR	MPI	Müller *et al.* (2018)
NORESM2-LM	NCC	Bentsen *et al.* (2013)
UKESM1-0-LL	MOHC	Sellar *et al.* (2019)

Two periods are considered on the future: the near future (2036–2065) and the far future (2071–2100).

Future rainfall anomalies are calculated in percentage over the June–September period using the following formula:

$$\text{Precipitation anomaly} = \frac{P_{\text{fut}\_i} - \overline{P_{\text{res}}}}{\overline{P_{\text{res}}}} \times 100, \quad (2)$$

where P_{fut_i} is the mean rainfall of the 24 departments over the year i during the future and $\overline{P_{\text{res}}}$ is the mean rainfall over the 24 departments averaged over the present time (2000–2013).

2.3 | Methods

2.3.1 | Preprocessing of data

Climate and vegetation datasets were first spatially interpolated at 0.5° spatial resolution and aggregated at

monthly timescale. Climate and satellite data are then averaged for each department using nearest neighbour method by taking into account the cultivated areas of each crop with the crop mask from MIRCA2000.

The time series of the 24 departments of the different data types are extracted over the common period of 14 years from 2000 to 2013. The matrixes are constructed by associating the predictors (climate+satellite data) with each response (peanut, maize, millet and sorghum) for each month (June, July, August and September). This gives us matrixes of dimensions 336 rows × 57 columns (Figure S1, Supporting Information).

The outliers found in this data are deleted by using the following confidence interval (CI) (Vinutha *et al*., 2018):

$$\mathrm{CI} = (25\text{th percentile} - 1.5 \times \mathrm{IQR},\ \ 75\text{th percentile} + 1.5 \times \mathrm{IQR}), \tag{3}$$

where $\mathrm{IQR} = 75\text{th percentile} - 25\text{th percentile}$ represents the interquantile.

Values outside CI are considered as outliers and are replaced by missing values.

The predictors data (climate and satellite data) were then normalized to have a standard deviation of 1 and a mean of 0. After treatment, we obtained respective percentages of missing values of 1.45, 2.09, 1.64 and 1.59% for matrixes constructed with peanut, maize, millet and sorghum. While there is no established threshold from the literature regarding an acceptable percentage of missing data in a data set for valid statistical inferences, it is commonly admitted that a missing rate of 5% or less has no impacts on findings (Schafer, 1999; Bennett, 2001).

2.3.2 | Selecting predictors for regression

Reducing the number of predictors is important before applying machine learning. There are several feature selection methods designed to identify irrelevant and redundant parameters that do not contribute significantly to the accuracy of predictive models. Applying such methods help to significantly improve accuracy, reduce learning times and simplify learning results. Here we used a neighbourhood component analysis (NCA; Rasmussen *et al*., 1996; Lichman, 2013) which is a non-parametric method of predictors selection with the aim of maximizing prediction accuracy of regression and classification algorithms (Yang *et al*., 2012). The NCA is based on a leave-one-out regression which enables to compute predictors weights for minimization of an objective function that measures the average leave-one-out regression loss. Highest values of weight are obtained by the most relevant predictors while the weight of the irrelevant predictors is close to zero. We applied this method independently to each type of crop we aim to predict (peanut, maize, millet and sorghum) in order to select the best predictors among the full list of 13 climate variables described in section 2.2.3. Only few climate variables were found to be irrelevant by the NCA. For instance, for peanut the potential evapotranspiration, the diffuse flux of surface PAR and the diffuse flux of shortwave radiation were not selected. For maize, only VPD was not selected. Concerning the millet, irrelevant predictors are the diffuse flux of surface PAR and the direct flux of shortwave radiation. The nonselected variables for sorghum are the mean temperature and the direct flux of surface PAR. Most of those variables presented nonsignificant correlations with crop yields and/or a high correlation with other variables, thus adding barely no information in the predictive model.

2.3.3 | The relative contribution of predictors

Although the weight of the different predictors can hardly be interpreted in terms of physical link in machine learning methods, we investigated the relative contribution (expressed in %) of the different explaining variables to the predictive model. For the Lasso regression we compute the relative *F*-value which is the ratio between the mean square and mean square error. For the neural network, the relative importance of a predictor is computed with respect to the output neuron (see Ibrahim, 2013 for more details). Decision trees are used for the random forest method where the algorithm computes estimate of predictor importance by summing these estimates over all weak learners in the ensemble. Finally, although relative contribution is not easy to determine with the SVM parameters unless the kernel function is linear, we estimated the features relative importance by using the neighbourhood component analysis (the "fscnca" function in MATLAB).

2.4 | Machine learning methods for estimating crop yield

We use three commonly used machine learning methods, namely support vector machine (SVM), random forest (RF) and neural network (NN), to predict agriculture yield (peanut, maize, millet and sorghum). In addition to these machine learning algorithms, we used a linear regression method called LASSO. Several studies have

been conducted with these same methods in a similar context (Cai *et al*., 2019; Abbas *et al*., 2020). The input data for these models are agronomic data (predictands), satellite and climate data (predictors).

Following the approach described by Cai *et al.* (2019), we first randomly divided the whole dataset into a training dataset (70% of the dataset) and a test dataset (the remaining 30% of the dataset) to generate one predicted R^2. We then apply a fivefold cross-validation only on the training data to find the optimal hyperparameters for each method from empirical candidates based on the cross-validated R^2, which were used to estimate the performance of the different models. Hyperparameters are basically the parameters which are related to the training or learning of the algorithm (learning rate, regularization constant, number of branches in a decision tree, etc.). The optimized model has been applied on testing data to calculate the predicted R^2. The process was repeated 100 times as recommended by several studies (Deist *et al*., 2018; Cai *et al*., 2019; Wilson *et al*., 2019) and we got 100 values of predicted R^2 per method. As done by Cai *et al*., 2019, we average the 100 predicted R^2 values to present a single average value to evaluate the performance of the prediction, named predicted R^2 in the rest of the paper.

2.4.1 | Least absolute shrinkage and selection operator

Least absolute shrinkage and selection operator (LASSO) is a type of linear regression in which the selection and regulation of variables take place simultaneously. It was developed by Robert Tibshirani (Tibshirani, 1996). It is largely used in the field of agronomy (Cai *et al*., 2019; Abbas *et al*., 2020). The LASSO model minimizes the usual sum of squared errors, with a bound α on the sum of the absolute values of the coefficients. The hyperparameter α assigns a value of exactly 0 to regression coefficients corresponding to nonsignificant or redundant predictors. Therefore, the main objective of LASSO regression is to "obtain the subset of predictors that minimizes prediction error for a quantitative response variable." It presents the advantage of being a simple linear regression model with regularization easy to generalize and powerful to extract few important features from large datasets. However, it presents some limitations. For instance, it needs feature scaling and adjustment of regularization parameter (Pereira *et al*., 2016).

2.4.2 | Random forest

Random forest (RF) are one type of machine learning algorithms. It consists of combining a large number of decision trees for classification or regression. It is also widely used in research, especially for crop prediction (Jeong *et al*., 2016; Abbas *et al*., 2020). It is used here for regression. RF is simply a collection of Decision Trees that have been generated using a random subset of data subset of data. The name "random forest" comes from junction of the randomness that is used to pick the subset of data with having a bunch of decision trees, hence a forest. It includes three hyperparameters, namely the number of trees in the "forest," the number of variables used to divide each node and the maximum depth of the tree. The main objective is to overcome the overfitting problem of the individual decision trees. In fact, the output of random forest is evaluated by taking average value of the prediction of individual trees (Breiman, 2001). Among the advantages of such a method, it also outputs the importance of important features in large datasets and it is less prone to overfitting than decision trees and other algorithms. However, RF algorithms present some limitations. Indeed, it may change considerably by a small change in the data and these computations can be much more complex than those of other algorithms (Pathak and Pathak, 2020).

2.4.3 | Support vector machine

Support vector machine (SVM) is a supervised machine-learning algorithm and can be used for both classification and regression (Vapnik, 1998). SVM uses kernel functions which can be linear or polynomial in order to obtain nonlinear function (Gunn, 1998). Kernel functions are one of the important hyperparameters that need tuning. SVM minimizes the error by adding the hyperplane and maximizing the margin between the prediction and the actual values (Karatzoglou *et al*., 2006). The goal of SVM is to identify an optimal separating hyperplane which maximizes the margin between different classes of the training data. It is among the most effective machine learning methods in various crop modelling studies for its high accuracy (Cai *et al*., 2019; Abbas *et al*., 2020). In fact, this model presents many advantages. It is very effective even with high dimensional data. It also works very well if the classes in the data are well separated points. SVM can also work with image data. Nevertheless, SVM model has some disadvantages. In fact, it is not easy to choose a good kernels function. Moreover, the training time has to be very long for large datasets. It is also difficult to understand and interpret the final model, the weights of the variables and the individual impact and to fine-tune these hyperparameters (Kirchner and Signorino, 2018).

2.4.4 | Neural network

Neural networks (NNs) are composed of simple elements working in parallel. These elements are inspired by biological neural network that constitutes human (Agatonovic-Kustrin and Beresford, 2000). As in nature, the connections between these elements largely determine the network function. A neural network contains layers of interconnected nodes (input nodes layer, hidden nodes layers and output node layers). The nodes known as perceptron are similar to a multiple linear regression. The perceptron feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear. A neural network can be operated by performing a particular function by adjusting the values (parameters) of the connections (weights) between its elements. The hyperparameters that need tuning include the number of neurons in each layers, the number of hidden layers and the transfers function. Like other three machine learning techniques, NN is also widely used in various fields (Awad, 2019). This model has some advantages. In fact, it stores information about the whole network. In addition, the disappearance of some information in one place does not prevent the network from working. It can work with incomplete knowledge. The loss of performance here depends on the amount of missing information. The corruption of one or more cells in the ANN network does not prevent it from generating an output. It can also do parallel processing to perform several tasks at the same time. NN has also many drawbacks. Indeed, it requires processors with parallel processing power, according to their structure. The biggest problem is the unexplained behaviour of the network. When it produces a convincing solution, it does not give a clue as to why and how. This can lead to uncertainties in the network. There is no specific rule for determining the structure of artificial neural networks. The appropriate network structure is obtained through experience and testing. The duration of the network is also unknown. Indeed, the network is reduced to a certain value of the error on the sample, which means that the learning is finished. This value does not give us optimal results (Mijwel, 2018).

2.4.5 | Linear regression

To study the future predictability we use a linear regression which is given by the following equation:

$$Y = aX + b, \tag{4}$$

where Y is the dependent variable which represents here R^2 of yield obtained with one of the four methods NN, SVM, RF or LASSO, X is the independent variable which represents here the summer annual rainfall per department. A and B represent the regressions parameters.

3 | RESULTS

3.1 | Performance of the prediction models

Table 3 shows the predicted R^2 for each crop by using only the selected climate variables for regression models. The results show that the R^2 value is generally greater than 0.5 for most of the regression models except for LASSO when we consider maize and sorghum yield and for NN with maize yield. The results show also that R^2 is higher for peanut yield compared to others. The performance (R^2) of the four methods were computed using three combinations of input data (NDVI only, Climate only and NDVI+Climate) to predict peanut, maize, millet and sorghum yields (Figure 2). Predictions using climate variables largely outperform those using NDVI whatever the crop and the statistical method we used for prediction. Combining NDVI to climate predictors slightly improved the performance of most models but the added value of NDVI remains low. In addition, the prediction obtained with the linear model (i.e., LASSO) is generally lower than the one obtained with the other methods (SVM, NN and RF) except for the millet where it presents a better prediction than the RF. Indeed, as noticed by Cai *et al.* (2019) who found similar results, the relationship between the predictors and the response is not necessarily linear. The NN model gives the better prediction for peanut and sorghum yields (Figure 2a,d). For maize yield (Figure 2b) the best prediction was obtained with RF followed by SVM and for millet yield (Figure 2c) the SVM gives the best performance which is the same with NN when considering the NDVI+Climate combination.

TABLE 3 Coefficient of determination (R^2) between the predicted yields obtained from selected variables and crop yields (peanut, maize, millet and sorghum) during the rainy season (June–September)

Crops	Peanut	Maize	Millet	Sorghum
Number of selected climate predictors	9	12	11	11
R^2 LASSO	0.56	0.41	0.50	0.47
R^2 NN	0.66	0.42	0.52	0.57
R^2 SVM	0.56	0.56	0.53	0.53
R^2 RF	0.65	0.64	0.51	0.55

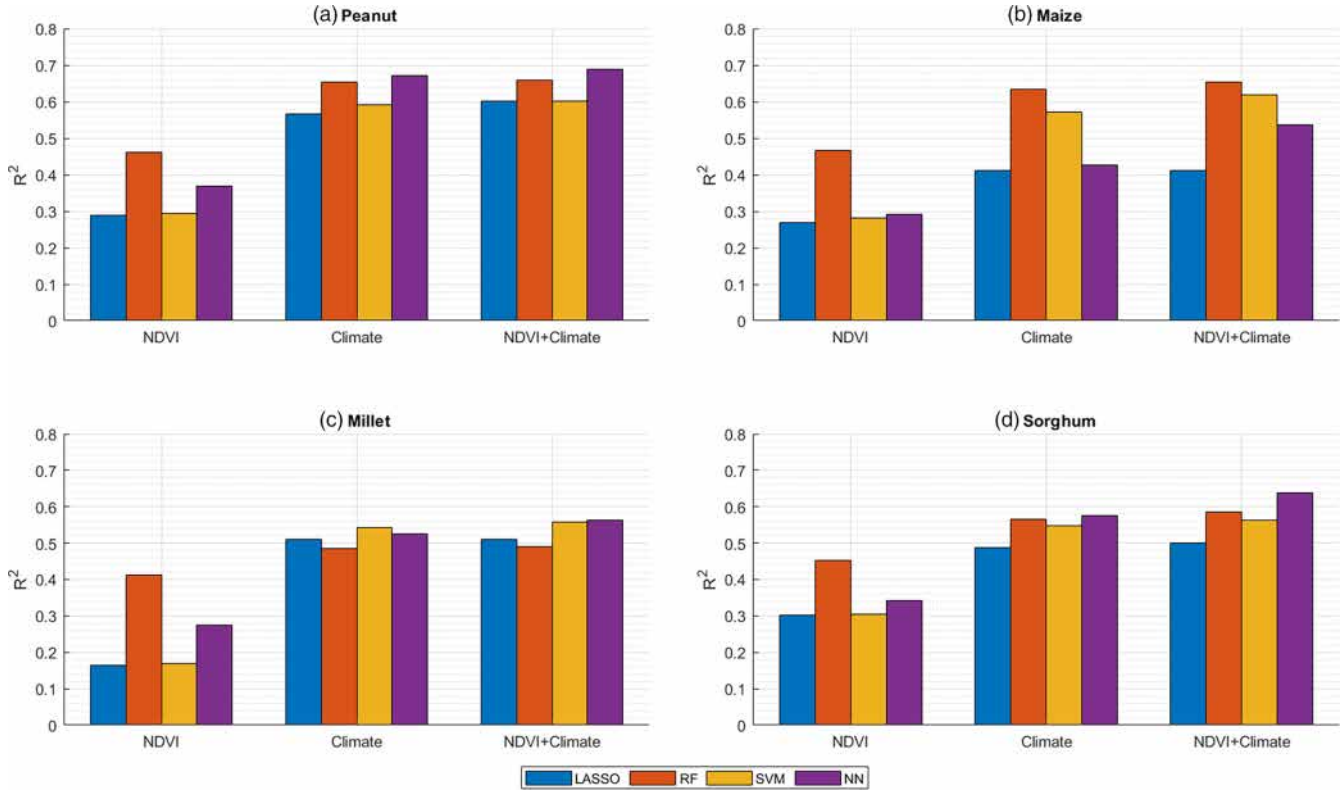

FIGURE 2 The models performance (mean predicted R^2) of four regression models using three combination of inputs (NDVI only, Climate only and NDVI+Climate) averaged over the rainy season (June–September) for peanut (a), maize (b), millet (c) and sorghum (d) [Colour figure can be viewed at wileyonlinelibrary.com]

Although the weight of the different predictors cannot really be interpreted in terms of physical link in machine learning methods, we investigated the relative contribution of the different explaining variables to the yield prediction (Figures S2–S5). We found that the cumulated rainfall is almost always the most important predictive variable regardless of the considered prediction method and crop. Cloud cover percentage and water vapour deficit, which are highly linked to rainfall, are to a lesser extent important predictive variables for several crops/predictive methods. Since those four crops are rainfed, it is not surprising that total rainfall and related variable (cloud cover percentage and water vapour deficit) explain an important part of the observed variance of crop yields.

Producing and disseminating accurate crop yield forecasts with a long lead-time could enable decision-makers to take adjustable, intervention options that can mitigate the severity of various scenarios of food insecurity or get benefits of forecasts of high-yield years. In order to evaluate the value of the statistical models in a forecasting context, we compute the temporal evolution of the skill of the full model (considering both NDVI and Climate variables) from June which corresponds to the start of the rainy season in the southern regions of Senegal to September which corresponds roughly to the cessation of the rains and the start of the harvest (Figure 3). The value of R^2 in June given in Figure 3 corresponds to the model builds with climate indices and NDVI in June. The value of R^2 in July given in Figure 3 corresponds to the model builds from climate indices and NDVI averaged or cumulated from June to July. The value of August is built from climate indices and NDVI averaged or cumulated from June to August. Finally the value of September in Figure 3 corresponds to the performance of the full model given in Figure 2 (NDVI+Climate). If, as expected, the skill is increasing from June to September as we integrate more climate variations during the rainy season, it is worth to notice that this increase is not linear. Indeed, there is generally a strong increase between June and July and a slight increase from July to August. The four crops we studied are rainfed and depend on the unique water supply from the monsoon season which usually starts in June and ends in late September. The crops are sown in early June and a late onset of the rains and/or a water shortage after the onset can have a detrimental effect on crop yields. This dependence of crop production to the rains early in the season could explain the high skill of the predictive models in June. Almost all the models reach saturation in August for the prediction of

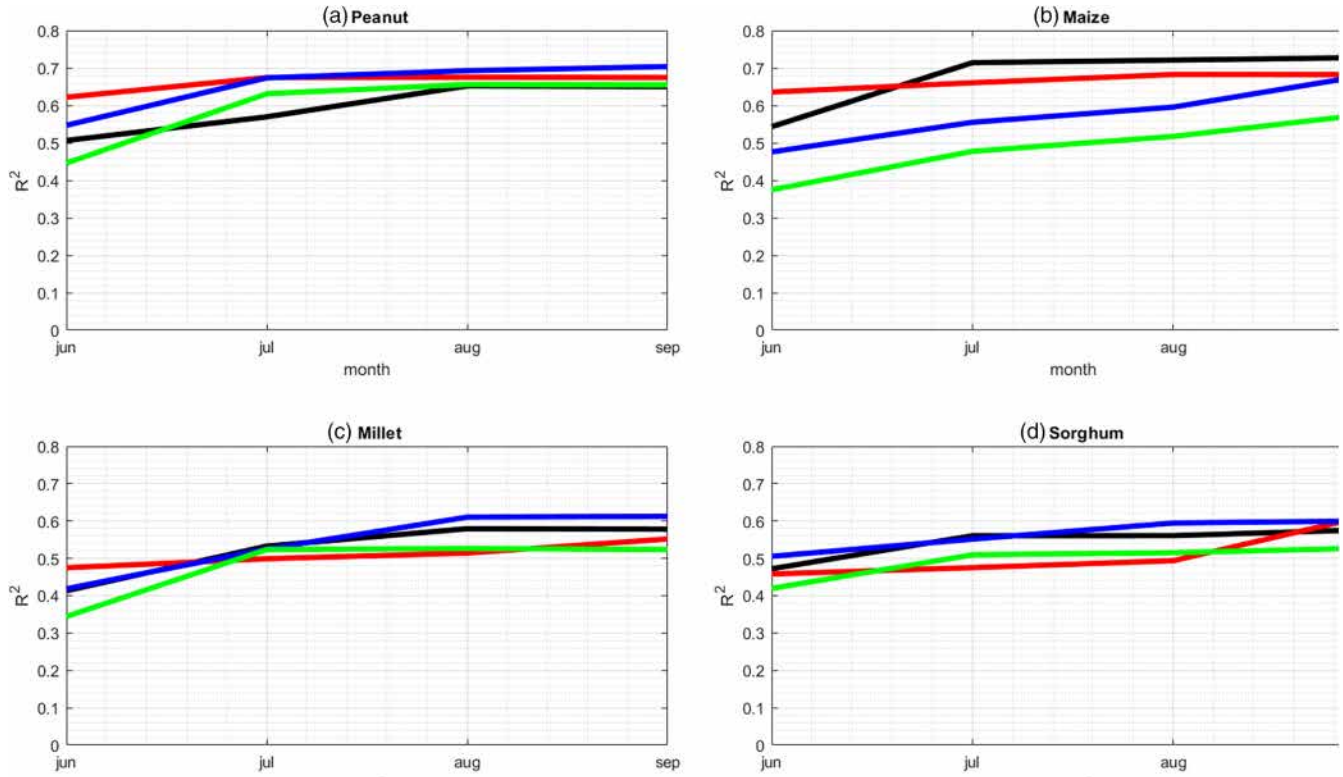

FIGURE 3 Temporal progression of yield prediction obtained with four regression models across the month of the rainy season (June, July, August and September) using NDVI and climate combination as predictors for peanut (a), maize (b), millet (c) and sorghum (d). Each predicted value is computed iteratively using NDVI–Climate of June only (the June R^2 in the figure), then using NDVI–Climate of June–July (the July R^2 in the figure), June–July–August predictors (the August R^2 in the figure) and finally using the NDVI–Climate of June, July, August and September to compute the September R^2 value. As we increase incrementally the inputs, the R^2 values are logically increasing from June to September and are maximal using the NDVI–Climate of June, July, August and September (the September R^2 in the figure) [Colour figure can be viewed at wileyonlinelibrary.com]

peanut, millet and sorghum prediction except for the RF model. It means that the models could provide skilful forecasts of crop yields using climate and satellite information averaged from first June to July 30th. The best predictions are generally obtained with NN in July for groundnut, millet and sorghum. However, the prediction with RF exceeds the one of NN in September for sorghum. For maize, the best predictions are obtained with SVM using climate and satellite information averaged from first June to July 30th. It is also interesting to notice that the RF method provides the best forecasts of peanut, maize and millet yields with early information (climate and NDVI in June). Moreover, the methods are generally more efficient in August or September.

3.2 | Spatiotemporal variability

In order to assess the spatial variability of the predicted yields, we extract the predicted yields for each department and calculate predicted R^2 separately. It reveals that the performance of the statistical models is not equal across Senegal (Figure 4). Indeed, regarding peanut yield prediction (Figure 4a), the highest coefficients of determination are found in the central west of the country. Concerning maize yield predictions (Figure 4b), the highest skills are obtained in the north of the peanut basin (Kébémer) and in the south of the basin (towards Nioro) and in the southwest of Senegal (towards Oussouye, Sédhiou and Kolda). Regarding millet yield predictions (Figure 4c), the performance is generally better in Kébémer and Sédhiou. As for sorghum yield (Figure 4d), the best predictions are recorded in Gossas, Mbacké, Kaffrine and Mbour.

Several reasons can explain the spatial variability of the performance of the model. First, the location of the cropped areas indicates areas where the crops are mainly grown and where the production is likely to be more intensive (Figure S6). In such a region, there are usually more capital, labour and inputs such as fertilizers, insecticides, pesticides and weedicides which results in more yield of the crop per hectare and less losses due to

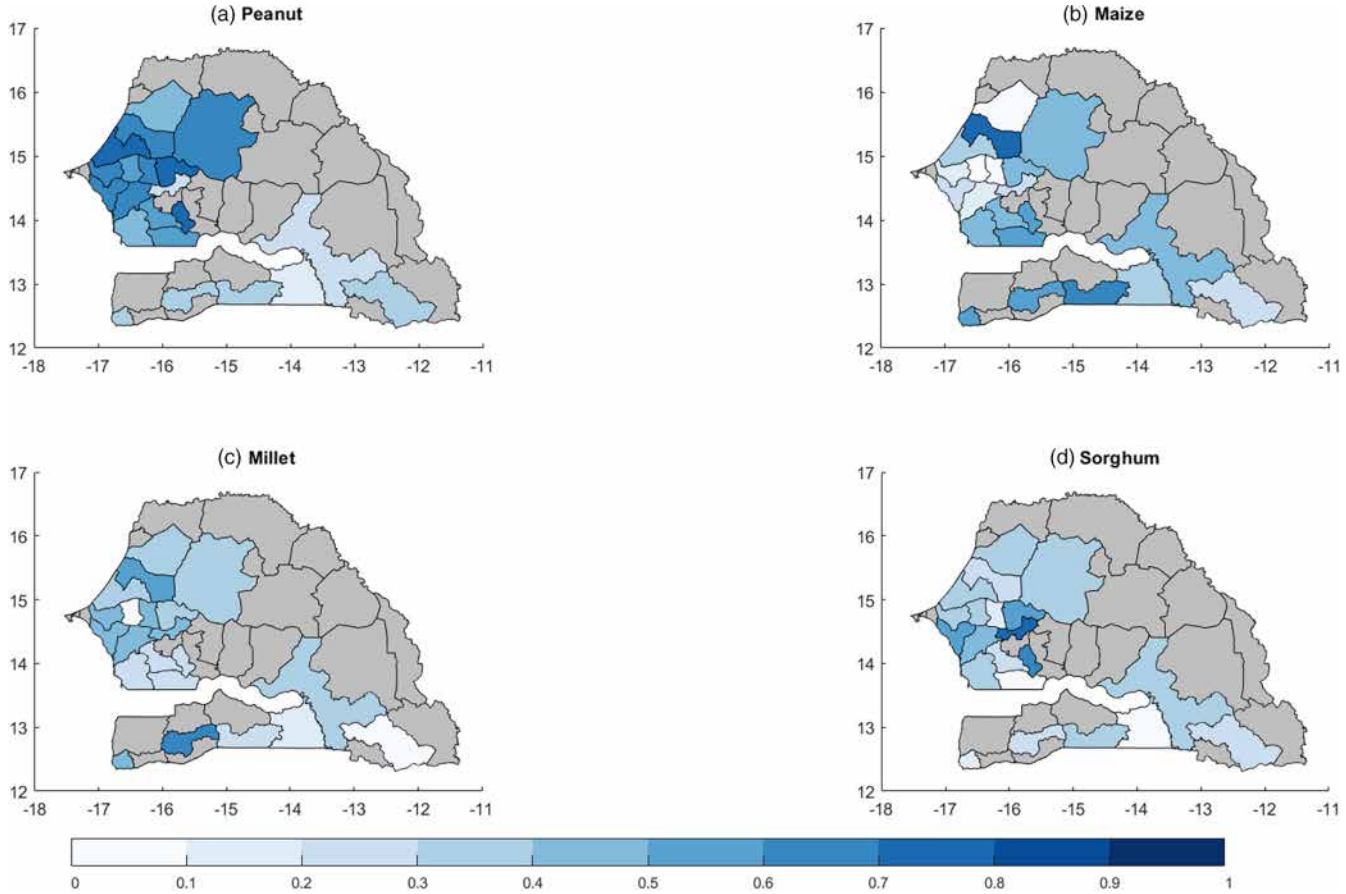

FIGURE 4 Spatial distribution of the skill of the regression models using NDVI+Climate combination from June to September for peanut (a), maize (b), millet (c) and sorghum (d). We fit/test the four models (NN, LASSO, RF and SVM) using data from all departments and years together. We thus have four mean predictions for each department and each year. We then extract the predictions for one department, calculate R^2 separately for each of the four methods and average the four R^2 values to produce the figure [Colour figure can be viewed at wileyonlinelibrary.com]

management issues. We thus expect that climate explain more variations of crop yields in these more intensive regions. It is the case for peanut where the highest coefficients of determination are found in the areas where peanut is mainly grown (Figure S6). Another reason that could be given is the size of the cropped areas. Land holding in more intensive cropping systems is generally smaller, expensive and in more densely populated region than in extensive cropping systems. Small cropped surface could be a marker of more intensive cropping areas with a higher relationship with climate variability. This hypothesis reveals to be true for peanut and millet where there is a linear negative relationship between the cropped surface and the prediction skill with the highest skills associated to the lowest surfaces (Figure S7).

In addition to cropped areas, another element which could explain the spatial distribution of the predictive score of the different methods is the spatial pattern of mean rainfall. Indeed, if crop yields in Senegal depend crucially on summer rainfall, the relationship between those two variables is not linear (Figure S8). Crop yields variability is clearly driven by the rainy season when the cumulative rain is less than roughly 800 mm per rainy season with a positive linear relationship showing that more or less rainfall could lead to respectively more or less yields.

When seasonal rainfall amount is above this threshold, Figure S8 shows a slightly decreasing plateau where cumulative rainfall becomes less important to explain crop yield variability. It depicts wet situations where water stress is less frequent and where other factors can limit crop yields such as radiation limitations (Baron *et al.*, 2005), farmers' practices or the occurrence of pest and diseases. As cumulative rainfall is one of the most important predictors in the four statistical models, it is likely that we could expect higher predictive skills in departments with less than 800 mm of rainfall per summer (Gossas, Tambacounda, Foundiougne, Bambey, Diourbel, Louga, Fatick, Kaffrine, Mbour, Thiès, Kébémer, Linguère, Kaolack, Nioro, Tivaouane, Mbacké and Bakel).

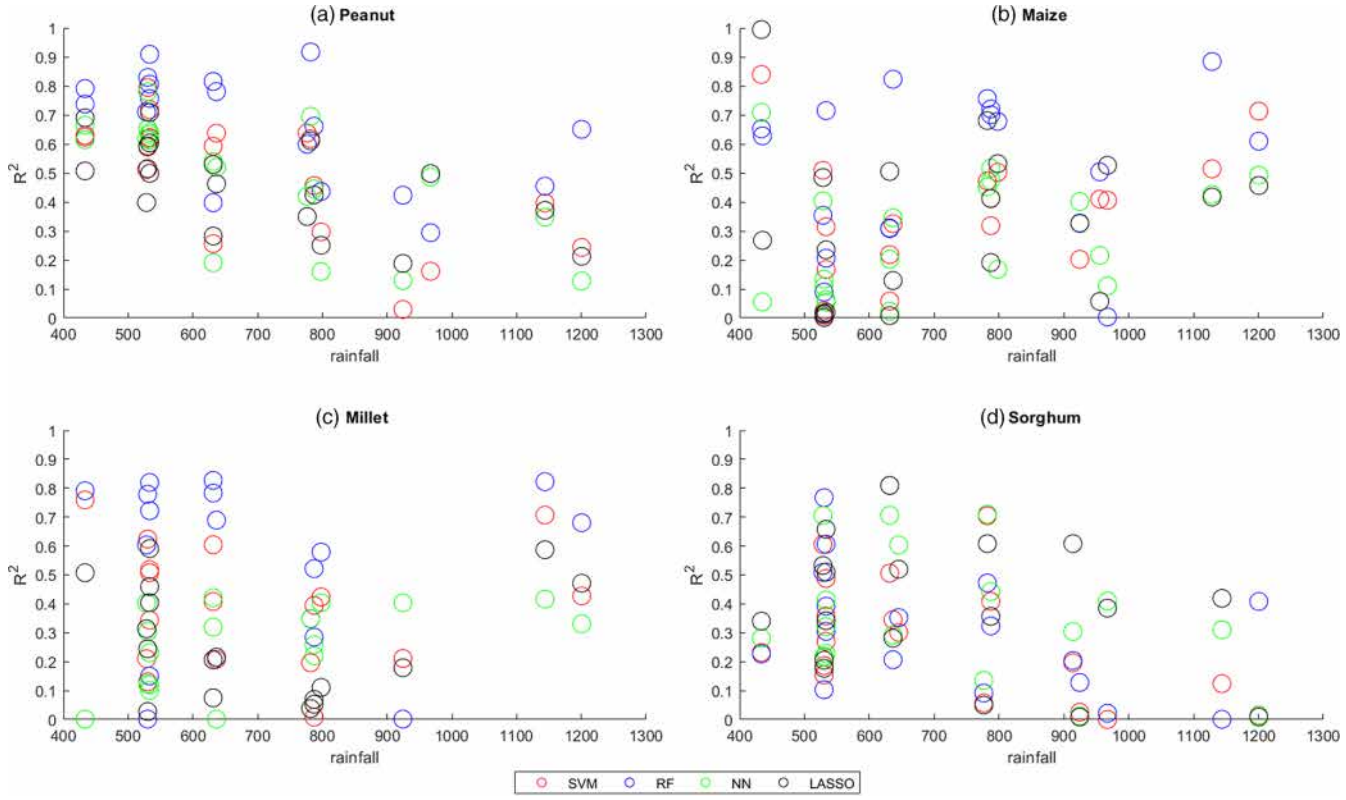

FIGURE 5 Predicted R^2 of the regression models at the department level (one colour per model) and June to September accumulated rainfall (mm·season^{-1}). As for Figure 4, the statistical models use NDVI+Climate combination from June to September to predict yield of peanut (a), maize (b), millet (c) and sorghum (d). We fit/test the four models (NN, LASSO, RF and SVM) using data from all departments and years together. We thus have four mean predictions for each department and each year. We then extract the predictions for one department and calculate R^2 separately for each of the four methods [Colour figure can be viewed at wileyonlinelibrary.com]

The predicted values versus annual rainfall at the departments scale are shown in Figure 5. Regarding peanut predictions, as expected, the highest prediction skills are observed in the drier departments and there is a decreasing trend of the forecast accuracy as the departments get wetter. Figure 5b–d does not show such a clear trend for maize, millet and sorghum even if we could suspect this negative relationship between prediction accuracy and annual rainfall for sorghum predictions. It is likely that those crops are more sensitive to intra-seasonal variations of rainfall rather than cumulative rainfall (Marteau *et al.*, 2011; Guan *et al.*, 2015).

Figures 6 and S9–S11 show the cumulative summer rainfall and the performance of the regression models using different combinations of input data (Climate only, NDVI only and Climate+NDVI) to predict peanut, maize, millet and sorghum yields respectively for each department. Generally, the best predictions are obtained when we have more input data (i.e., Climate+NDVI). If we consider the different inputs data taken individually, the results show that the best performance is obtained with climate data in most departments except for NN with maize, millet and sorghum crops (Figures S9–S11) where NDVI gives on average the best predictions. Regarding peanut prediction (Figure 6), the R^2 from the two combinations Climate and NDVI+Climate are the highest (lowest) in the departments where cumulative summer rainfall is the lowest (highest), especially for the NN and SVM models. However, for the other crops, we do not see such a trend, except for the prediction from the combination NDVI+Climate using NN in the case of millet.

Finally, we show the time series of observed and mean predicted yield from the 100 simulations for peanut, maize, sorghum and millet using the NN method (Figure 7) in average over Senegal. Even if the sample is limited to 14 years, it is clear from Figure 7 than the observed and predicted time series of crop yield of peanut are very close. In particular the model is able to simulate yield losses due to abnormally dry years in 2002, 2007 and 2013 but also high yields during wet years in 2005, 2008, 2009 and 2010. Although maize, millet and sorghum observations indicate a yield drop during the dry years 2002 and 2007 which is well predicted by the NN method, the agreement between the observed and predicted time series is less satisfying for these three crops. It is also clear from Figure 7 than observed yield of maize,

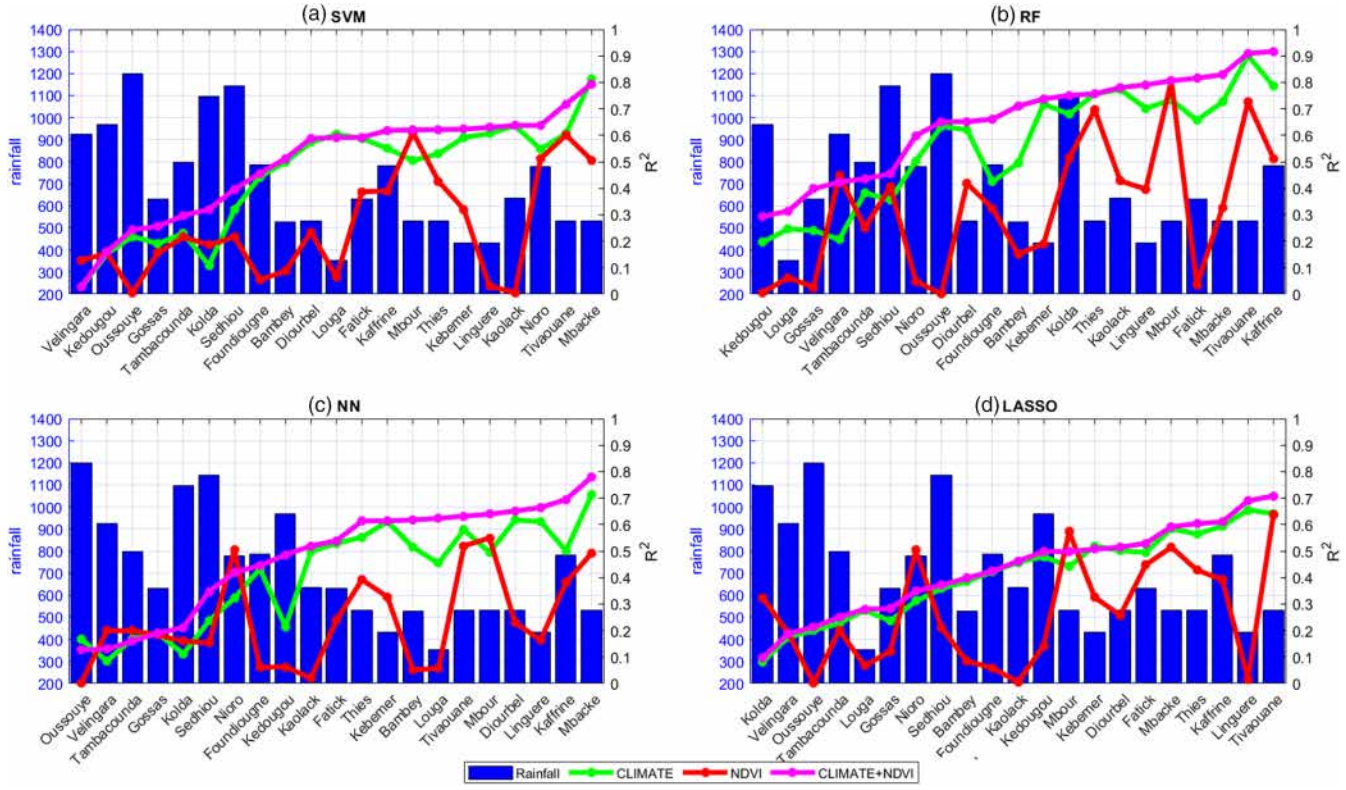

FIGURE 6 Predicted R^2 of the regression models at the department level to predict peanut yields using June to September climate (green), NDVI (red) and Climate+NDVI combination (purple) as inputs. Blue bars represent June to September accumulated rainfall (mm·season^{-1}) [Colour figure can be viewed at wileyonlinelibrary.com]

millet and sorghum are less correlated with annual rainfall, being more sensitive to other climatic (intraseasonal distribution of rainfall, seasonality of the rains, radiation...) or nonclimatic factors (management, pest and diseases...). It could explain the relatively lower performances of the prediction model. Similar results can be found using the three other approaches SVM, LASSO and RF (Figures S12–S14).

4 | SUMMARY AND DISCUSSIONS

Our study investigates the predictability of crop yields in Senegal using four machine learning algorithms, namely three nonlinear models (random forest, singular vector machine and neural network) and the linear regression model LASSO to predict yields of peanut, maize, millet and sorghum. Either observed climate variables and/or vegetation data estimated from satellite retrievals were used as explaining variables to predict crop yields.

Our results showed that the combination of climate and vegetation with the nonlinear models (RF, SVM and NN) gives the best performance for crop yield prediction even if the contribution of satellite vegetation data remains low compared to climate data. These results are consistent with the ones of Cai *et al.* (2019) and Leroux *et al.* (2020). Both studies showed that nonlinear models outperform linear methods (LASSO regression in Cai *et al.*, 2019; multiple linear regression in Leroux *et al.*, 2020) and found that vegetation indices from satellite data explain less crop variability compared to climate data. However, it does not call into question the use of satellite data for crop prediction. Indeed, promising results for maize yield estimation in West Africa are found using satellite indices of surface soil moisture and temperature of canopy (Leroux *et al.*, 2020). The integration of such satellite retrievals of climate conditions experienced by the plant (drought and heat) could be a good way to improve the performance of the prediction models we developed in this study. Furthermore, it is important to highlight that the approach is still experimental since predictor selection and model test are performed on a very short period of data (14 years) with the use of cross-validation which could lead to overestimate the performance of the prediction methods. However, using a similar experimental framework, Cai *et al.* (2019) provided strong evidences of the practical performance of such validation to assess the out-of-sample predictions. Another important limitation induced by the small sample of the yield dataset is that we combined yield values from the

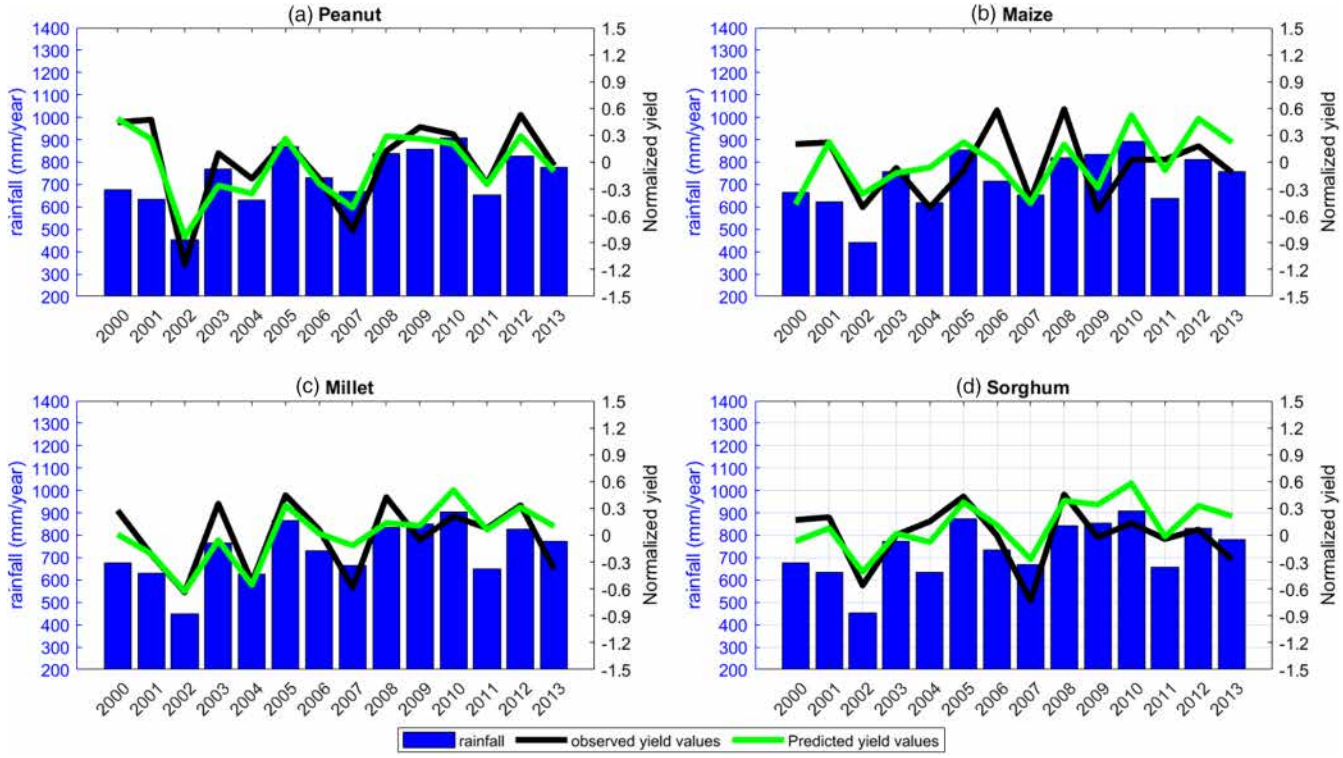

FIGURE 7 Observed (black) and predicted (green) yield using NN method (standardized anomalies) in average across Senegal. We fit/test the NN model using data from all departments and years together. We thus have mean predictions for each department and each year. We then averaged the predictions of the 24 departments to get predicted values in Senegal. Blue bars represent June–September accumulated rainfall (mm·season^{-1}). Similar comparisons between observed and predicted values have been done for the three other machine learnings methods (Figures S12–S14) [Colour figure can be viewed at wileyonlinelibrary.com]

14 years and the 24 departments to get a larger dataset to train and validate the machine learning methods. It implicitly assumes that the same model applies for each department and may account for some of the spatial variation in skill.

The results also showed that the best yield predictions are obtained between 1 and 2 months before the end of the rainy season which is also the harvest time. Such information before harvest could be a valuable information to be included into an early warning system linked to early action in case of harvest shortfalls. However, this lead time might be too short to support decision making of farmers who need to make critical climate-sensitive decisions months before the rainy season starts (Roudier *et al.*, 2012). A skilful prediction for longer lead-time could be reached by applying the same kind of statistical models developed in this study but using numerical climate forecasts instead of climate observation. Even if the use of such climate forecasts is still challenging for specific sectors or regions (Doblas-Reyes *et al.*, 2013), there are promising examples of crop yield predictions based on such climate forecasts (see for instance Iizumi *et al.*, 2021). However, the use of such forecasts is not straightforward for crop yield prediction. In our case, it would require first to compare the climate observations and climate hindcasts at different lead time to select the best forecasted climate variables. These variables could then be used to train a new statistical model as done in this study to predict crop yields.

The best prediction skill is obtained for peanut with a coefficient of determination up to 0.66 when comparing observed and predicted yields. Indeed, peanut shows the highest correlation between yields and summer rainfall and several studies reported the high sensitivity of this crop to interannual climate variability not only in Africa (Prasad *et al.*, 2010; Hamidou *et al.*, 2013) but also in India (Bhatia *et al.*, 2009; Challinor *et al.*, 2009; Singh *et al.*, 2012). The performance of statistical models to predict peanut yields has important societal implications in Senegal since peanut is an important oil seed and food crop grown by small-holder and resource-poor farmers, providing the main source of income in rural areas (Tarawali and Quee, 2014; Faye *et al.*, 2018a). Together with Nigeria, Senegal is one of the largest producers in the West African region and a skilful crop yield prediction system could have an important economic value.

The increase of anthropic greenhouse gases emissions affects global surface temperatures but has also an effect

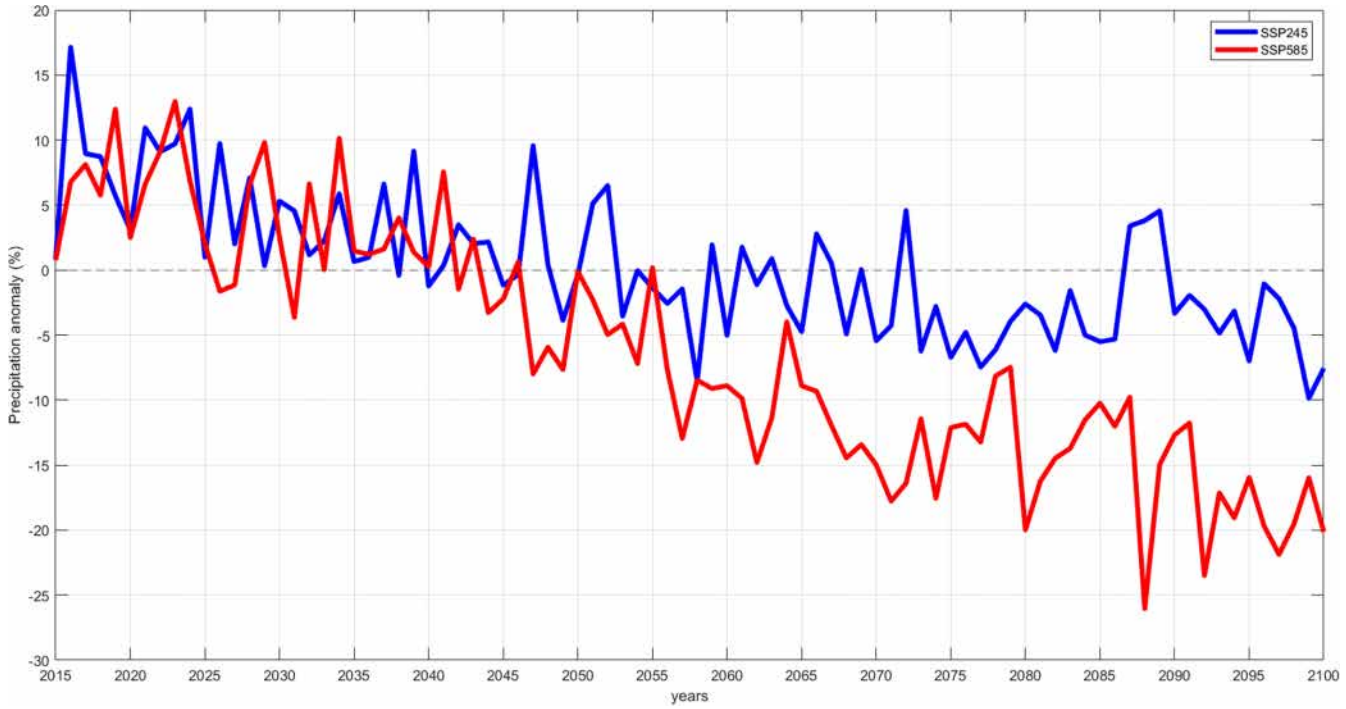

FIGURE 8 Temporal evolution of the mean JJAS (June–July–August–September) rainfall relatively to baseline 2000–2013 period (%) as simulated by the ensemble mean of CMIP6 model bias-corrected under the SSP2-45 and SSP5-85 scenarios [Colour figure can be viewed at wileyonlinelibrary.com]

on the hydrological cycle in West Africa (Sultan *et al.*, 2014) with regions experiencing more rainfall under future climate scenarios such as Central Sahel and less rain in the Western part of the Sahel (Senegal, Southwest Mali). Since rainfall is expected to change under global warming scenarios, one can expect changes in crop yields but also in crop yields predictability. The rainfall projections in Senegal from the latest CMIP6 exercise show a clear downward trend in June to September rainfall amount across time (Figure 8). Although this rainfall reduction occurs in both low- and high-emission scenarios (SSP2-45 and SSP5-85, respectively), the amplitude of the rainfall deficit is more pronounced under the SSP5-85 with rainfall anomalies of −20% by the end of the century. This trend was also depicted in the previous CMIP5 simulations (see for instance Sylla *et al.*, 2016; Diallo *et al.*, 2016 who analysed rainfall trends in the Sahel). Under the scenario SSP2-45, the multimodel mean simulates a slight increase in rainfall during near future (Figure 9a) in Bakel, but also in the centre and north of the peanut basin (central-western part of the country) except for the departments of Kaffrine and Nioro. Rainfall is expected to decrease in the south of the country except over the lower Casamance (Ziguinchor, Bignona and Oussouye). The decrease is much more significant with the SSP5-85 scenario (Figure 9b) with a rainfall deficit in all 24 departments. The greatest decrease is found in Bambey, Kaffrine, Nioro and Tambacounda. The situation is aggravated by 2,100 horizon (Figure 8c,d for SSP2-45 and SSP5-85, respectively). Indeed, under both scenarios, the average of the CMIP6 models simulates strong decreases in rainfall that could exceed 20% in Tambacounda, Vélingara, Kaffrine, Louga and Nioro. Although the simulations are not the same, this effect of anthropic greenhouse gases on rainfall in the Sahel is consistent with the results from Giorgi *et al.*, 2014; Mariotti *et al.*, 2014; Diallo *et al.*, 2016; Sarr and Camara, 2017.

This rainfall decrease could lead to yield losses, especially in departments where rainfall is limiting agricultural yields. Such detrimental effects on crop yields have been largely described in the literature (see for instance Sultan *et al.*, 2014; Faye *et al.*, 2018b). On the other hand, a reduction in seasonal rainfall could modify the predictability of crop yields as the performance of yield predictions depends on summer rainfall amount for some regions and some crops (Figure 5a). To have a rough estimate of these changes in predictability, we build on the results from Figure 5 to develop a simple linear model based on the relationship between the coefficient of determination of our prediction models and cumulated summer rainfall. Table 4 presents the correlation between the coefficient determination of each regression models based on climate and NDVI indices and cumulated seasonal rainfall at the departmental scale over the

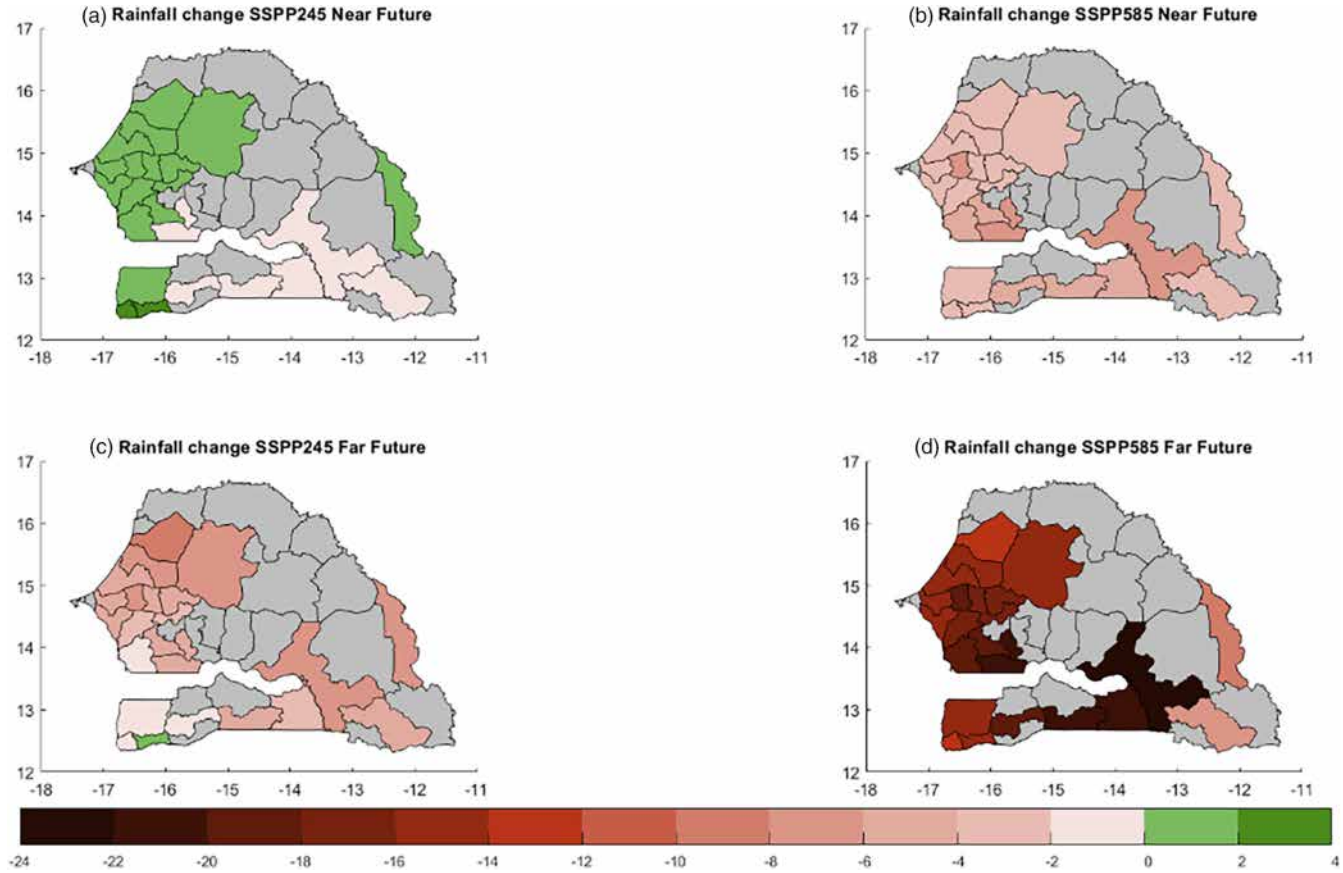

FIGURE 9 Change future (near future and far future) minus present (2000–2013) during JJAS period in % simulated by ensemble mean of CMIP6 models bias-corrected under the SSP2-45 and SSP5-85 scenarios [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Correlation between the predicted coefficient determination (R^2) by each regression model and rainfall during rainy season (June–September) from 2000 to 2013 at the departmental level

Models	LASSO	NN	RF	SVM
Peanut	**−0.57**	**−0.72**	−0.28	**−0.69**
Maize	0.19	0.24	0.34	0.12
Millet	−0.10	**0.45**	0.13	−0.25
Sorghum	−0.27	−0.33	−0.16	−0.38

Note: Significant values at 5% are highlighted in bold.

baseline period 2000–2013. As highlighted by Figure 5 there is a negative correlation between the coefficient of determination of most statistical prediction models of peanut (LASSO, NN and SVM) and annual rainfall which means that the models perform better in departments where rainfall is lower. However, such relationships are not significant for other crops although similar negative correlations are found for sorghum predictions. We will thus focus on peanut to roughly estimate future predictability by using a simple linear regression model (see Equation (4)) where the dependent variable to predict is R^2 of peanut yield obtained with one of the four methods NN, SVM, RF or LASSO and the independent explaining variable represents the summer annual rainfall per department. Once the regression parameters computed using the baseline rainfall from 2000 to 2013, we estimate future predictability by rescaling the baseline rainfall by future climate anomalies (see section 2) and replacing the baseline rainfall by this new future rainfall. From this estimated predictability in the future, we can compute the relative change in predictability by department between the future (near and far future) and the baseline period (Figure 10). Globally, our results show that the strongest changes are obtained with the regression models characterized by the strongest correlation between R^2 and rainfall during the reference period (i.e., NN model, followed by SVM and LASSO). In the near future, changes are relatively small, especially in the medium scenario (SSP2-45) (Figure 10a), which shows both negative and positive changes, with a relatively large decrease in the department of Oussouye

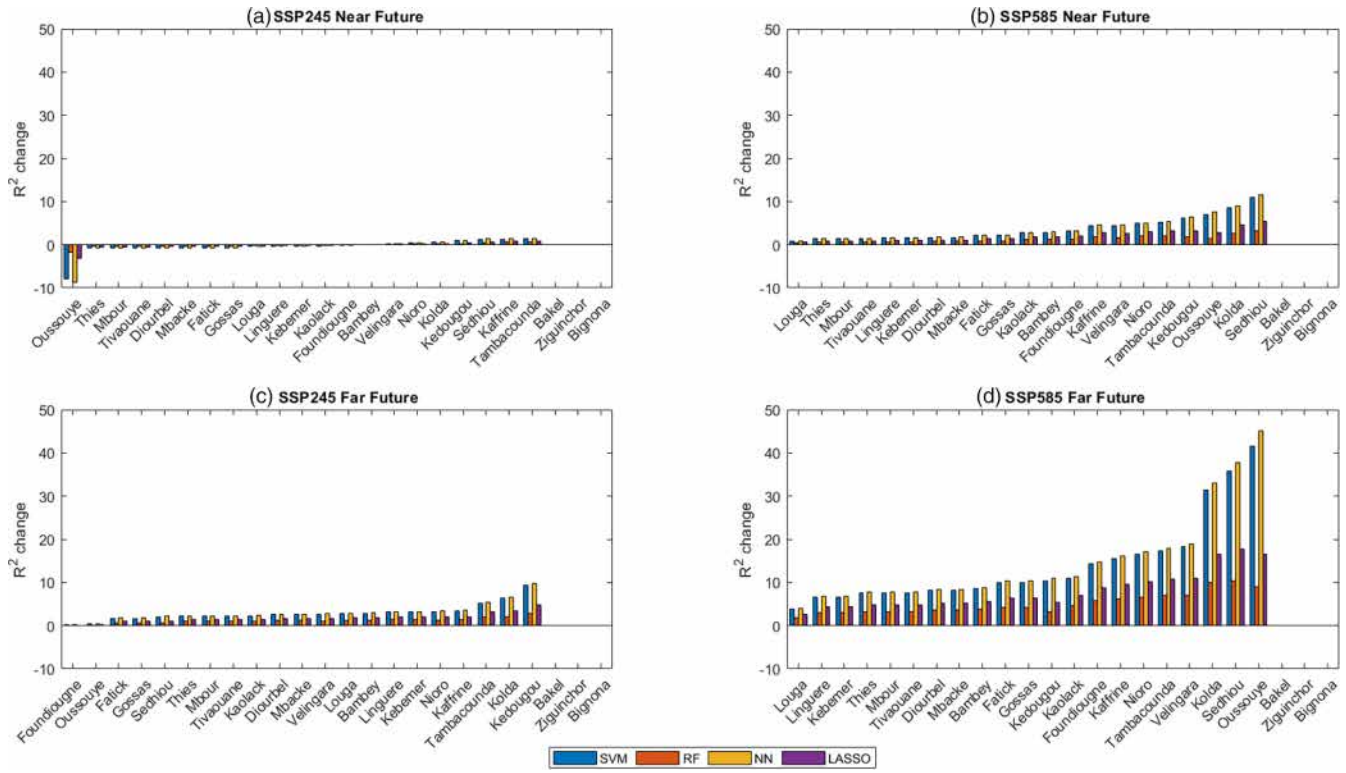

FIGURE 10 Predictability change during rainy season (JJAS) between the future (near future and far future) with respect to the baseline (2000–2013) for each department under the different scenarios (SSP2-45 and SSP5-85) [Colour figure can be viewed at wileyonlinelibrary.com]

and an upward trend in the models for the departments of Tambacounda, Kaffrine, Kédougou, Kolda, Nioro and Vélingara. Under the high-emission scenario (SSP5-85) (Figure 10b), the prediction performance could experience a small increase compared to the SSP2-45 scenario. This increase could reach 10% in the departments of Tambacounda, Kédougou, Oussouye, Kolda and Sédhiou depending on the prediction models. During the far future (Figure 10c for SSP2-45 scenario and Figure 10d for SSP5-85 scenario), an increase in predictability is expected regardless of the type of scenario considered. This increase remains relatively low in the medium scenario and does not exceed 10%. The largest increases are recorded in the department of Kédougou, followed by Bakel and Tambacounda, and the smallest (<1%) in the departments of Foundiougne and Oussouye. A large increase could be obtained if we consider the SSP5-85 scenario (Figure 10d). This could exceed 40% in the department of Oussouye and 30% in the departments of Sédhiou and Kolda with the NN and SVM models. This estimation of the changes of predictability under climate change scenarios has some limitations like most modelling studies. Among the most important ones, it only considers the changes in mean rainfall as it is likely that the frequency and/or the seasonality of rainfall could also change in the future and affect the crops (Guan *et al.*, 2015). The expected increase in mean temperature can also shorten crop cycle length, modify evapotranspiration and change the relationship between crop yield and summer rainfall as shown in several studies (see for instance Sultan *et al.*, 2013; 2014). An increased of mean temperature would thus likely corroborate the conclusions of our study on more predictable yields under future climate. Furthermore, even without considering temperature changes, it remains very likely that since rainfall is expected to decrease, water stress will increase, strengthening the relationship between crop and rainfall.

The use of crop yield forecasts will be more and more relevant to compensate yield losses anticipated under climate change scenarios. The need of such early forecasts to support adaptation to climate change is well acknowledged by Least Developed Countries and Small Island Developing States who were nearly 90% to identify early warning systems as a top priority in their Nationally Determined Contributions on climate change (World Meteorological Organisation, 2020).

5 | CONCLUSION

This study evaluated the potential of climate and vegetation indices to predict crop yields of peanut, maize, millet and sorghum in Senegal. The comparison of multiple statistical methods (RF, SVM, NN and LASSO) and combination of predictands show promising prediction skills 1–2 months before harvest. Overall, predictions of peanut yields using nonlinear statistical methods give the best results compared to the three other crops under current climate and the performances may even increase for peanut yields predictions under climate change scenarios. Although experimental yet, this approach could be extended to be included into an early warning system linked to early action in case of harvest shortfalls. Although further work is needed to implement such predictions into an operational forecast system, we believe that such prediction and their integration into operational early warning systems could increase resilience of Senegal to climate change and contribute to food security.

AUTHOR CONTRIBUTIONS

Alioune Badara Sarr: Conceptualization; methodology; resources; software; writing – original draft; writing – review and editing. **Benjamin Sultan:** Conceptualization; methodology; supervision; writing – original draft; writing – review and editing.

ACKNOWLEDGEMENTS

The authors would like to thank the IRD-ESPACE-DEV Laboratory and the MOPGA (Make Our Planet Great Again) project for their support. The authors appreciate the help of Moctar Camara, Samo Diatta, Bamol Ali Sow, Mamadou Lamine Mbaye of Physics Department/UASZ.

ORCID

Alioune Badara Sarr https://orcid.org/0000-0002-7196-3918

REFERENCES

Abbas, F., Afzaal, H., Farooque, A.A. and Tang, S. (2020) Crop yield prediction through proximal sensing and machine learning algorithms. *Agronomy*, 2020(10), 1046. https://doi.org/10.3390/agronomy10071046.

Agatonovic-Kustrin, S. and Beresford, R. (2000) Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, 22(5), 717–727. https://doi.org/10.1016/s0731-7085(99)00272-1.

Awad, M.M. (2019) Toward precision in crop yield estimation using remote sensing and optimization techniques. *Agriculture*, 9(3), 54. https://doi.org/10.3390/agriculture9030054.

Baron, C., Sultan, B., Balme, M., Sarr, B., Traoré, S., Lebel, T., Janicot, S. and Dingkuhn, M. (2005) From GCM grid cell to agricultural plot: scale issues affecting modelling of climate impact. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 2095–2108.

Bennett, D.A. (2001) How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25(5), 464–469.

Bhatia, V.S., Singh, P., Kesava Rao, A.V.R., Srinivas, K. and Wani, S.P. (2009) Analysis of water non-limiting and water limitingyields and yield gaps of groundnut in India using CROPGRO-peanut model. *Journal of Agronomy and Crop Science*, 195, 455–463.

Bi, D., Dix, M.R., Marsland, S.J., O'Farrell, S., Rashid, H.A., Uotila, P., Hirst, A.C., Kowalczyk, E.A., Golebiewski, M., Sullivan, A., Yan, H., Hannah, N., Franklin, C., Sun, Z., Vohralik, P.F., Watterson, I.G., Zhou, X., Fiedler, R.A., Collier, M.A., Ma, Y., Noonan, J.A., Stevens, L., Uhe, P., Zhu, H.Q., Griffies, S.M., Hill, R., Harris, C. and Puri, K. (2013) The ACCESS coupled model: description, control climate and evaluation. *Australian Meteorological and Oceanographic Journal*, 63, 41–64.

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L. and Peng, B. (2019) Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*, 274, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010.

Challinor, A., Wheeler, T., Hemming, D. and Upadhyaya, H. (2009) Ensemble yield simulations: crop and climate uncertainties, sensitivity to temperature and genotypic adaptation to climate change. *Climate Research*, 38, 117–127.

Chang, J., Hansen, M.C., Pittman, K., Carroll, M. and Di Miceli, C. (2007) Corn and soybean mapping in the United States using Modis time-series data sets. *Agronomy Journal*, 99(6), 1654–1664.

Crane-Droesch, A. (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, 13, 114003.

Deist, T.M., Dankers, F.J.W.M., Valdes, G., Wijsman, R., Hsu, I.C., Oberije, C., Lustberg, T., Soest, J., Hoebers, F., Jochems, A., el Naqa, I., Wee, L., Morin, O., Raleigh, D.R., Bots, W., Kaanders, J.H., Belderbos, J., Kwint, M., Solberg, T., Monshouwer, R., Bussink, J., Dekker, A. and Lambin, P. (2018) Machine learning algorithms for outcome prediction in (chemo) radiotherapy: an empirical comparison of classifiers. *Medical Physics*, 45, 3449–3459. https://doi.org/10.1002/mp.12967.

Di Paola, A., Valentini, R. and Santini, M. (2016) An overview of available crop growth and yield models for studies and assessments in agriculture. *Journal of the Science of Food and Agriculture*, 96, 709–714. https://doi.org/10.1002/jsfa.7359.

Diallo, I., Giorgi, F., Deme, A., Tall, M., Mariotti, L. and Gaye, A.T. (2016) Projected changes of summer monsoon extremes and hydroclimatic regimes over West Africa for the twenty-first century. *Climate Dynamics*, 47, 3931–3954. https://doi.org/10.1007/s00382-016-3052-4.

Doblas-Reyes, F.J., Andreu-Burillo, I., Chikamoto, Y., Garcia-Serrano, J., Guemas, V., Kimoto, M., Mochizuki, T., Rodrigues, L.R.L. and van Oldenborgh, G.J. (2013) Initialized

near-term regional climate change prediction. *Nature Communications*, 4, 1715. https://doi.org/10.1038/ncomms2704.

Dufresne, J.L., Foujols, M.A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., de Noblet, N., Duvel, J.P., Ethé, C., Fairhead, L., Fichefet, T., Flavoni, S., Friedlingstein, P., Grandpeix, J.Y., Guez, L., Guilyardi, E., Hauglustaine, D., Hourdin, F., Idelkadi, A., Ghattas, J., Joussaume, S., Kageyama, M., Krinner, G., Labetoulle, S., Lahellec, A., Lefebvre, M.P., Lefevre, F., Levy, C., Li, Z.X., Lloyd, J., Lott, F., Madec, G., Mancip, M., Marchand, M., Masson, S., Meurdesoif, Y., Mignot, J., Musat, I., Parouty, S., Polcher, J., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Talandier, C., Terray, P., Viovy, N. and Vuichard, N. (2013) Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. *Climate Dynamics*, 40(9–10), 2123–2165.

Dunne, J.P., Horowitz, L.W., Adcroft, A.J., Ginoux, P., Held, I.M., John, J.G., Krasting, J.P., Malyshev, S., Naik, V., Paulot, F., Shevliakova, E., Stock, C.A., Zadeh, N., Balaji, V., Blanton, C., Dunne, K.A., Dupuis, C., Durachta, J., Dussin, R., Gauthier, P. P.G., Griffies, S.M., Guo, H., Hallberg, R.W., Harrison, M., He, J., Hurlin, W., McHugh, C., Menzel, R., Milly, P.C.D., Nikonov, S., Paynter, D.J., Ploshay, J., Radhakrishnan, A., Rand, K., Reichl, B.G., Robinson, T., Schwarzkopf, D.M., Sentman, L.T., Underwood, S., Vahlenkamp, H., Winton, M., Wittenberg, A.T., Wyman, B., Zeng, Y. and Zhao, M. (2020) The GFDL Earth System Model version 4.1 (GFDL-ESM 4.1): overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems*, 12, e2019MS002015. https://doi.org/10.1029/2019MS002015.

Egbebiyi, T.S., Lennard, C., Crespo, O., Mukwenha, P., Lawal, S. and Ouagraine, K. (2019) Assesing future spatio-temporal changes in crop suitability and planting season over West Africa: using the concept of crop-climate departure. *Climate*, 7, 102. https://doi.org/10.3390/cli7090102.

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. and Taylor, K.E. (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9, 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016.

Famien, A.M., Janicot, S., Ochou, A.D., Vrac, M., Defrance, D., Sultan, B. and Noël, T. (2018) A bias-corrected CMIP5 dataset for Africa using the CDF-t method—a contribution to agricultural impact studies. *Earth System Dynamics*, 9, 313–338. https://doi.org/10.5194/esd-9-313-2018.

Faye, B., Webber, H., Diop, M., Mbaye, M.L., Owusu-Sekyere, J.D., Naab, J.B. and Gaiser, T. (2018a) Potential impact of climate change on peanut yield in Senegal, West Africa. *Field Crops Research*, 219, 148–159. https://doi.org/10.1016/j.fcr.2018.01.034.

Faye, B., Webber, H., Naab, J.B., MacCarthy, D.S., Adam, M., Ewert, F., Lamers, J.P.A., Schleussner, C.F., Ruane, A., Gessner, U., Hoogenboom, G., Boote, K., Shelia, V., Saeed, F., Wisser, D., Hadir, S., Laux, P. and Gaiser, T. (2018b) Impacts of 1.5 versus 2.0 °C on cereal yields in the West African Sudan Savanna. *Environmental Research Letters*, 13, 034014. https://doi.org/10.1088/1748-9326/aaab40.

Funk, C., Shukla, S., Thiaw, W.M., Rowland, J., Hoell, A., McNally, A., Husak, G., Novella, N., Budde, M., Peters-Lidard, C., Adoum, A., Galu, G., Korecha, D., Magadzire, T., Rodriguez, M., Robjhon, M., Bekele, E., Arsenault, K., Peterson, P., Harrison, L., Fuhrman, S., Davenport, F., Landsfeld, M., Pedreros, D., Jacob, J. P., Reynolds, C., Becker-Reshef, I. and Verdin, J. (2019) Recognizing the famine early warning systems network: over 30 years of drought early warning science advances and partnerships promoting global food security. *Bulletin of the American Meteorological Society*, 100, 1011–1027.

Gavaud, M. (1988) *Nature et localisation de la dégradation des sols au Sénégal*. Dakar: ORSTOM, p. 15.

Gcayi, S.R., Chirima, G.J., Adelabu, S.A., Adam, E. and Abutaleb, K. (2019) Evaluating the potential of narrow-band indices to predict soybean glycine Max L. Merr grain yield in the free state and mpumalanga of South Africa. *Open Access Journal Of Environmental & Soil Science*, 3(1), 265–278.

Giorgi, F., Coppola, E., Raffael, F., Diro, G.B.T., Fuentes-Franco, R., Giuliani, G., Mamgain, A., Llopart, M.B.P., Mariotti, L. and Torma, C. (2014) Changes in extremes and hydroclimate regimes in the CREMA ensemble projections. *Climate Change*, 125, 39–51. https://doi.org/10.1007/s10584-014-1117-0.

Guan, K., Sultan, B., Biasutti, M., Baron, C. and Lobell, D.B. (2015) What aspects of future rainfall changes matter for crop yields in West Africa? *Geophysical Research Letters*, 42(19), 8001–8010. https://doi.org/10.1002/2015GL063877.

Guan, K., Wu, J., Kimball, J.S., Anderson, M.C., Frolking, S., Li, B., Hain, C.R. and Lobell, D.B. (2017) The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. *Remote Sensing of Environment*, 199, 333–349. https://doi.org/10.1016/j.rse.2017.06.043.

Gunn, S. (1998) *Support vector machines for classification and regression*. Southampton: ISIS Research Group, Department of Electronics and Computer Science, University of Southampton, UK. Technical report, p. 11.

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., Ohgaito, R., Ito, A., Yamazaki, D., Okajima, H., Ito, A., Takata, K., Ogochi, K., Watanabe, S. and Kawamiya, M. (2020) Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks. *Geoscientific Model Development*, 13, 2197–2244. https://doi.org/10.5194/gmd-13-2197-2020.

Hamidou, F., Halilou, O. and Vadez, V. (2013) Assessment of groundnut under combined heatand drought stress. *Journal of Agronomy and Crop Science*, 199, 1–11.

Hansen, J.W. (2002) Realizing the potential benefits of climate prediction to agriculture: issues, approaches, challenges. *Agricultural Systems*, 74, 309–330. https://doi.org/10.1016/S0308-521X(02)00043-4.

Harris, I., Jones, P.D., Osborn, T.J. and Lister, D.H. (2014) Updated high-resolution grids of monthly climatic observations—the CRU TS3.10 dataset. *International Journal of Climatology*, 34, 623–642. https://doi.org/10.1002/joc.3711.

Heck, E., de Beurs, K.M., Owsley, B.C. and Henebry, G.M. (2019) Evaluation of the MODIS collections 5 and 6 for change analysis of vegetation and land surface temperature dynamics in North and South America. *ISPRS Journal of Photogrammetry and Remote Sensing*, 156, 121–134.

Held, I.M., Guo, H., Adcroft, A., Dunne, J.P., Horowitz, L.W., Krasting, J., Shevliakova, E., Winton, M., Zhao, M., Bushuk, M., Wittenberg, A.T., Wyman, B., Xiang, B., Zhang, R., Anderson, W., Balaji, V., Donner, L., Dunne, K., Durachta, J., Gauthier, P.P.G., Ginoux, P., Golaz, J.C., Griffies, S.M., Hallberg, R., Harris, L., Harrison, M., Hurlin, W., John, J., Lin, P., Lin, S.J., Malyshev, S., Menzel, R., Milly, P.C.D., Ming, Y., Naik, V., Paynter, D., Paulot, F., Ramaswamy, V., Reichl, B., Robinson, T., Rosati, A., Seman, C., Silvers, L.G., Underwood, S. and Zadeh, N. (2019) Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, 11, 3691–3727. https://doi.org/10.1029/2019MS001829.

Holzman, M.E., Rivasa, R. and Piccolo, M.C. (2014) Estimating soil moisture and the relationship with crop yield using surface temperature and vegetation index. *International Journal of Applied Earth Observation and Geoinformation*, 28, 181–192.

Ibrahim, O.M. (2013) A comparison of methods for assessing the relative importance of input variables in artificial neural networks. *Journal of Applied Sciences Research*, 9(11), 5692–5700.

Iizumi, T., Shin, Y., Choi, J., van der Velde, M., Nisini, L., Kim, W. and Kim, K.-H. (2021) Evaluating the 2019 NARO-APCC Joint Crop Forecasting Service yield forecasts for Northern Hemisphere countries. *Weather and Forecasting*, 86, 879–891. https://doi.org/10.1175/WAF-D-20-0149.1.

Ingram, K.T., Roncoli, M.C. and Kirshen, P.H. (2002) Opportunities and constraints for farmers of West Africa to use seasonal precipitation forecasts with Burkina Faso as a case study. *Agricultural Systems*, 74, 331–349.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.M., Gerber, J.S., Reddy, V.R. and Kim, S.H. (2016) Random forests for global and regional crop yield predictions. *PLoS One*, 11(6), e0156571. https://doi.org/10.1371/journal.pone.0156571.

Jones, P. and Harris, I. (2013) *CRU TS3.21: Climatic Research Unit (CRU) Time-Series (TS) version 3.21 of high resolution gridded data of month-by-month variation in climate (Jan 1901–Dec 2012)*. Leeds: NCAS British Atmospheric Data Centre.

Jones, P. and Thornton, P. (2003) The potential impacts of climate change on maize production in Africa and Latin America in 2055. *Global Environmental Change*, 13, 51–59.

Karatzoglou, A., Meyer, D. and Hornik, K. (2006) Support Vector Machines in R. *Journal of Statistical Software*, 15, 1–28.

Kim, K.H., Shim, P.S. and Shin, S. (2019) An alternative bilinear interpolation method between spherical grids. *Atmosphere*, 2019(10), 123. https://doi.org/10.3390/atmos10030123.

Kim, N.L. and Lee, Y.W. (2016) Machine learning approaches to corn yield estimation using satellite images and climate data: a case of Iowa State. *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography*, 34(4), 383–390.

Kirchner, A. and Signorino, C.S. (2018) Using Support Vector Machines for survey research. *Survey Practice*, 11(1), 1–14. https://doi.org/10.29115/sp-2018-0001.

Lanzante, J.R., Nath, M.J., Whitlock, C.E., Dixon, K.W. and Adams-Smith, D. (2019) Evaluation and improvement of tail behaviour in the cumulative distribution function transform downscaling method. *International Journal of Climatology*, 39, 2449–2460. https://doi.org/10.1002/joc.5964.

Law, R.M., Ziehn, T., Matear, R.J., Lenton, A., Chamberlain, M.A., Stevens, L.E., Wang, Y.P., Srbinovsky, J., Bi, D., Yan, H. and Vohralik, P.F. (2017) The carbon cycle in the Australian Community Climate and Earth System Simulator (ACCESS-ESM1)—part 1: model description and pre-industrial simulation. *Geoscientific Model Development*, 10(7), 2567–2590.

Lee, J., Kim, J., Sun, M.A., Kim, B.H., Moon, H., Sung, H.M., Kim, J. and Byun, Y.H. (2019) Evaluation of the Korea Meteorological Administration Advanced Community Earth-System model (K-ACE). *Asia-Pacific Journal of Atmospheric Sciences*, 2020(56), 381–395. https://doi.org/10.1007/s13143-019-00144-7.

Leroux, L., Falconnier, G.N., Diouf, A.A., Ndao, B., Gbodjo, J.E., Tall, L., Balde, A.B., Clermont-Dauphin, C., Bégué, A., Affholder, F. and Roupsard, O. (2020) Using remote sensing to assess the effect of trees on millet yield in complex parklands of Central Senegal. *Agricultural Systems*, 184, 102918. https://doi.org/10.1016/j.agsy.2020.102918.

Li, L., Yu, Y., Tang, Y., Lin, P., Xie, J., Song, M., Dong, L., Zhou, T., Liu, L., Wang, L., Pu, Y., Chen, X., Chen, L., Xie, Z., Liu, H., Zhang, L., Huang, X., Feng, T., Zheng, W., Xia, K., Liu, H., Liu, J., Wang, Y., Wang, L., Jia, B., Xie, F., Wang, B., Zhao, S., Yu, Z., Zhao B. and Wei, J. (2020) The flexible global ocean-atmosphere-land system model grid-point version 3 (FGOALS-g3): description and evaluation. *Journal of Advances in Modeling Earth Systems*, 12, e2019MS002012.

Li, X., Xiao, J.F., He, B.B., Arain, M.A., Beringer, J., Desai, A.R. and Varlagin, A. (2018) Solar-induced chlorophyll fluorescence is strongly correlated with terrestrial photosynthesis for a wide variety of biomes: first global analysis based on OCO-2 and flux tower observations. *Global Change Biology*, 24, 3990–4008. https://doi.org/10.1111/gcb.14297.

Lichman, M. (2013) *UCI machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science. Available at: http://archive.ics.uci.edu/ml.

Mariotti, L., Diallo, I., Coppola, E. and Giorgi, F. (2014) Seasonal and intraseasonal changes of African monsoon climates in 21st century CORDEX projections. *Climatic Change*, 125, 53–65. https://doi.org/10.1007/s10584-014-1097-0.

Marteau, R., Sultan, B., Moron, V., Alhassane, A., Baron, C. and Traoré, S.B. (2011) The onset of the rainy season and farmers' sowing strategy for pearl millet cultivation in Southwest Niger. *Agricultural and Forest Meteorology*, 151(10), 1356–1369.

Mijwel, M. (2018) *Artificial neural networks advantages and disadvantages*. Available at: available:https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel/.

Müller, W.A., Jungclaus, J.H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., Bunzel, F., Esch, M., Ghosh, R., Haak, H., Ilyina, T., Kleine, T., Kornblueh, L., Li, H., Modali, K., Notz, D., Pohlmann, H., Roeckner, E., Stemmler, I., Tian, F. and Marotzke, J. (2018) A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1. 2-HR). *Journal of Advances in Modeling Earth Systems*, 10, 1383–1413. https://doi.org/10.1029/2017MS001217.

Ndiaye, O. (2018) *Analyse des politiques agricoles et commerciales au Sénégal: Sécurité et souveraineté alimentaires compromises?* Master's degree, University of Sherbrooke.

Ngom, M. M. (2014) *Agrobusiness versus agriculture familiale. L'État dans le tourbillon d'impératifs contradictoires*. Avalaible at: https://www.seneplus.com/article/l'état-dans-le-tourbillon-dimpératifs-contradictoires.

Nigam, S., Jain, R., Marwaha, S., Arora, A. and Singh, K.V. (2019) Plant disease identification using deep learning: a review. *Indian Journal of Agricultural Sciences*, 90(2), 249–257.

Pathak, A. and Pathak, S. (2020) Study of machine learning algorithms for stock market prediction. *International Journal of Engineering Research & Technology*, 9(6), IJERTV9IS060064.

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

Pereira, J.M., Basto, M. and Silva, A.F. (2016) The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance*, 39, 634–641.

Petersen, L.K. (2018) Real-time prediction of crop yields from MODIS relative vegetation health: a continent-wide analysis of Africa. *Remote Sensing*, 10(11), 1726.

Portmann, F.T., Siebert, S. and Döll, P. (2010) MIRCA2000—global monthly irrigated and rainfed crop areas around the year 2000: a new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles*, 24, 1–24.

Prasad, P.V.V., Kakani, V.G. and Upadhyaya, H.D. (2010) Growth and production of groundnut. In: *Soils, Plant Growth and Crop Production*. Oxford: Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, pp. 1–26.

Rapport National sur le Développement Humain au Sénégal. (2009) *Changement Climatique, Sécurité Alimentaire et Développement Humain*. Available at: http://www.pnud.org/content/dam/senegal/docs/OMD/undp-sn-RapportNationalDeveloppementHumainSenegal2009.pdf.

Rasmussen, C.E., Neal, R.M., Hinton, G.E., van Campand, D., Revow, M., Ghahramani, Z., Kustra, R. and Tibshirani, R. (1996) *The DELVE Manual*. Toronto, ON: The University of Toronto. Available at: http://mlg.eng.cam.ac.uk/pub/pdf/RasNeaHinetal96.pdf.

Roudier, P., Sultan, B., Quirion, P., Baron, C., Alhassane, A., Traoré, S.B. and Muller, B. (2012) An ex-ante evaluation of the use of seasonal climate forecasts for millet growers in SW Niger. *International Journal of Climatology*, 32, 759–771.

Sarr, A.B. and Camara, M. (2017) Evolution des indices pluviométriques extrêmes par l'analyse de modèles climatiques régionaux du programme CORDEX: Les projections climatiques sur le Sénégal. *European Scientific Journal*, 13(17), 1857–7881. https://doi.org/10.19044/esg.2017.v13n17p206.

Schafer, J.L. (1999) Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1), 3–15. https://doi.org/10.1191/096228099671525676.

Schlenker, W. and Lobell, D. (2010) Robust negative impacts of climate change on African agriculture. *Environmental Research Letters*, 5, 014010.

Séférian, R., Nabat, P., Michou, M., Saint-Martin, D., Voldoire, A., Colin, J., Decharme, B., Delire, C., Berthet, S., Chevallier, M., Sénési, S., Franchisteguy, L., Vial, J., Mallet, M., Joetzjer, E., Geoffroy, O., Guérémy, J.F., Moine, M.P., Msadek, R., Ribes, A., Rocher, M., Roehrig, R., Salas-y-Mélia, D., Sanchez, E., Terray, L., Valcke, S., Waldman, R., Aumont, O., Bopp, L., Deshayes, J., Éthé, C. and Madec, G. (2019) Evaluation of CNRM Earth-System Model, CNRM-ESM2-1: role of Earth system processes in present-day and future climate. *Journal of Advances in Modeling Earth Systems*, 11, 4182–4227. https://doi.org/10.1029/2019MS001791.

Sellar, A.A., Jones, C.G., Mulcahy, J., Tang, Y., Yool, A., Wiltshire, A., O'Connor, F.M., Stringer, M., Hill, R., Palmieri, J., Woodward, S., Mora, L., Kuhlbrodt, T., Rumbold, S.T., Kelley, D.I., Ellis, R., Johnson, C.E., Walton, J., Abraham, N.L., Andrews, M.B., Andrews, T., Archibald, A.T., Berthou, S., Burke, E., Blockley, E., Carslaw, K., Dalvi, M., Edwards, J., Folberth, G.A., Gedney, N., Griffiths, P.T., Harper, A.B., Hendry, M.A., Hewitt, A.J., Johnson, B., Jones, A., Jones, C.D., Keeble, J., Liddicoat, S., Morgenstern, O., Parker, R.J., Predoi, V., Robertson, E., Siahaan, A., Smith, R.S., Swaminathan, R., Woodhouse, M.T., Zeng, G. and Zerroukat, M. (2019) UKESM1: description and evaluation of the UK Earth System Model. *Journal of Advances in Modeling Earth Systems*, 11, 4513–4558.

Singh, P., Boote, K.J., Kumar, U., Srinivas, K., Nigam, S.N. and Jones, J.W. (2012) Evaluation of genetic traits for improving productivity and adaptationof groundnut to climate change in India. *Journal of Agronomy and Crop Science*, 198, 399–413.

Sultan, B., Barbier, B., Fortilus, J., Mbaye, S.M. and Leclerc, G. (2010) Estimating the potential economic value of the seasonal forecasts in West Africa: a long-term ex-ante assessment in Senegal. *Weather, Climate, and Society*, 2, 69–87.

Sultan, B. and Gaetani, M. (2016) Agriculture in West Africa in the twenty-first century: climate change and impacts scenarios, and potential for adaptation. *Frontiers in Plant Science*, 7, 1–20.

Sultan, B., Guan, K., Kouressy, M., Biasutti, M., Piani, C., Hammer, G., McLean, G. and Lobell, D. (2014) Robust features of future climate change impacts on sorghum yields in West Africa. *Environmental Research Letters*, 9(10), 104006.

Sultan, B. and Janicot, S. (2003) The West African monsoon dynamics. Part II: the "pre-onset" and "onset" of the summer monsoon. *Journal of Climate*, 16, 3407–3427.

Sultan, B., Roudier, P., Baron, C., Quirion, P., Muller, B., Alhassane, A., Ciais, P., Guimberteau, M., Traoré, S.B. and Dingkuhn, M. (2013) Assessing climate change impacts on sorghum and millet yields in West Africa. *Environmental Research Letters*, 8, 014040.

Swart, N.C., Cole, J.N.S., Kharin, V.V., Lazare, M., Scinocca, J.F., Gillett, N.P., Anstey, J., Arora, V., Christian, J.R., Hanna, S., Jiao, Y., Lee, W.G., Majaess, F., Saenko, O.A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D. and Winter, B. (2019) The Canadian Earth System Model version 5 (CanESM5.0.3). *Geoscientific Model Development*, 12, 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019.

Sylla, M.B., Nikiema, P.M., Gibba, P., Kebe, I. and Klutse, N.A.B. (2016) Climate change over West Africa: recent trends and future projections. In: Yaro, J. and Hesselberg, J. (Eds.) *Adaptation to Climate Change and Variability in Rural West Africa*. Cham: Springer. https://doi.org/10.1007/978-3-319-31499-0_3.

Tarawali, A.R. and Quee, D.D. (2014) Performance of groundnut (*Arachis hypogaea* L.) varieties in two agro-ecologies in

Sierra Leone. *African Journal of Agricultural Research*, 9, 1442–1448.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogocho, K., Takemura, T., Sudo, K.K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizutki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M. and Kimoto, M. (2019) Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, 12, 2727–2765. https://doi.org/10.5194/gmd-12-2727-2019.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1), 267–288.

Touré, A.K., Diakhaté, M., Gaye, A.T., Diop, M. and Ndiaye, O. (2020) Sensivity of crop yields to temperature and rainfall daily metrics in Senegal. *America Journal of Rural Development*, 8(1), 1–11. https://doi.org/10.12691/ajrd-8-1-1.

Traoré, S.B., Alhassane, A., Muller, B., Kouressy, M., Somé, L., Sultan, B., Oettli, P., Siéné, L., Ambroise, C., Sangaré, S., Vaksmann, M., Diop, M., Dingkhun, M. and Baron, C. (2010) Characterizing and modeling the diversity of cropping situations under climatic constraints in West Africa. *Atmospheric Science Letter*, 12, 89–95. https://doi.org/10.1002/asl.295.

van Oort, P.A.J. and Zwart, S.J. (2018) Impacts of climate change on rice production in Africa and causes of simulated yield changes. *Global Change Biology*, 24, 1029–1045. https://doi.org/10.1111/gcb.13967.

Vapnik, V. (1998) *Statistical Learning Theory*. New York, NY: Wiley.

Vigaud, N., Vrac, M. and Caballero, Y. (2013) Probabilistic downscaling of GCM scenarios over southern India. *International Journal of Climatology*, 33, 1248–1263.

Vinutha, H.P., Poornima, B. and Sagar, B.M. (2018) Detection of outliers using interquartile range technique from intrusion dataset. In: Satapathy, S., Tavares, J., Bhateja, V. and Mohanty, J. (Eds.) *Information and Decision Sciences. Advances in Intelligent Systems and Computing*, Vol. 701. Singapore: Springer. https://doi.org/10.1007/978-981-10-7563-6_53.

Voldoire, A. (2019a) CNRM-CERFACS CNRM-CM6-1 model output prepared for CMIP6 HighResMIP. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1925.

Voldoire, A. (2019b) *CNRM-CERFACS CNRM-CM6-1-HR model output prepared for CMIP6 ScenarioMIP ssp245*. Earth System Grid Federation. https://doi.org/10.22033/ESGF/CMIP6.4190.

Volodin, E.M., Mortikov, E.V., Kostrykin, S.V., Galin, V.Y., Lykossov, V.N., Gritsun, A., Diansky, N.A., Gusev, A.V. and Iakovlev, N. (2017) Simulation of the present-day climate with the climate model INMCM5. *Climate Dynamics*, 49(11–12), 3715–3734.

Volodin, E.M., Mortikov, E.V., Kostrykin, S.V., Galin, V.Y., Lykossov, V.N., Gritsun, A., Diansky, N.A., Gusev, A.V., Iakovlev, N., Shestakova, A.A. and Emelina, S.V. (2018) Simulation of the modern climate using the INM-CM48 climate model. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 33, 367–374.

Vrac, M. and Ayar, P.V. (2016) Influence of bias correcting predictors on statistical downscaling models. *Journal of Applied Meteorology and Climatology*, 56, 5–26.

Wardlow, B.D., Egbert, S.L. and Kastens, J.H. (2008) Large area crop mapping using timeseries MODIS 250 m NDVI data: an assessment of the U.S. Central Great Plains. *Remote Sensing of Environment*, 112, 1096–1116.

WASP. (2021) *WASP Brief #4 - Early Warning Systems for Adaptation*. Available at: https://wasp-adaptation.org/wasp-publications/wasp-brief-4-early-warning-systems-for-adaptation.

Wei, J., Tang, X., Gu, Q., Wang, M., Ma, M. and Han, X. (2019) Using solar-induced chlorophyll fluorescence observed by OCO-2 to predict autumn crop production in China. *Remote Sensing*, 2019(11), 1715. https://doi.org/10.3390/rs11141715.

Wielicki, B.A., Barkstrom, B.R., Harrison, E.F., III, Smith, G.L. and Cooper, J.E. (1996) Clouds and the earth's radiant energy system (CERES): an earth observing system experiment. *Bulletin of the American Meteorological Society*, 77, 853–868. https://doi.org/10.1175/1520-0477(1996)077<0853:CATERE>2.0.CO;2.

Wilson, R.C., Shenhav, A., Straccia, M. and Cohen, J.D. (2019) The eighty five percent rule for optimal learning. *Nature Communications*, 10, 4646. https://doi.org/10.1038/s41467-019-12552-4.

World Meteorologycal Organisation. (2020) *State of climate services report: move from early warnings to early action*. Geneva: WMO. WMO No. 1252. Available at: https://library.wmo.int/index.php?lvl=notice_display&id=21777#.YNNWrGgzbb2.

Yang, W., Wang, K. and Zuo, W. (2012) Neighborhood component feature selection for high-dimensional data. *Journal of Computers*, 7, 161–168.

Yao, L., Yang, D., Liu, Y., Wang, J., Liu, L., du, S., Cai, Z., Lu, N., Lyu, D., Wang, M., Yin, Z. and Zheng, Y. (2021) A new global solar-induced chlorophyll fluorescence (SIF) data product from TanSat measurements. *Advances in Atmospheric Sciences*, 38, 341–345. https://doi.org/10.1007/s00376-020-0204-6.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sarr, A. B., & Sultan, B. (2023). Predicting crop yields in Senegal using machine learning methods. *International Journal of Climatology*, *43*(4), 1817–1838. https://doi.org/10.1002/joc.7947