# Application of Statistical Models to the Prediction of Seasonal Rainfall Anomalies over the Sahel

HAMADA S. BADR AND BENJAMIN F. ZAITCHIK

*Department of Earth and Planetary Sciences, The Johns Hopkins University, Baltimore, Maryland*

SETH D. GUIKEMA

*Department of Geography and Environmental Engineering, The Johns Hopkins University, Baltimore, Maryland*

(Manuscript received 28 May 2013, in final form 15 October 2013)

ABSTRACT

Rainfall in the Sahel region of Africa is prone to large interannual variability, and it has exhibited a recent multidecadal drying trend. The well-documented social impacts of this variability have motivated numerous efforts at seasonal precipitation prediction, many of which employ statistical techniques that forecast Sahelian precipitation as a function of large-scale indices of surface air temperature (SAT) anomalies, sea surface temperature (SST), surface pressure, and other variables. These statistical models have demonstrated some skill, but nearly all have adopted conventional statistical modeling techniques—most commonly generalized linear models—to associate predictor fields with precipitation anomalies. Here, the results of an artificial neural network (ANN) machine-learning algorithm applied to predict summertime (July–September) Sahel rainfall anomalies using indices of springtime (April–June) SST and SAT anomalies for the period 1900–2011 are presented. Principal component analysis was used to remove multicollinearity between predictor variables. Predictive accuracy was assessed using repeated *k*-fold random holdout and leave-one-out cross-validation methods. It was found that the ANN achieved predictive accuracy superior to that of eight alternative statistical methods tested in this study, and it was also superior to that of previously published predictive models of summertime Sahel precipitation. Analysis of partial dependence plots indicates that ANN skill is derived primarily from the ability to capture nonlinear influences that multiple major modes of large-scale variability have on Sahelian precipitation. These results point to the value of ANN techniques for seasonal precipitation prediction in the Sahel.

## 1. Introduction

Seasonal prediction of precipitation is a core challenge for applied climatology. The attempt to forecast statistics of rainfall several months in advance requires that the forecaster engage with the theory of memory in the climate system, consider trade-offs between physically based dynamical methods and empirically grounded statistical methods, and decide between models that are generalizable and those that provide the best fit to recent observations. The forecaster also faces challenges in evaluation—do we optimize for fit or for predictive skill, and over what time horizon?—and must balance between predictive performance and physical interpretability of any given modeling system. Within this context, the proliferation of dynamical models, statistical techniques, satellite products, and data assimilation systems represents a tremendous opportunity to improve seasonal predictive skill, but it also amplifies the challenge of selecting and applying the appropriate predictive tools for each research question or application.

This challenge is particularly pressing for regions that are highly ecologically or socioeconomically vulnerable to climate variability. The Sahel, an ecoclimatic zone located on the southern edge of the Sahara Desert, stands out in this regard (Kandji et al. 2006). The Sahel extends for approximately 4500 km, from Cape Verde in the west through Senegal, Mauritania, Mali, Burkina Faso, Niger, and Chad to the east. It is limited by the Sahara to the north and by the more humid Sudano–Sahelian belt to the south. Climatically, the Sahel is a

Denotes Open Access content.

*Corresponding author address:* Hamada Badr, The Johns Hopkins University, Olin Hall, 3400 N. Charles St., Baltimore, MD 21218.
E-mail: badr@jhu.edu

DOI: 10.1175/JAMC-D-13-0181.1

transitional zone between the arid Sahara and the tropical forest that borders the maritime coast. Annual rainfall varies from approximately 200 mm in the north of the Sahel to approximately 600 mm in the south and is subject to large variability at intraseasonal to decadal time scales.

Sahelian precipitation variability and its impacts have motivated a significant number of studies on the drivers of precipitation in the region. Researchers have consistently found that the El Niño–Southern Oscillation (ENSO) has a significant impact on rainy season precipitation (Folland et al. 1991; Giannini et al. 2003; Rowell et al. 1995). Influences of the tropical Atlantic Ocean (Vizy and Cook 2002) and tropical Indian Ocean (Palmer 1986; Rowell 2001) have also been identified, and the extratropical influence also appears to be significant, including anomalies in Mediterranean Sea temperatures (Rowell 2003). This diversity of external drivers suggests that the processes governing precipitation variability in the Sahel may interact in complex and nonlinear ways, posing a challenge for seasonal forecast systems that derive their skill from memory in large-scale SST or surface pressure patterns.

Nevertheless, numerous statistical and dynamical efforts at seasonal prediction in the Sahel have been proposed, and some have had considerable success. Statistical methods, defined here as models based on purely empirical relationships derived from climate reanalyses or general circulation models (GCMs), can take any number of structural forms, but most models proposed for the Sahel are based on multivariate linear regression schemes, using standard climate indices, SST anomalies averaged over selected domains, or principal components of global SST or sea level pressure (SLP) variability as inputs. For example, Fontaine and Philippon (2000) and Fontaine et al. (1999) used statistical approaches to predict July–September (JAS) Sahelian rainfall for the period 1968–97 using April–May SST patterns and regional moist static energy. They showed that SST predictors are more useful for predicting longer-term trends than interannual variability; better hindcast skills were obtained by adding PBL energy content to the predictors. Garric et al. (2002) used both statistical and dynamical approaches to predict monsoon rainfall anomalies over the central Sahel. Their statistical predictions utilized linear regressions from SST and rainfall predictors over 1968–97, which is a relatively dry period, and their dynamical predictions used the Action de Recherche Petite Echelle Grande Echelle (ARPEGE) atmospheric model forced by observed SST over the 1979–93 period. They found that the dynamical approach was less skillful overall but was better in simulating the variability of the large-scale monsoon circulation; therefore, they adopted a hybrid statistical–dynamical approach to achieve better predictions.

Ndiaye et al. (2009, 2011) developed and applied a model output statistics (MOS) approach to correct poor GCM seasonal predictions of Sahelian rainfall over 1981–2008. Their MOS method used empirical orthogonal functions of the model's regional (tropical Atlantic and West Africa) 925-hPa wind as precipitation predictors in a regression model. The MOS system was applied to GCM experiments using SST anomalies from the months of June, May, and April, respectively, to estimate the potential of the system to make forecasts with lead times between 0 and 2 months ahead of the JAS season. They explored the ability of several atmosphere-only and coupled ocean–atmosphere GCMs for the prediction of seasonal JAS Sahel rainfall and found that an MOS approach that uses predicted low-level winds over the tropical Atlantic and the western part of West Africa yields good Sahel rainfall skill for all models.

As is evident from even this brief review of seasonal prediction efforts, researchers have utilized a diversity of modeling tools, study periods, and evaluation metrics to derive skillful predictions. Moreover, investigators working on seasonal prediction have differed in the relative emphasis they place on predictive skill versus explanatory mechanism (McIntosh et al. 2005).

Here, we consider the problem of developing a statistical model capable of predicting summertime (JAS) Sahelian precipitation as a function of springtime (April–June) predictors. The goal is to produce a model that provides skillful prediction for the complete period of data availability, 1900–2011, which includes the pre-1950 regime of weak interannual persistence and the post-1950 period of stronger persistence (Brooks 2004; Nicholson 1995), such that processes associated with this shift will be included in the model. In this respect, our analysis complements recent seasonal prediction studies that have focused exclusively on prediction in the post-1950 regime (Fontaine et al. 1999; Garric et al. 2002). Further, in order to facilitate interpretation within the context of previous literature, we employ only previously defined and widely used climate indices as predictors. Under these constraints, we test the predictive skill of a diverse set of models: regression, tree based, and artificial neural network machine learning. The predictive skill of each model is compared against the others, as well as against the average of all models, the null model (the mean of the response variable in the training dataset), and a memory model that uses last year's rainfall as this year's prediction. To our knowledge, no such rigorous methodological comparison has been performed for the Sahel. High-performing models are then further

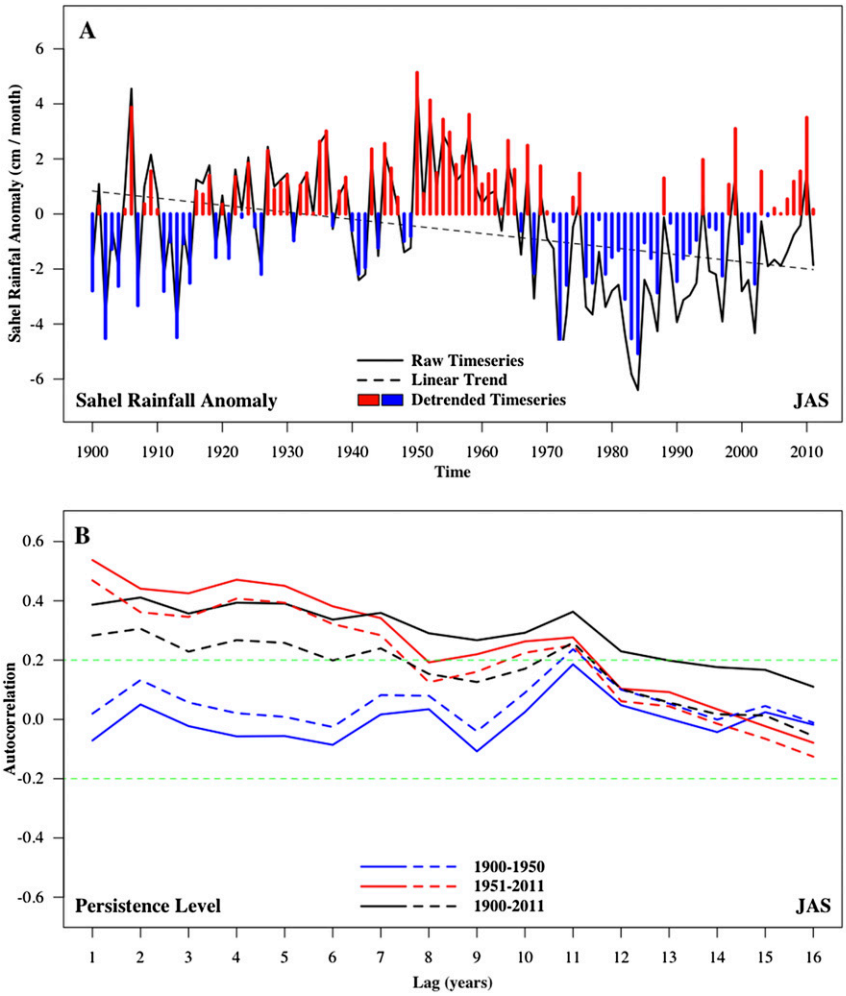


FIG. 1. Sahelian rainfall and its persistence. (a) Summer (JAS) time series for the period 1900–2011 and (b) autocorrelation vs lag (yr) of raw (solid lines) and detrended (dashed lines) time series over different time periods: 1900–50, 1951–2011, and the entire 1900–2011 record.

investigated to identify the physical climate mechanisms that underlie model predictions, in order to provide confidence in the model results and to contribute to our basic understanding of the drivers of Sahelian precipitation variability.

The description of the data is presented in section 2 and methods used are presented in section 3. Section 4 presents the results and associated discussions including model comparisons, cross validation, and identification of various potential predictors. Conclusions are offered in section 5.

## 2. Data description

This study focuses on the prediction of variability in total summertime Sahel rainfall, defined as the anomaly in July–September precipitation, averaged over the geographic region 10°–20°N and 10°W–20°E, which was defined using a rotated principal component analysis (PCA) of African precipitation (Janowiak 1988). This definition approximates that used in previous studies (Dezfuli and Nicholson 2011; Nicholson et al. 2012) and there is strong correlation in interannual precipitation variability across the area. The raw data for the Sahel rainfall anomaly are obtained from University of Washington's Joint Institute for Study of the Atmosphere and Ocean (JISAO; jisao.washington.edu/data_sets/sahel). The JISAO Sahelian rainfall anomaly index is derived from the well-established National Oceanic and Atmospheric Administration's (NOAA) Global Historical Climatology Network (GHCN) dataset and has been used in previously published studies of the Sahel, such

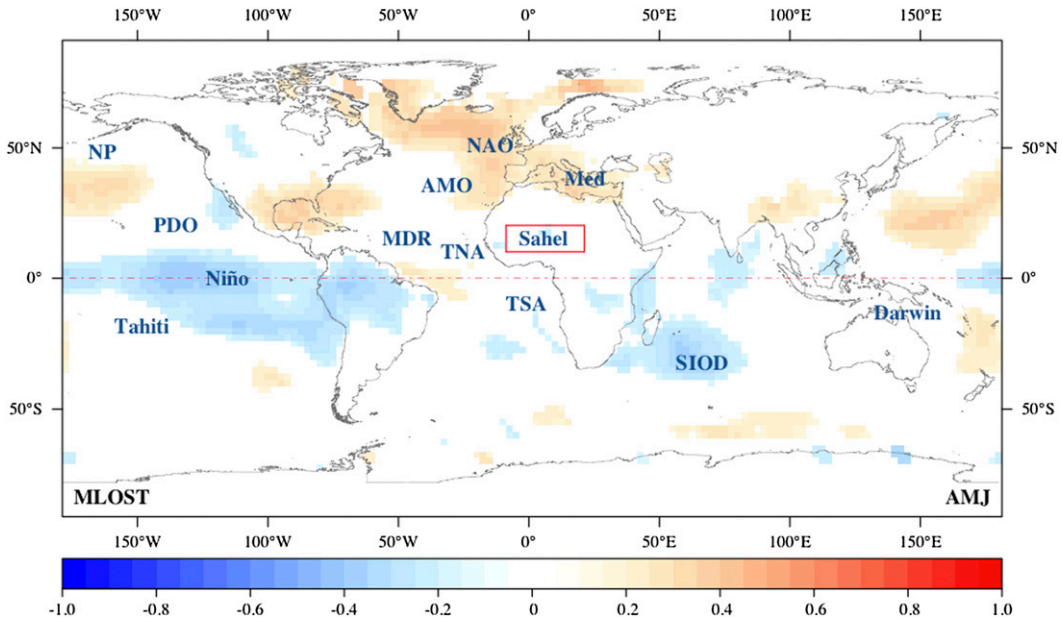


FIG. 2. Correlation patterns of summer (JAS) Sahelian rainfall anomaly with spring (AMJ) MLOST data for the period 1900–2011. All correlations are statistically significant at the 95% confidence level; insignificant correlations are masked out. Centers of action of large-scale climate indices used in this study are labeled in blue, and the Sahel region is marked by a red box.

as Haywood et al. (2013). It uses the raw precipitation stations in the GHCN product, without homogeneity adjustments. Here, we consider the entire historical record, 1900–2011.

GHCN data show that precipitation was above the long-term mean from 1915 through the late 1930s and during the 1950s–1960s, after which it was persistently below the long-term mean, with the largest negative anomalies in the early 1980s (Fig. 1a). The dry episode at the end of the twentieth century was of a particularly long duration, and has been described by Hulme (2001) as unprecedented in the Sahel and also in any other dryland region within the context of the modern observational record (Brooks 2004). There is also a notable shift in the nature of persistence in seasonal precipitation totals between the pre-1950 period, when interannual correlation in precipitation was quite low, and the post-1950 period, which is characterized by significant correlation on interannual time scales and weak anticorrelation at decadal time scales (Hulme 2001). This shift in persistence patterns (Fig. 1b) has been noted in previous studies (Brooks 2004; Nicholson 1995), and it has been associated with global SST (Brooks 2004).

The correlation patterns of summertime rainfall anomalies and springtime global SSTs and land temperatures are shown in Fig. 2. It is clear that the correlations are weak in general, and that regions with relatively high correlation are distributed in the Pacific, North Atlantic, and Indian Oceans in patterns that align with conventional climate indices such as ENSO, the Atlantic multidecadal oscillation (AMO), and the subtropical Indian Ocean dipole (SIOD). While one could use the correlations shown in Fig. 2 to define customized predictors for Sahel precipitation, the fact that zones of correlation generally align with well-known indices of climate variability indicates that combinations of these standard indices can serve as a basis for skillful prediction. The use of standard indices is advantageous in that it allows predictions to be understood within the context of extensive existing research on prevailing modes of global climate variability.

Based on this reasoning, we chose to use a set of climate indices and global temperature indicators to construct predictive models. Figure 2 shows the centers of action for the climate indices used in this study, which are summarized in Table 1. Several seasonal averages of 1–5 months (ranging from the previous season to a year lag) were tested as predictors of seasonal rainfall anomalies over the Sahel. It was found that all models perform best in fit and predictive skill when April–June (AMJ) predictors are used to predict JAS rainfall. For all models, the JAS Sahel rainfall index is the response variable, while the oceanic indices are computed as the mean of the previous season (AMJ). The SST indices are calculated from the extended reconstructed SST

TABLE 1. Brief description and geographic domain of the variables used in model development. The response variable is Sahel_Precip, while other variables are used as raw predictors for PCA. The boldfaced variables are derived from subsets of variables that represent the same physics/region. The latitude and longitude pairs give the bounds of the domain.

| Variable | Description | Geographic domain | | | |
|---|---|---|---|---|---|
| | | Lat | | Lon | |
| Sahel_Precip | Sahel rainfall anomaly | 20°N | 10°N | 20°E | 10°W |
| SATA_LNH | Global mean SATA over NH land | 90°N | 0° | Land only | |
| SATA_LSH | Global mean SATA over SH land | 0° | 90°S | Land only | |
| SATA_ONH | Global mean SATA over NH ocean | 90°N | 0° | Ocean only | |
| SATA_OSH | Global mean SATA over SH ocean | 0° | 90°S | Ocean only | |
| **SATA** | The PC1 of the SATA indices (represents ~88% of the total variance) | | | | |
| Niño-1.2 | Niño region 1 + 2 SST | 0° | 10°S | 80°W | 90°W |
| Niño-3 | Niño region 3 SST | 5°N | 5°S | 90°W | 150°W |
| Niño-3.4 | Niño region 3.4 SST | 5°N | 5°S | 120°W | 170°W |
| Niño-4 | Niño region 4 SST | 5°N | 5°S | 160°E | 150°W |
| **Niño** | The PC1 of Niño indices and SOI (represents ~83% of the total variance) | | | | |
| SLP_Darwin | Sea level pressure at Darwin | 13°S | | 131°E | |
| SLP_Tahiti | Sea level pressure at Tahiti | 18°S | | 150°W | |
| **SOI** | Southern Oscillation index (the difference between the SLP at Tahiti and Darwin) | | | | |
| **ENSO** | The PC1 of Niño indices and SOI (represents ~70% of the total variance) | | | | |
| AMO | Atlantic multidecadal oscillation | 70°N | 0° | 10°W | 75°W |
| NAO | North Atlantic Oscillation index (PC) | 80°N | 20°N | 40°E | 90°W |
| SST_MDR | Hurricane main development region SST | 20°N | 10°N | 20°W | 85°W |
| TNA | Tropical northern Atlantic SST | 25°N | 5°N | 15°W | 55°W |
| **NAI** | The PC1 of North Atlantic indices (represents ~74% of the total variance) | | | | |
| TSA | Tropical southern Atlantic SST | 0° | 20°S | 10°E | 30°W |
| SIOD_E | Eastern subtropical Indian Ocean SST | 18°S | 28°S | 100°E | 90°E |
| SIOD_W | Western subtropical Indian Ocean SST | 27°S | 37°S | 65°E | 55°E |
| **SIOD** | Subtropical Indian Ocean dipole (the difference between SIOD_E and SIOD_W) | | | | |
| NP | Northern Pacific pattern | 65°N | 30°N | 160°E | 140°W |
| PDO | Pacific decadal oscillation (PC) | 70°N | 60°S | 60°W | 100°E |
| SST_Med | Mediterranean Sea SST | 45°N | 30°N | 25°E | 0° |

(ERSST), version 3b (v3b), dataset (Smith et al. 2008). The surface air temperature anomaly (SATA) indices are obtained from NOAA/National Climatic Data Center (www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php). The other indices listed in Table 1 [North Atlantic Oscillation (NAO); Pacific decadal oscillation (PDO); and SLP at Darwin, Northwest Territory, Australia, and Tahiti, French Polynesia] are obtained from the NOAA/Office of Oceanic and Atmospheric Research/Earth System Research Laboratory Physical Sciences Division archive (www.esrl.noaa.gov/psd/gcos_wgsp/Timeseries).

NOAA Merged Land–Ocean Surface Temperature Analysis (MLOST, 1900–2011), version 3.5.3—from the GHCN land surface temperature and NOAA ERSST v3b dataset—and the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis (Kalnay et al. 1996) atmospheric fields (1948–2011) were used to support a physically based explanation of the statistical results and to test associations with mechanisms of interannual precipitation variability over the Sahel.

## 3. Methods

Multiple statistical models (see appendix A Table A1) were developed to predict seasonal rainfall anomalies over the Sahel as a function of large-scale indices of SATA, SST, surface pressure, and other predictors. All models were compared with each other and with three reference models: average model (mean predictions from models), null/climatology model (mean rainfall), and memory/persistence model (last year's rainfall). All models were assessed and compared in terms of the predictive skill as well as the goodness of fit. We use these terms in the manner that they are typically applied in the statistical modeling literature: goodness of fit is evaluated for in-sample accuracy while predictive skill is evaluated for out-of-sample accuracy through cross validation (Arlot and Celisse 2010; Von Storch 1999; Von Storch and Zwiers 2001).

The final statistical computations and graphics were carried out in R open-source software version 3.0.2 (25 September 2013, ''Frisbee Sailing,'' 64 bit) on a server of 4 six-core Intel Xeon X5670 CPUs (2.93 GHz). A script

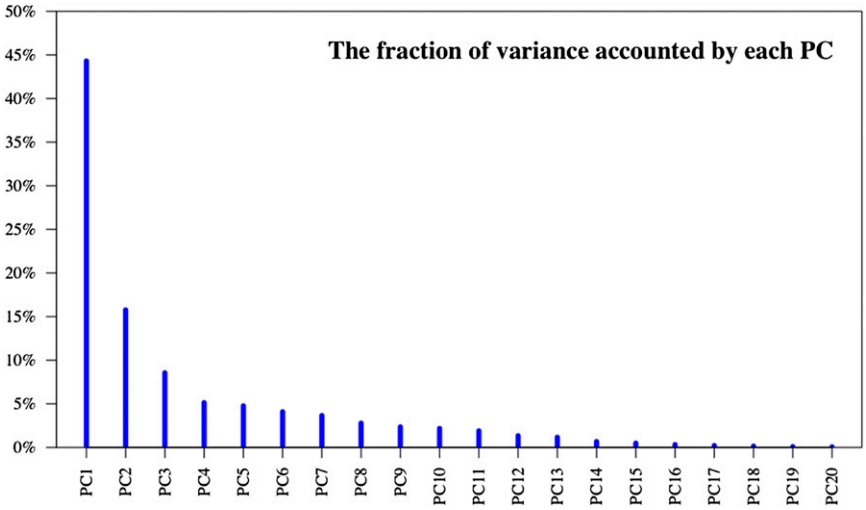


FIG. 3. Percentage variance explained by each PC of the spring (AMJ) large-scale climate indices listed in Table 4 for the period 1900–2012. The first three PCs (PC1–3) explain the majority of the variance (~69%), and the first eight PCs (PC1–8) explain ~89% of the total variance.

was written for model fitting, cross validation, and visualization utilizing the set of R packages discribed in appendix A (Table A2), including a function for multidimensional partial dependence plots (PDPs). The computational time for the cross-validation analysis was about 15 min per holdout; the number of holdouts was chosen equal to the sample size (112) for consistency with the cross-validation methods used.

The preprocessing of the data including the multicollinearity problem and PCA is presented in section 3a. Descriptions of the statistical models are briefly presented in section 3b and are extensively defined in appendix A. Section 3c presents the methods used for model assessment and comparison including variable selection, cross validation, and hypothesis tests for model comparison.

### *a. Data preprocessing*

Modes of large-scale climate variability are characterized by a number of interrelated and often highly correlated indices. ENSO, for example, has multiple definitions that capture distinct but related aspects of eastern and central Pacific SST and SLP variability. This presents a methodological challenge for seasonal prediction, as multicollinearity between predictor variables can inhibit the identification of significant predictor variables and undermine the performance of some statistical algorithms (Farrar and Glauber 1967). As expected, the multiple, overlapping indices listed in Table 1 exhibit significant multicollinearity [variance inflation factor (VIF) $> 5$]. Multicollinearity could be solved by simply selecting a subset of predictors, avoiding those that show significant VIF. This approach, however, does not make maximum use of the information contained in the full suite of indices. Moreover, removing a variable or more outside the cross-validation analysis—based on fit/correlation with respect to the response variable—biases the results. For this reason, we conducted a PCA on the full list of proposed predictors. All principal components (PCs) were retained, to allow for objective downselection of PCs in the statistical modeling process. The percentage of variance in each PC is shown in Fig. 3. Models were constructed using both raw and detrended PCs. The raw PCs provided slightly better performance, since low-frequency variability captured by the trend in certain PCs does contribute to the skill for the long time period considered in this study. Only the results from models with raw PCs are presented in this paper.

From the perspective of model robustness, there are numerous methods through which variables can be selected for a predictive model, and the orthogonality of PCA does not in its own right guarantee that a PCA-based model will be superior to other approaches. For this reason we tested multiple approaches to covariate definition and selection. This includes a model based on the physically based selection of climate indices that we include in the model evaluation presented in the paper. As described in the results, this model underperforms relative to PCA-based models. We found this to be true for all models constructed using raw indices rather than PCs of indices.

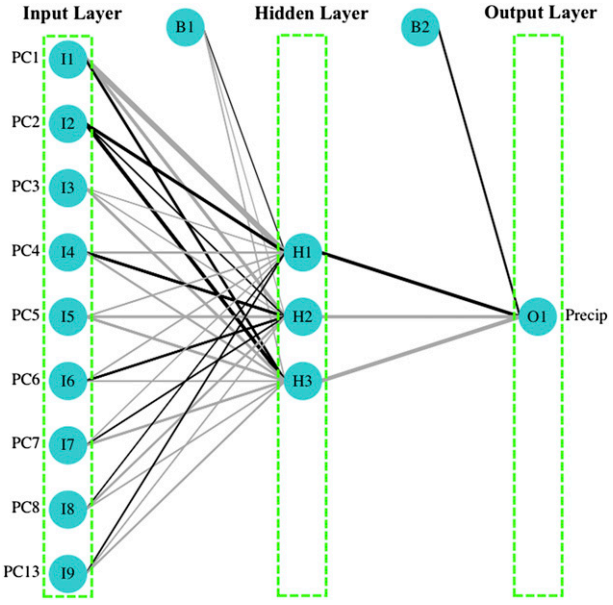


FIG. 4. Schematic diagram for the ANN in the final ANN model. It is a feed-forward single-layer network with three units in the hidden layer. The connections between nodes are shown in black (gray) lines for positive (negative) weights. Line thickness is proportional to relative weight. The nine nodes in the input layer units (labeled as I1–I9) represent the nine PCs used as covariates in the model, and the response variable is shown in the output layer unit (labeled as O1). The hidden layer units are labeled as H1–H3 (three units), which were specified using cross-validated tuning of the hidden layer size. B1 and B2 are bias layers that apply constant values to the nodes, similar to intercept terms in a regression models.

### *b. Statistical models*

Ten different statistical models, representing both linear and nonlinear statistical techniques, were developed to predict seasonal rainfall anomalies over the Sahel (appendix A). For comparison and evaluation, three other models were included: average, null, and memory models. The average model predicts the response by averaging the predictions of all models, while the null model uses the mean of the response variable in the training data. The memory model was included to capture the persistence of precipitation from the previous year; it uses last year's rainfall as this year's prediction. All models were tuned inside the cross-validation analysis in order to find the optimal parameters for each model type.

Generalized linear models (GLMs) and generalized additive models (GAMs) are likelihood-based regression models. GLM generalizes the standard ordinary least squares (OLS) model (Nelder and Wedderburn 1972) by adding a link function, which relates the mean of the response to the predictor variables (Cameron and Trivedi 1998). GAM replaces the GLM link function with a nonparametric smoothing function, allowing for a nonlinear relationship between the response and the predictors (Hastie and Tibshirani 1986). The Sahel rainfall anomalies exhibited a normal Gaussian distribution and therefore a normal identity link function was used in the construction of the GLM and the cubic regression spline was used for the GAM.

Multivariate adaptive regression spline (MARS) is a form of regression analysis introduced by Friedman (1991) that is similar in theory to GLM but is based on automatically selected basis functions. MARS models are constructed in the same fashion as recursive partitioning trees, with the addition of an option to allow a forward and backward pass (Hastie et al. 2009). A MARS model was developed and tested for the dataset.

In addition to GLM, GAM, and MARS, four tree-based modeling techniques were applied to the data: a classification and regression tree (CART; Breiman 1984), Bayesian additive regression trees (BART; Chipman et al. 2010), bagged categorical and regression trees (BCART; Sutton 2005), and a random forest (RF) model (Breiman 2001). Cost-complexity pruning of tree objects was considered. For the BART model, appropriate prior information for the error variance was used.

Finally, an artificial neural network (ANN) was implemented. An ANN is a network of many simple units, each with an associated threshold value, connected by weighted communication channels. The signals propagate through the network (Collier 1994). An ANN has at least three basic layers: the inputs, the hidden layer, and the outputs. For regularization, weight decay adds a penalty term to the error function. Feed-forward ANNs generalize linear regression functions (Venables et al. 1994). The ordinary least squares training criterion is $(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w})$, where $\mathbf{y}$ is the response vector, $\mathbf{X}$ is a matrix of predictors, and $\mathbf{w}$ is the weight vector to be computed. The penalty term in weight decay uses the training criterion $(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + d^2\mathbf{w}^{\mathrm{T}}\mathbf{w}$, where $d$ is the decay rate. In this study, a single-layer ANN model with different sizes (the final model has a single hidden layer with three units) and unit weight decay was used.

The input units distribute information from inputs to the units in the hidden layer, which compute a fixed function $\phi_h$ of the summation of the inputs plus a bias (constant $\alpha$). Similarly, the output units have the same form as the output function $\phi_o$.

The response is then computed by

$$y_k = \phi_o\left[\alpha_k + \sum_h w_{hk}\phi_h\left(\alpha_h + \sum_i w_{ih}x_i\right)\right].$$

The hidden layer has a logistic function with linear output units in the form

TABLE 2. Model comparisons based on predictive skill (mean $\mu$ and standard deviation $\sigma$). The table summarizes cross-validation results for the best-performing models of each type in terms of standard correlation and error measures. RRHCV estimates the model performance in terms of out-of-sample COR and standard errors (MAD, MAE, MSE, and RMSE), and LOOCV estimates model performance in terms of out-of-sample AE. Better models have lower error measures and higher correlation coefficients. The boldface rows are the final simplest and best-performing models: ANN and GLM with nine PCs (PC1–8 and PC13).

| | LOOCV AE | | RRHCV | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | COR | | MAD | | MAE | | MSE | | RMSE | |
| Model | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| **GLM** | **0.62** | **0.44** | **0.66** | **0.16** | **0.72** | **0.23** | **0.61** | **0.12** | **0.56** | **0.20** | **0.73** | **0.14** |
| SGLM | 0.64 | 0.47 | 0.60 | 0.19 | 0.75 | 0.26 | 0.64 | 0.13 | 0.62 | 0.23 | 0.78 | 0.14 |
| GAM | 0.63 | 0.45 | 0.65 | 0.16 | 0.72 | 0.24 | 0.61 | 0.12 | 0.57 | 0.21 | 0.74 | 0.14 |
| SGAM | 0.63 | 0.45 | 0.64 | 0.16 | 0.72 | 0.25 | 0.62 | 0.13 | 0.57 | 0.21 | 0.74 | 0.14 |
| MARS | 0.81 | 0.59 | 0.36 | 0.24 | 0.86 | 0.33 | 0.78 | 0.17 | 0.92 | 0.39 | 0.94 | 0.20 |
| CART | 0.77 | 0.56 | 0.38 | 0.22 | 0.88 | 0.27 | 0.78 | 0.17 | 0.92 | 0.36 | 0.94 | 0.19 |
| BCART | 0.68 | 0.54 | 0.51 | 0.21 | 0.76 | 0.26 | 0.68 | 0.15 | 0.75 | 0.31 | 0.85 | 0.18 |
| BART | 0.63 | 0.48 | 0.63 | 0.17 | 0.71 | 0.25 | 0.62 | 0.13 | 0.61 | 0.23 | 0.77 | 0.15 |
| RF | 0.67 | 0.50 | 0.58 | 0.18 | 0.75 | 0.24 | 0.66 | 0.13 | 0.68 | 0.26 | 0.81 | 0.16 |
| **ANN** | **0.58** | **0.42** | **0.71** | **0.14** | **0.65** | **0.25** | **0.57** | **0.13** | **0.50** | **0.19** | **0.69** | **0.14** |
| Avg | 0.61 | 0.45 | 0.68 | 0.16 | 0.70 | 0.24 | 0.61 | 0.13 | 0.56 | 0.21 | 0.74 | 0.14 |
| Null | 0.83 | 0.56 | — | — | 0.98 | 0.30 | 0.83 | 0.16 | 0.99 | 0.35 | 0.98 | 0.18 |
| Memory | 0.90 | 0.65 | 0.41 | 0.24 | 1.01 | 0.31 | 0.86 | 0.17 | 1.16 | 0.42 | 1.06 | 0.20 |

$$l(z) = \exp(z)/[1 + \exp(z)].$$

A schematic for the final ANN model is shown in Fig. 4.

### c. *Model assessment and comparison*

The statistical models were assessed and compared for goodness of fit and predictive skill in terms of standard in-sample and out-of-sample measures, respectively. Here, we used the correlation coefficient (COR; larger is better), median absolute deviation (MAD; smaller is better), mean absolute error (MAE; smaller is better), mean square error (MSE; smaller is better), and root-mean-square error (RMSE; smaller is better). The RMSE and MSE are popular because of their theoretical relevance in statistical modeling, but they are more sensitive to outliers than MAE or MAD (Hyndman and Koehler 2006). Sections 3c(1), 3c(2), and 3c(3) present the variable selection, cross validation, and hypothesis tests for the model comparison, respectively.

#### 1) VARIABLE SELECTION

Models with many covariates tend to have low bias and high variance, while models with few covariates have high bias and low variance—the bias-variance trade-off—and the best predictions result from balancing these two extremes. For this reason we performed variable selection on candidate models and include both full-covariate models and nested, reduced-variable models in the cross-validation analysis. Covariates for the GLM were chosen through stepwise selection. For GAMs, the full-covariate model and a nested model selected by penalizing smoothing terms are also included in the cross-validation analysis with all other models. For the MARS model, the model is selectively pruned using the maximum number of terms based on a bootstrap cross-validation approach. For the tree-based data mining techniques, the trees are also pruned by recursively snipping off the least important splits. For the ANN model, the size of the hidden layer and the weight decay are selected based on bootstrap cross validation.

To ensure a complete evaluation of all potential models, we also test models that use various subsets of both original and PC covariates, as described in appendix A. Through this process we are able to select the simplest model (fewest covariates) that achieves predictive skill scores that are statistically indistinguishable from other top-performing models. The variable selection is done within the cross-validation analysis based on the training dataset for each round. In addition, these models are compared to models that used the original climate indices instead of principal components of the indices (see appendix A). This index-based model is included as a reference to assess the value of performing PC transformations on the physically based climate predictors. The predictive accuracy of all models is assessed through the cross-validation analysis.

#### 2) CROSS-VALIDATION ANALYSIS

The skill of the model predictions was assessed through repeated $k$-fold random-holdout (RRHCV) and leave-one-out cross-validation (LOOCV) methods. In RRHCV, a random holdout of 10% of the data was used for validation (test data) and the remaining 90% of

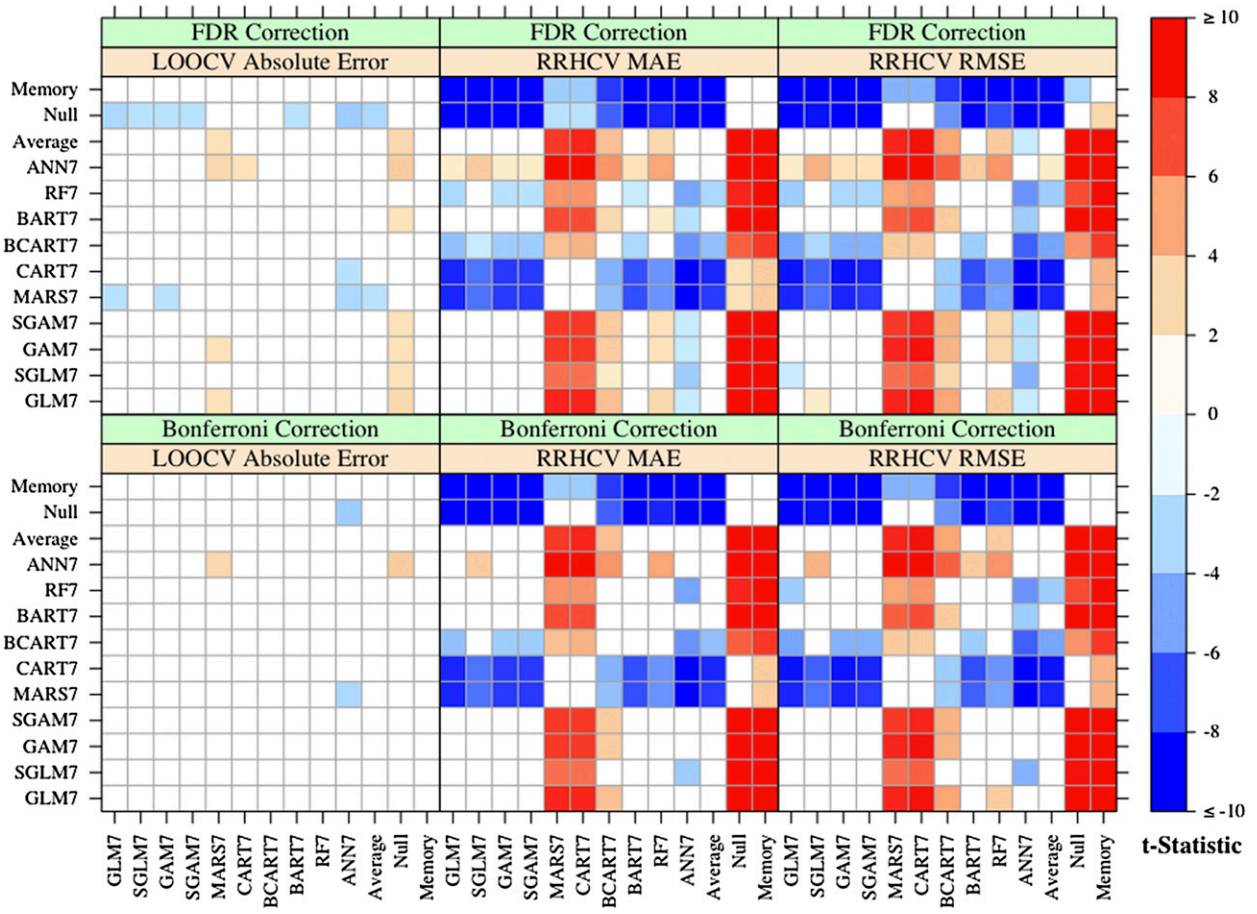


FIG. 5. Model comparisons for the best-performing models of each type. The "7" in model labels indicates that this is model subset ID 7 as described in appendix A. Grids summarize the cross-validation results in terms of standard out-of-sample error measures: LOOCV absolute error and RRHCV MAE and RMSE. The color of each cell represents the value of the $t$ statistic obtained from a Student's $t$ test between the model along the $x$ axis ($M_x$) and the model along the $y$ axis ($M_y$). Colored cells are significant model differences at the 95% confidence level. Positive values (red colors) indicate positive differences (i.e., $M_y$ has lower errors and outperforms $M_x$), and vice versa. For example, the large number of red cells in the ANN7 rows indicates strong performance relative to other models while the large number of blue cells in the MARS7 rows indicates the opposite. The $p$ values are corrected using both (top) FDR and (bottom) Bonferroni corrections.

the data were used to train the models (training data). The data are randomly split into two segments, followed by training on the larger segment, and predicting the holdout segment; this is done repeatedly $k$ times, with $k$ chosen equal to the sample size (112). RRHCV provides a large number of performance estimates but underestimates the performance variance for comparisons, because of the overlap between the training and test datasets. In LOOCV, a single observation from the sample was used as test data and the remaining observations as training data. This provides unbiased performance estimation but with a very large variance. Both RRHCV and LOOCV provide a measure of the generalizability of the predictive model. The results of the RRHCV are presented in terms of the out-of-sample COR, MAD, MAE, MSE, and RMSE while the results of the LOOCV method are presented in terms of the out-of-sample absolute error (AE).

### 3) MODEL COMPARISON

The out-of-sample error measures for all statistical models were compared using all possible combinations of pairwise $t$ tests. The total number of models compared was 103 (10 model types listed in Table A1 by 10 subsets of covariates listed in Table A3 plus three models: average, null, and memory models). The total number of comparisons for the final models of each type is 78 selected through covariate selection inside the same cross-validation analysis, which equals the number of all combinations of the 13 models (Table A1) comparing 2 models at a time. All significance tests were corrected for multiple hypotheses testing using the Bonferroni

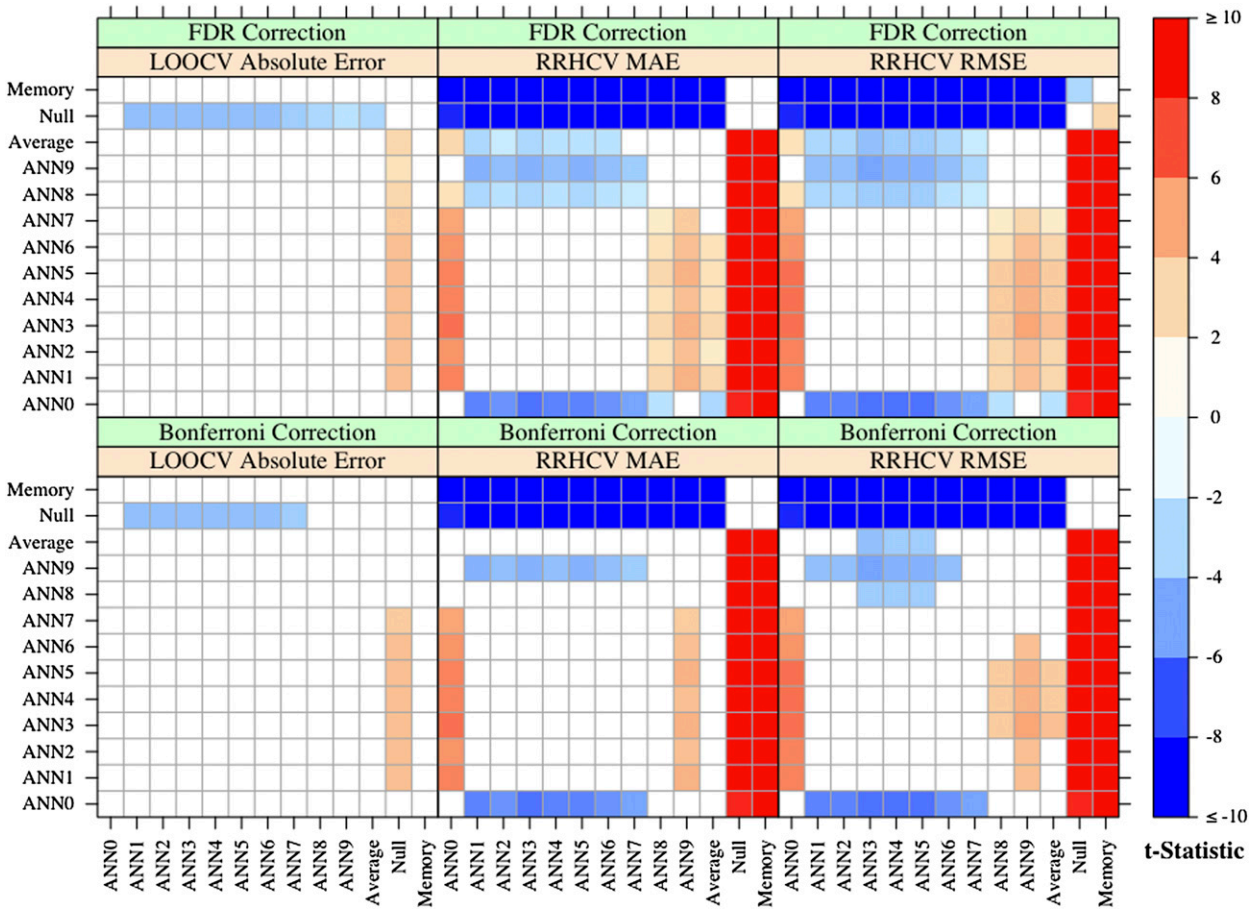


FIG. 6. As in Fig. 5, but for comparisons between ANN models with different sets of variables. Model numbers refer to the subsets described in appendix A. PC-based ANN models (ANN1–9) outperform the original-indices reference model (ANN0). ANN7 is the simplest model (using nine PCs), and it is statistically indistinguishable from the larger-size PC-based models (ANN1–6).

correction and the less-conservative false discovery rate (FDR) method (Benjamini and Hochberg 1995), defined as the expected proportion of false positives among all significant tests. The comparisons were assessed at the 95% confidence level ($\alpha = 0.05$).

## 4. Results and discussion

This section presents only the most relevant results, while more details can be found in the appendixes. Section 4a presents the model comparison in terms of predictive skill, while section 4b shows the goodness-of-fit results. Variable influence and importance together with the associated implications for the best-performing models are presented in section 4c. Appendix B shows supplementary results for model comparisons from the hypothesis tests.

### *a. Predictive skill*

Table 2 summarizes the model comparisons from cross validation. Only the best-performing model of each type is included in these comparisons. RRHCV estimates model performance in terms of out-of-sample COR and standard errors (MAD, MAE, MSE, and RMSE), and LOOCV estimates model performance in terms of out-of-sample absolute error. Better models have lower error measures and higher correlation coefficients. The statistical significance of the differences in the means between the different models is considered in the results of hypothesis tests based on the 95% confidence level. As this large number of significance tests is difficult to show in table form, Fig. 5 presents a graphical representation of the significant differences between models according to LOOCV absolute error and RRHCV MAE and RMSE. The final simplest and best-performing models are ANN and GLM with nine PCs (PC1–8 and PC13).

The statistical significance of differences between models was calculated using Bonferroni- and FDR-adjusted $p$ values for each of these metrics. The results of the statistical hypothesis tests for the other metrics

TABLE 3. Model comparisons based on goodness of fit. The table summarizes the overall quality of fit for the best-performing models of each type in terms of correlation coefficients and standard in-sample errors. Better models have higher correlation coefficient and lower error measures. Boldface rows are the final best-performing models: ANN and GLM with nine PCs (PC1–8 and PC13).

| Model | COR | MAD | MAE | MSE | RMSE |
|---|---|---|---|---|---|
| **GLM** | **0.71** | **0.70** | **0.57** | **0.49** | **0.70** |
| SGLM | 0.71 | 0.76 | 0.58 | 0.50 | 0.71 |
| GAM | 0.72 | 0.71 | 0.57 | 0.48 | 0.70 |
| SGAM | 0.72 | 0.71 | 0.57 | 0.48 | 0.70 |
| MARS | 0.41 | 0.93 | 0.74 | 0.83 | 0.91 |
| CART | 0.48 | 0.80 | 0.69 | 0.77 | 0.87 |
| BCART | 0.82 | 0.49 | 0.47 | 0.36 | 0.60 |
| BART | 0.86 | 0.53 | 0.44 | 0.30 | 0.55 |
| RF | 0.97 | 0.34 | 0.28 | 0.12 | 0.35 |
| **ANN** | **0.81** | **0.59** | **0.48** | **0.34** | **0.59** |
| Avg | 0.81 | 0.64 | 0.50 | 0.39 | 0.63 |
| Null | — | 1.10 | 0.83 | 0.99 | 1.00 |
| Memory | 0.39 | 1.19 | 0.89 | 1.21 | 1.10 |

(COR, MAD, and MSE) from RRHCV are shown in appendix B (Fig. B1). The ANN model significantly outperforms all models except for the average model in terms of RMSE, MSE, and MAE, and it outperforms all except for GLMs, GAMs, and BART in terms of MAD. Every model outperforms the null and memory model in terms of all error measures. Overall, the best-performing model is the ANN model followed in order by the GLMs, the GAMs, the BART model, the other tree-based models, and finally MARS and CART without bagging, which was statistically indistinguishable from the null model. The memory model—based on persistence of precipitation from the previous year—performed worse than the null model (see Fig. 5). The poor performance of the memory model is an indicator that low-frequency variability alone is not a basis for skillful predictions of Sahel precipitation.

As the ANN was the best-performing model type, we further explore the sensitivity of ANN performance to the choice of predictor variables. Figure 6 summarizes these model comparisons using the same metrics as in Fig. 5. Results of additional hypothesis tests of these ANN models are provided in appendix B (Fig. B2). All model differences are significant at the 95% confidence level. All PC-based ANN models (ANN1–9) outperform the ANN that uses climate indices rather than PCs of indices (ANN0), again demonstrating that the PC transformation adds predictive value relative to models based on selected raw indices. Figure 6 also shows that PC-based ANN models that make use of nine or more predictor PCs (i.e., ANN1–7) are statistically indistinguishable from one another according to standard performance estimates, such that the model that uses nine PCs (ANN7) can be identified as the simplest high-performing ANN.

### b. Goodness of fit

It is important to note that we have ranked the model performance on the basis of predictive skill, not goodness of fit. Goodness-of-fit results are summarized in Table 3, which lists COR, MAD, MAE, MSE, and RMSE between the observed rainfall anomalies and the fitted values from each model. It is clear that the goodness of fit is not necessarily an indication for the predictive accuracy, as a number of models that performed poorly in terms of prediction show high correlation and low error estimates for in-sample evaluation. The tree-based models tend to fit the data well, though the ANN model also fits fairly well. The observed inconsistency is not surprising, since goodness-of-fit measures such as correlation coefficient with observations do not penalize models for overfitting the data, though overfitting can negatively affect predictive performance.

In summary, the overall performance of models confirms that the nine-PC ANN model is the best in terms of predictive accuracy, and that it also provides relatively strong goodness of fit. The nine-PC GLM is considered to be the preferred linear model on account of its simplicity, and the BART model is the best-performing tree-based model. In section 4c, we consider the results of the best-performing nonlinear (ANN) and linear (GLM) models, in order to understand the statistical and underlying physical bases for their strong predictive performance.

### c. Variable importance and influence

To understand the physical basis for model predictions, it is first necessary to examine the meaning of the principle components used to generate predictive models. Transforming standard climate indices into PC space has the advantage of maximizing information content while eliminating multicollinearity, but interpreting results of models constructed from PCs does require that each PC can be related meaningfully to the original predictor variables and to coherent patterns of SST.

Figures 7a–c show raw and detrended time series plots for the three leading principal components, PC1, PC2, and PC3, respectively, while Table 4 lists their correlations with the original variables and Fig. 8 shows correlation patterns between each of PC1–3 with MLOST and selected atmospheric fields from the NCEP–NCAR reanalysis. The three leading PCs together represent approximately 70% of the total variance. The first principal component (PC1), which explains approximately

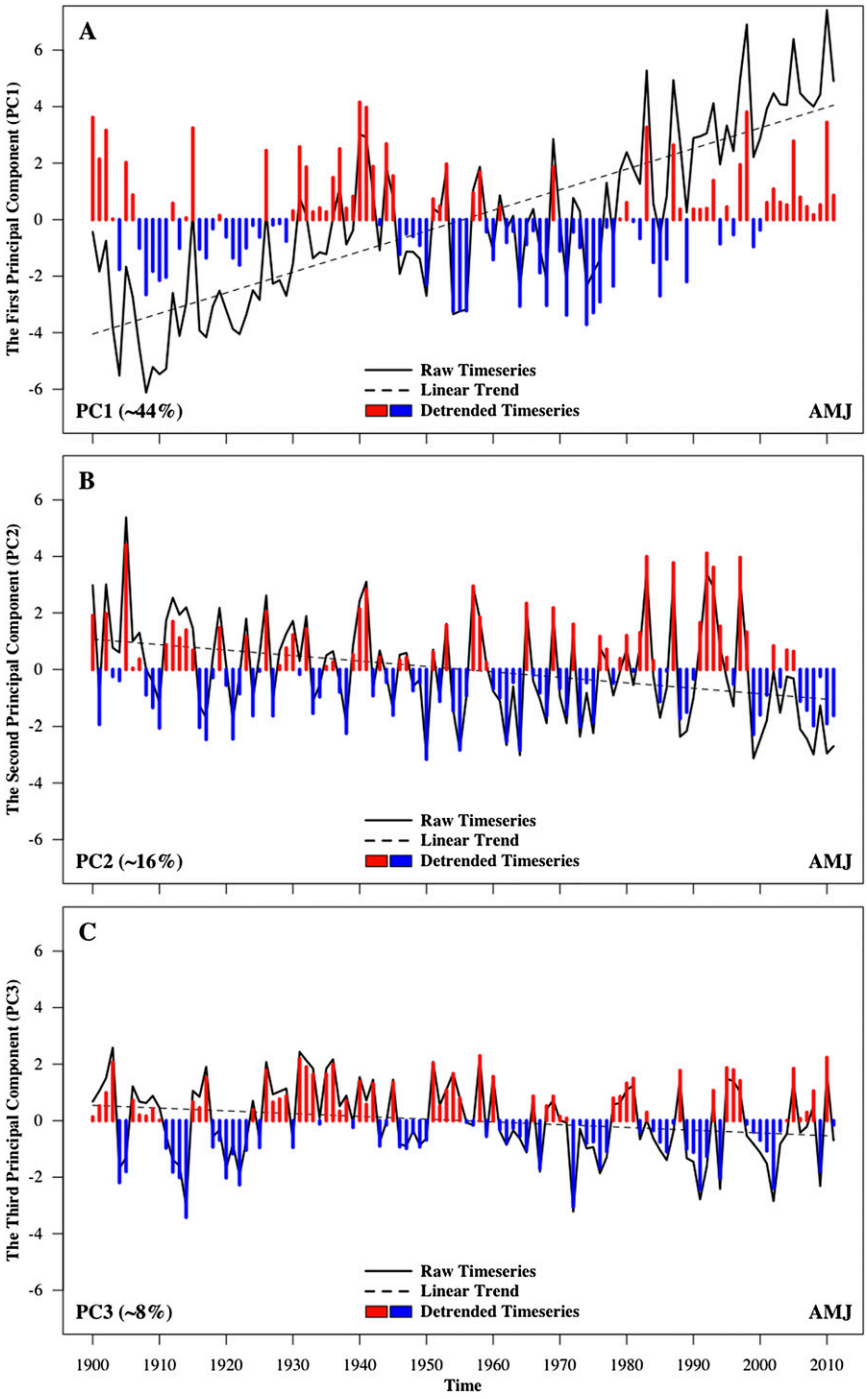


FIG. 7. Raw and detrended time series plots for the leading principal components (PC1–3). The three leading PCs explain ~70% of the total variance.

44% of the total variance, is loaded by almost all predictor variables (Table 4) and is generally representative of synchronous warming and cooling of the tropical oceans (Fig. 8a), with a particularly strong signal in the eastern Pacific Ocean. Not surprisingly, the time series of this PC (Fig. 7a) shows both interannual variability and a twentieth-century trend associated with warming SST and SATA. This general warming of the tropical oceans has been linked to Sahel drying in previous studies (e.g., Nicholson and Selato 2000; Giannini et al.

TABLE 4. Correlation coefficients of the leading three PCs with the original variables (large-scale climate indices) using both raw and detrended time series. All correlations are statistically significant at the 95% confidence level; correlations in boldface are statistically significant at the 99% confidence level.

| | Raw time series | | | Detrended time series | | |
|---|---|---|---|---|---|---|
| Variable | PC1 | PC2 | PC3 | PC1 | PC2 | PC3 |
| Sahel_Precip | **−0.38** | −0.21 | **0.29** | | **−0.40** | 0.22 |
| AMO | **0.81** | **−0.30** | **0.37** | **0.66** | | **0.68** |
| NAO | | | **−0.69** | −0.19 | | **−0.71** |
| SST_MDR | **0.79** | | **0.45** | **0.69** | | **0.71** |
| TNA | **0.78** | | **0.47** | **0.69** | | **0.74** |
| TSA | **0.61** | **−0.48** | | | **−0.34** | |
| Niño-1.2 | **0.56** | **0.57** | | **0.60** | **0.74** | |
| Niño-3 | **0.63** | **0.67** | | **0.71** | **0.85** | |
| Niño-3.4 | **0.63** | **0.66** | | **0.70** | **0.84** | |
| Niño-4 | **0.94** | | | **0.87** | **0.62** | |
| SLP_Darwin | **0.52** | **0.54** | | **0.57** | **0.68** | |
| SLP_Tahiti | | **−0.47** | 0.24 | | **−0.49** | **0.25** |
| NP | −0.23 | **−0.50** | **−0.35** | **−0.48** | **−0.50** | **−0.34** |
| PDO | **0.36** | **0.50** | **0.40** | **0.59** | **0.53** | **0.42** |
| SIOD_E | **0.61** | | **−0.28** | **0.41** | 0.23 | −0.19 |
| SIOD_W | **0.67** | | | **0.37** | | |
| SST_Med | **0.50** | **−0.51** | | | **−0.40** | |
| SATA_LNH | **0.79** | **−0.37** | | **0.47** | | |
| SATA_LSH | **0.85** | | −0.22 | **0.63** | **0.39** | |
| SATA_ONH | **0.92** | **−0.25** | | **0.76** | | **0.33** |
| SATA_OSH | **0.91** | −0.21 | −0.20 | **0.76** | **0.32** | |
| PC1 | | — | | **0.62** | **0.45** | **0.31** |
| PC2 | | — | | **0.30** | **0.93** | |
| PC3 | | — | | 0.20 | | **0.97** |

2008). Proposed mechanisms include the influence of Atlantic Ocean temperature gradients on moisture supply (e.g., Hoerling et al. 2006), and the stabilizing influence that higher Indo-Pacific SSTs has on the tropical troposphere over Africa (e.g., Giannini et al. 2003; Herceg et al. 2007).While a full exploration of these mechanisms is beyond the scope of the present paper, we do see that PC1 is associated with pan-tropical tropospheric warming stabilization patterns (Fig. 8b) that are consistent with earlier studies, pointing to a physical mechanism that potentially links the PC to Sahel precipitation variability.

The second PC (PC2), which explains about 16% of the total variance, is mostly loaded by ENSO (Table 4) and correlates with tropical Pacific SST variability (Fig. 8c). The PC time series is notable for its negative values over the past decade, a period during which the Sahel has seen a relative recovery from the dry period in the late twentieth century (Fig. 7b). ENSO is thought to influence the Sahel through both an atmospheric teleconnection and through its link to Indian Ocean temperatures (Nicholson 2013). PC2 has the potential to capture the atmospheric teleconnection, as it is associated with a global warming of the tropical troposphere that can lead to suppressed precipitation over Africa and to a weakening of winds in the region of the tropical easterly jet [TEJ; Fig. 8d—note that the positive correlation in this figure is due to weakening of easterly winds (i.e., reduced negative wind) in the TEJ region during positive ENSO events], which is consistent with known relationships between TEJ and ENSO and has been associated with dry years in the Sahel (Grist and Nicholson 2001).

The third PC (PC3), which explains about 9% of the total variance, shows a strong association with Northern Hemisphere Atlantic SST (Fig. 8e). Northern Hemisphere warming of the Atlantic has been associated with increases in Sahel precipitation, with proposed mechanisms including enhanced moisture flux into the region from the south (Biasutti et al. 2008) and stronger easterlies carrying moisture into the Sahel (Neupane and Cook 2013). Consistent with this proposed link, PC3 is associated with enhanced lower-tropospheric humidity in portions of the Sahel (Fig. 8f).

#### 1) ARTIFICIAL NEURAL NETWORK

The best-performing model in the multimodel comparison is an ANN that includes one input layer with nine units for the nine PCs selected from cross validation, a single hidden layer with three units, and an output layer with one unit for the response variable, Sahel rainfall anomalies. The prediction of this model is the average of 10 repeated neural networks (Fig. 4) with different random number seeds. It was tuned inside the cross-validation analysis for hidden layer size, weight decay, and bagging for each repeat. Rigorous variable selection analysis demonstrated that this nine-PC ANN model (PC1–8 and PC13) was statistically indistinguishable from the full 20-PC ANN and that it also outperformed all other model types (Figs. 5, 6, B1, and B2). As a simpler model is preferable for purposes of interpretation, we use the nine-PC ANN for all subsequent analyses. All units in the network are connected as shown in Fig. 4. To fit the full and nine-PC ANN models, 111 and 34 weights are to be calculated, respectively. This number of weights is large compared to the sample size (112 yr, 1900–2011), and has the potential to produce an overfitting problem. However, penalizing the weights using unit weight decay regularizes the model and prevents overfitting. The maximum absolute value of the weights is about 1.8.

Figure 9 shows scatterplots for the fitted and cross-validated predictions of the final ANN model against the summer rainfall anomalies over the Sahel, Q–Q plots for the residuals, and the fitted/predicted versus observed rainfall anomalies. Residuals are normally distributed.

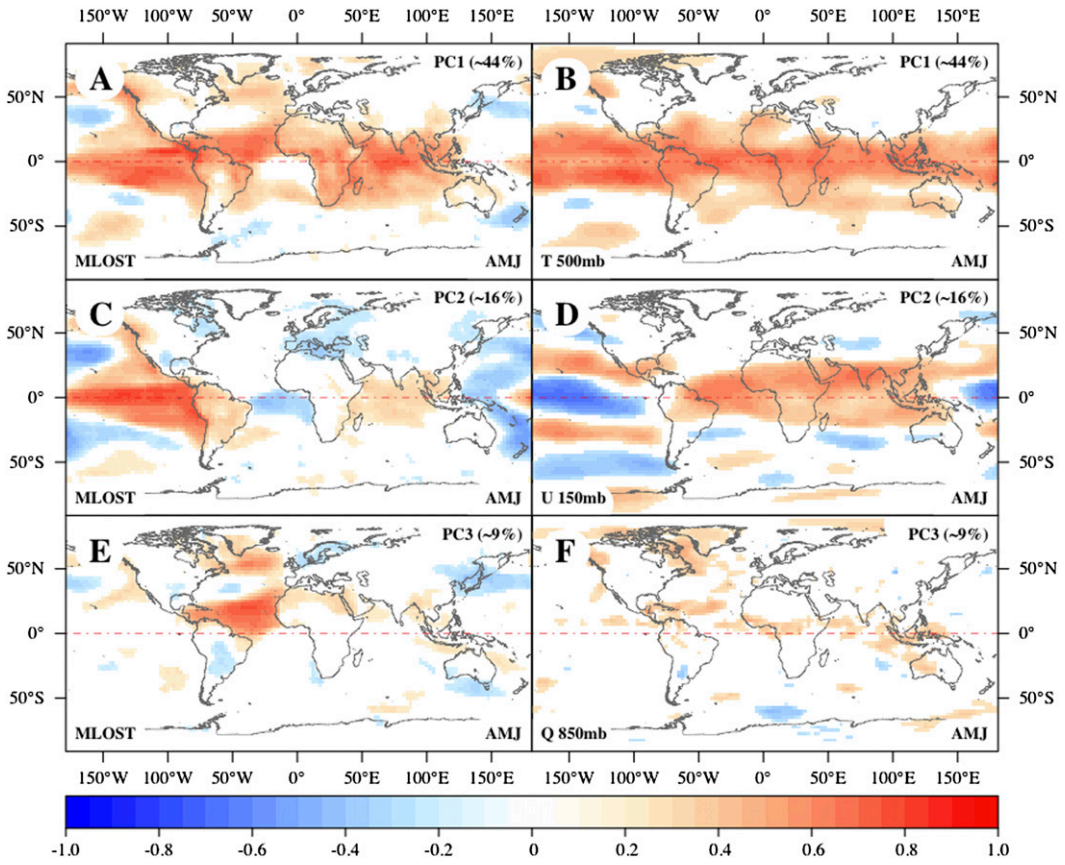


FIG. 8. Correlation patterns of the leading principal components (PC1–3) with spring (AMJ) (a),(c),(e) MLOST data for the period 1900–2011 and (b),(d),(f) NCEP–NCAR reanalysis atmospheric fields for the period 1948–2011. Shown are the MLOST correlations with (a) PC1, (c) PC2, and (e) PC3. Shown are correlations of (b) PC1, (d) PC2, and (f) PC3 with temperature at 500 mb, zonal wind at 150 mb (TEJ), and specific humidity at 850 mb (1 mb = 1 hPa). All correlations are statistically significant at the 95% confidence level.

The scatterplot shows the skill of that model in predicting Sahel rainfall anomalies. The correlation coefficient between the fitted (predicted) and observed rainfall anomalies is 0.82 (0.70).

Partial dependence plots for the final ANN model are shown in Fig. 10. It is clear that some of the covariates have nonlinear effects on the response variable while some other variables have linear effects. This confirms the ability of ANNs to uncover nonlinear interactions. An increase in the first PC (PC1) is associated with a nonlinear decrease in rainfall, with maximum sensitivity for PC1 values near zero (from −2 to 2). This is consistent with Sahel drying under pan-tropical tropospheric warming. The second PC (PC2) also has a drying effect, consistent with known ENSO impacts on Sahel rainfall. The third PC (PC3) has the opposite effect to PC2, since Northern Hemisphere Atlantic warming is associated with increased Sahelian precipitation, likely because of an increase in the moisture supply to the region.

Beyond these relatively obvious characteristics of the three leading PCs, the ANN based on nine PCs draws skill from a number of other mechanisms that link input indices to Sahel rainfall. Most notably, PC5—which captures an SST gradient between the North and South Atlantic (warmer north and cooler south)—is associated with strengthened low-level westerlies over the Gulf of Guinea and the western Sahel, which transport oceanic moisture into the region. Other higher-order PCs capture the influence of the subtropical Indian Ocean dipole (SIOD_W − SOID_E), Mediterranean SSTs, and South Atlantic SSTs (see loading matrix in appendix B), each of which has been associated with Sahel precipitation in previous studies (e.g., Giannini et al. 2003; Lu 2009; Lu and Delworth 2005; Nicholson 2013). The fact that these influential higher-order PCs are associated with known physical mechanisms of Sahel precipitation variability enhances our confidence in the model predictions.

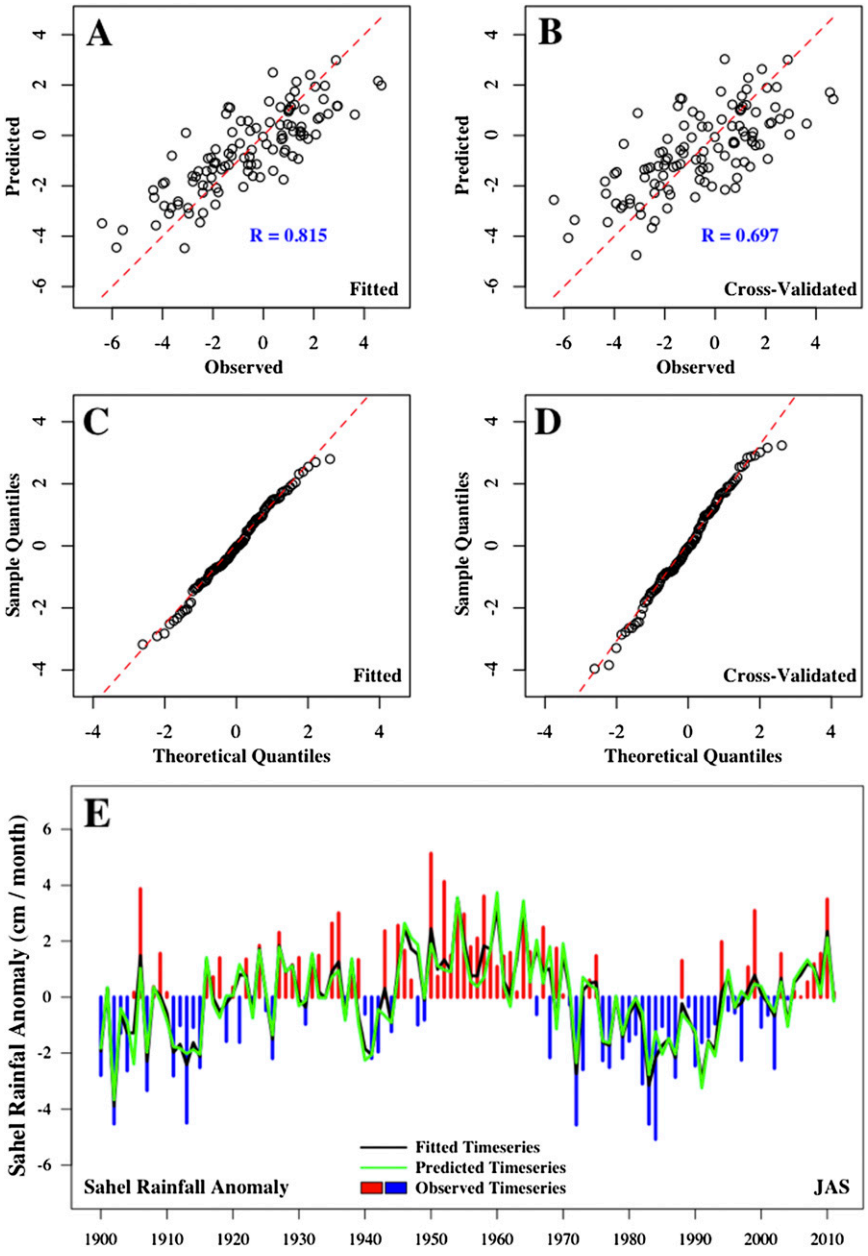


FIG. 9. Descriptive performance plots for the final ANN model. (a),(b) Scatterplots and (c),(d) Q–Q plots for the residuals of fitted and predicted rainfall. (e) The full time series of fitted, predicted, and observed Sahel rainfall anomalies.

### 2) GENERALIZED LINEAR MODEL

The second best-performing linear model is the nine-PC GLM. It predicts the seasonal rainfall anomalies using nine PCs (PC1–8 and PC13) from the 20 large-scale SST and SATA predictors described in Table 1. The intercept is not statistically significant and it was removed from the beginning and within the cross-validation analysis. The reason for this is that the response variable, Sahel rainfall anomalies, and predictor variables are centered (zero mean). Figure 11 shows the partial dependence and variable importance plots for the GLM. All coefficients are statistically significant at $\alpha = 0.05$. The residual deviance of this model is about 54.4 on 103 degrees of freedom. The residuals Q–Q plot shows that they are normally distributed. The residual mean is almost constant around zero for all values of the response variable.

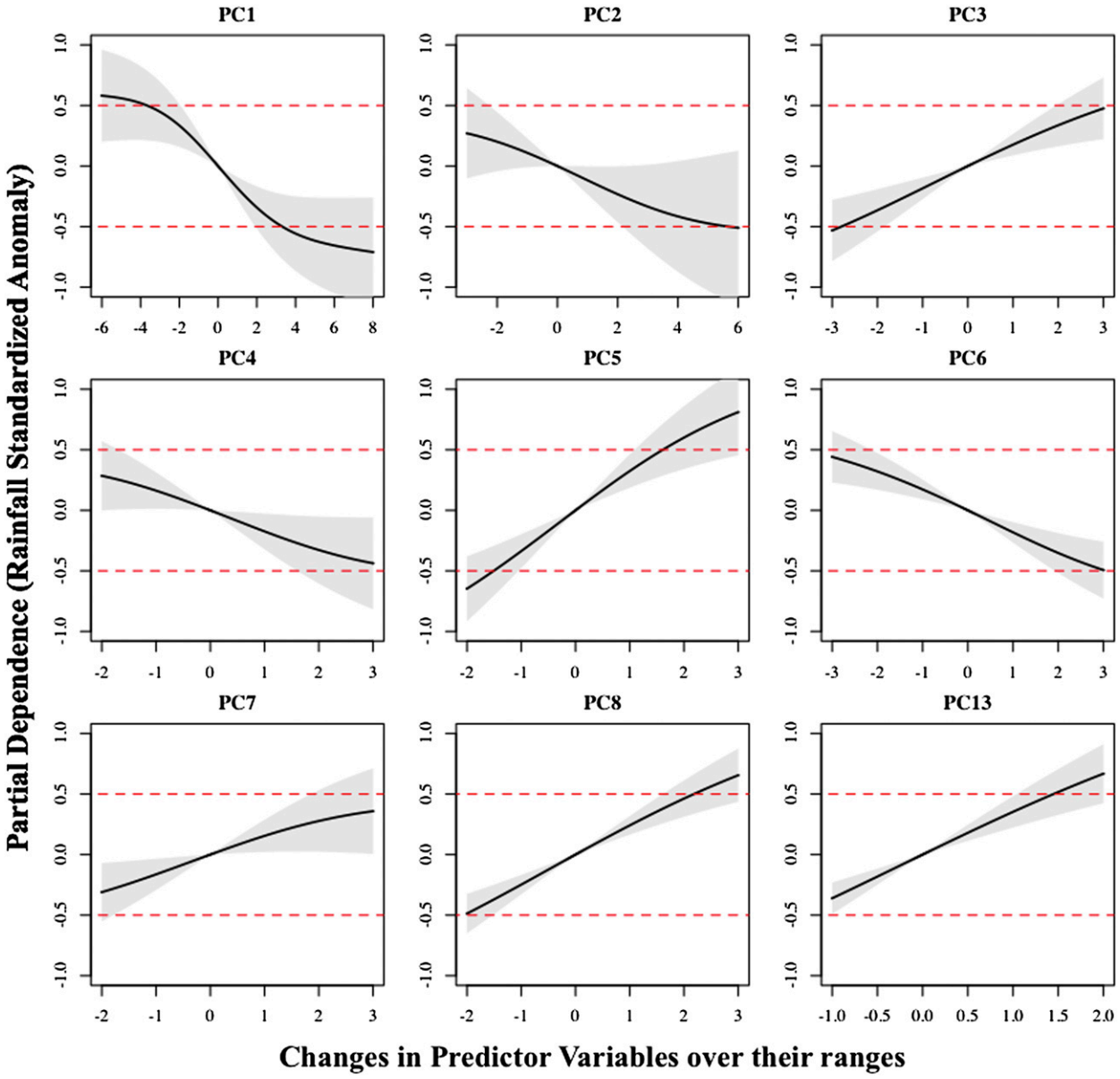


FIG. 10. Partial dependence plots for the ANN model. Gray shading represents the range of all possible effects of PCs on the response variable. The dashed red lines represent the partial dependence of ±0.5 standard deviation.

All variables in the GLM have similar effects as they do in the ANN model, but without nonlinearities, as shown in the ANN versus GLM bivariate partial dependence plot of PC1 and PC2 (Fig. 12). Similar physical insights can be drawn from GLM as from the ANN: negative rainfall anomalies in the Sahel are associated with ENSO, positive anomalies in the tropical oceans, and negative anomalies in North Atlantic, while the wet events are associated with the SIOD, tropical South Atlantic, and Mediterranean. Notably, the skill of ANN in covering the nonlinear interactions is advantageous and provides more details for the proposed mechanisms. For instance, the partial dependence of Sahelian rainfall on ENSO in Fig. 10 suggests a smaller effect of La Niña relative to that of El Niño, in contrast to its constant slope in Fig. 11, while Fig. 12 suggests nonlinear interaction between ENSO and tropospheric warming that cannot be recorded by the GLM.

### *d. Persistence*

The shift in the interannual persistence of rainfall anomalies between the first half of the twentieth century

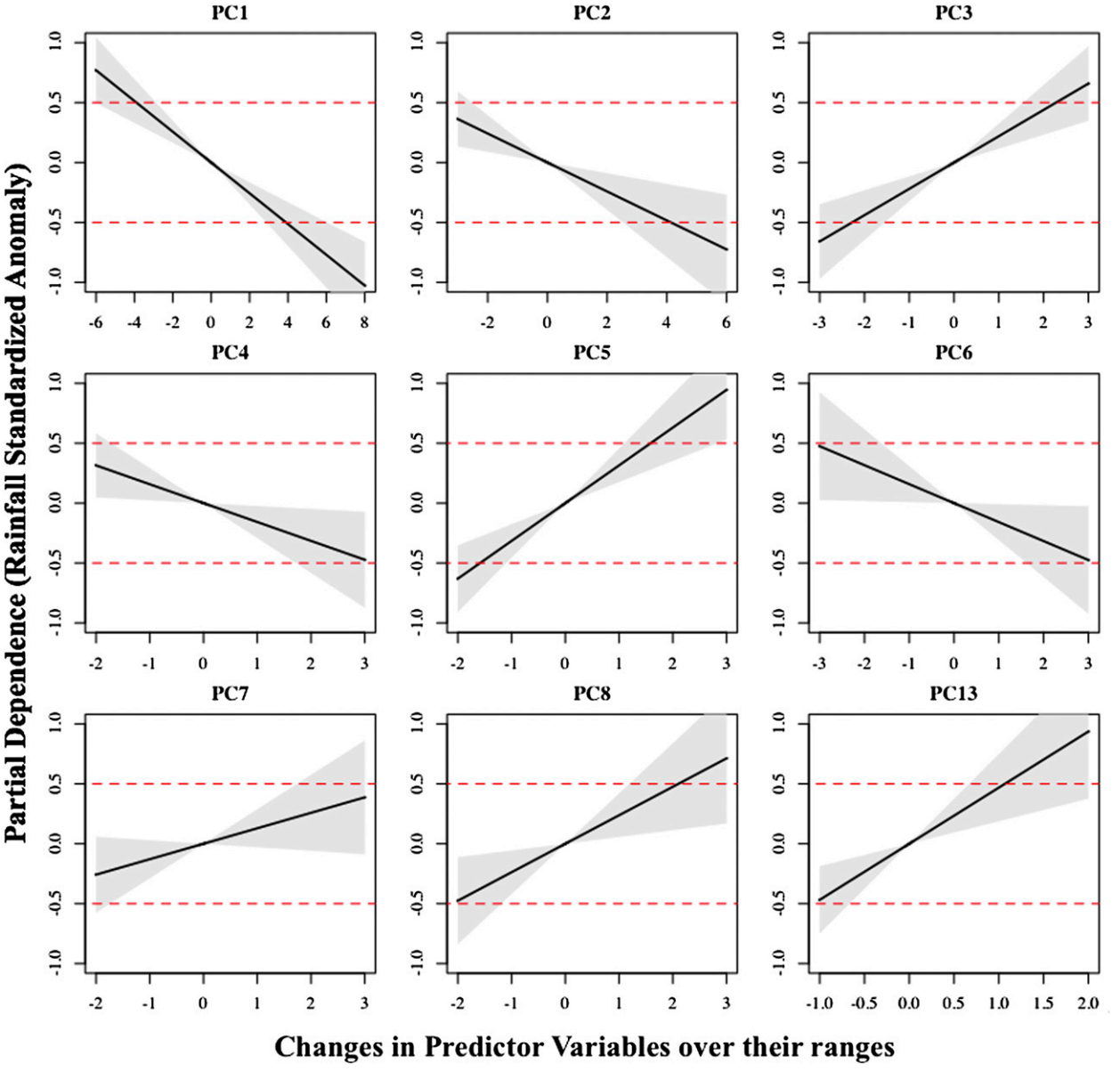


FIG. 11. Partial dependence plots for the selected Gaussian GLM. Gray shading indicates the confidence intervals for the model parameters. The dashed red lines represent the partial dependence of ±0.5 standard deviation.

and the second half is one of the most dramatic features in the modern precipitation record for the Sahel. Predictive models that are calibrated exclusively for the post-1950 period—including several tested in our own research but not included in this paper—draw some skill from this year-to-year persistence. The fact that the memory model underperformed the null model in our analysis, and the fact that including the previous year's precipitation anomaly as a predictor in more complex models did not significantly improve skill, is because we have constructed models that address the full precipitation record, 1900–2011.

There are both advantages and disadvantages to using the entire 1900–2011 record in our predictive models. The primary advantage is that the resulting models capture the predictors responsible for the shift in persistence between pre- and post-1950 regimes, such that the models do not assume the continuing stability of post-1950 conditions. In this sense a model trained to a longer baseline period may be more robust. The primary disadvantage in using the full record is that models trained to a shorter, recent baseline can achieve higher skill scores, as variability in the post-1950 period is relatively simple and persistent relative to the full historical

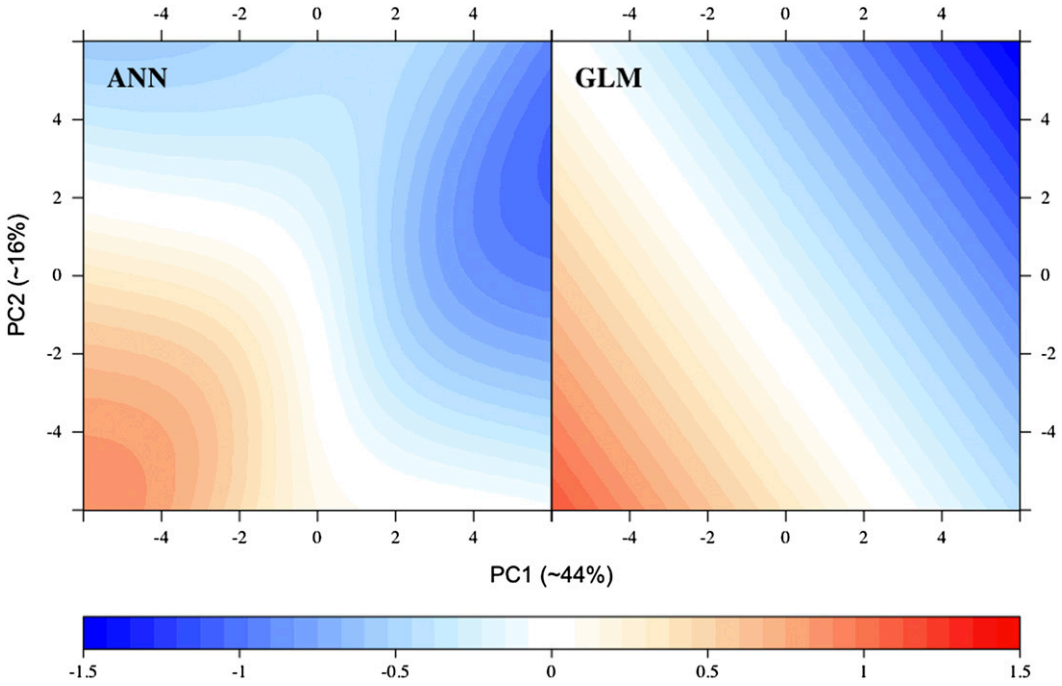


FIG. 12. Two-dimensional partial dependence of the Sahelian rainfall anomaly as a function of changes in both PC1 and PC2. ANN captures the nonlinear interactions among variables, while GLM approximates all relations in linear forms.

record. The choice of baseline period, then, depends on the objective of the modeling exercise.

Additionally, a comparison of models trained to different baseline periods can offer insights into the nonstationarity in large-scale predictors and their relationship to Sahel precipitation over the available historical record. For example, we found that ANN and GLM trained to the 1900–50 period were similar to models trained for 1951–2011 in most respects, but that the relationship between Sahel precipitation and certain subtropical and higher-latitude SST indices reversed between the two periods: the influence of PCs loaded by North Atlantic, North Pacific, and Mediterranean SSTs, and the PDO changed in magnitude and sign of variable influence for models trained to 1900–50 versus those trained to 1951–2011, suggesting that connections between Sahel precipitation and extratropical variability might have changed between the first and second halves of the twentieth century. These changes are the subject of ongoing research.

## 5. Conclusions

We applied 10 different statistical modeling approaches to the problem of seasonal precipitation prediction for the Sahel. A consistent set of large-scale predictors was used for all models, and it was found that the artificial neural network model offered the best performance in terms of predictive accuracy while providing reasonably strong goodness of fit. Moreover, the ANN is able to uncover nonlinear interactions; some variables have nonlinear effects on the rainfall anomalies

TABLE A1. Statistical models applied in this study. They include likelihood-based (GLMs and GAMs), MARS, tree-based (CARTs, BART, and RF), and neural network (ANN) approaches, along with model average, null, and memory models. The boldface rows are for the final best-performing models: ANN and GLM with nine PCs (PC1–8 and PC13).

| Model | Description |
|---|---|
| **GLM** | **Full-covariate generalized linear model** |
| SGLM | Selected generalized linear model based on stepwise selection |
| GAM | Full-covariate generalized additive model |
| SGAM | Selected generalized additive model based on penalized terms |
| MARS | Multivariate adaptive regression spline |
| CART | Classification and regression trees model |
| BCART | Bagged classification and regression trees model |
| BART | Bayesian additive regression trees model |
| RF | Random forest model |
| **ANN** | **Artificial neural network** |
| Avg | Prediction is the mean of the predictions of all models |
| Null | Prediction is the mean of the response variable in the training data |
| Memory | Model using last year's rainfall as this year's prediction |

TABLE A2. The R packages used for model development and visualization in this study. The boldface row is for the nnet package that was utilized to fit the best-performing ANN model with different sets of covariates.

| Package | Description |
|---|---|
| BayesTree | Implementation of BART models |
| caret | Training and plotting statistical models |
| earth | Implementation of MARS models |
| HH | Statistical analysis and data display |
| ipred | Improved predictive models by indirect classification and bagging |
| mgcv | Routines for GAMs and other generalized ridge regressions |
| **nnet** | **Feed-forward neural networks models** |
| party | Computational toolbox for recursive partitioning |
| randomForest | Breiman and Cutler's random forests for classification and regression |
| tree | Classification and regression trees |

TABLE A3. Top-performing combinations of predictor variables selected for different-size models inside cross-validation analyses. Subset 0 is a model based on selected climate indices and is used as a reference to test the robustness of PC-based models. The boldface row represents the set of covariates used in the final selected models.

| Subset | Size | Candidate variables |
|---|---|---|
| 0 | 9 | SATA, ENSO, NAI, NP, PDO, SIOD, SST_Med, and TSA |
| 1 | 20 | PC1–20 |
| 2 | 14 | PC1–8, PC10, PC11, PC13, PC14, PC15, and PC16 |
| 3 | 13 | PC1–11, PC13, and PC16 |
| 4 | 12 | PC1–8, PC10, PC11, PC13, and PC16 |
| 5 | 11 | PC1–8, PC11, PC13, and PC16 |
| 6 | 10 | PC1–8, PC13, and PC16 |
| **7** | **9** | **PC1–8, and PC13** |
| 8 | 8 | PC1–8 |
| 9 | 7 | PC1–7 |

response, while other variables have linear or even constant/zero effects. The tuning of the parameters for this model, and all other models, is an essential step toward developing a predictive model. It was found that the ''optimal'' tuning parameters for the full ANN model in this application included a three-unit (size = 3) single hidden layer with weights penalized using unit decay. However, this should be examined for different data and/or regions.

The application of machine-learning algorithms such as ANN to climate prediction should be approached cautiously and with some skepticism. These algorithms are complex and, rightly or wrongly, are often viewed as ''black boxes'' that obscure a physically based interpretation of the results. This study demonstrates that when properly applied, ANN can provide skillful predictions without overfitting and that an ANN can aid in identifying and explaining predictive patterns that include nonlinear relationships between predictors and response. It is noteworthy, for example, that the first principal component (PC1), which includes a significant tropical ocean SST loading, has a nonlinear impact on Sahel precipitation that tails off at extreme values of the PC. Furthermore, El Niño and La Niña impacts were identified differently in the nonlinear partial dependence plot of PC2, in contrast to the linear response suggested by GLMs. Linear regression models cannot capture these nonlinearities. At the same time, by examining partial dependence plots and comparing ANN with the simpler stepwise-selected GLM, we were able to confirm that the ANN was constructed on the basis of meaningful and robust relationships between Sahel rainfall and large-scale predictors.

Another conclusion from this study is that it is critical to distinguish between measures of predictive skill and measures of goodness of fit when developing seasonal prediction models. This may be an obvious point from a statistical perspective, but it is not uncommon to see models in the climate prediction literature that are optimized for goodness of fit when the applications goal of the model is to generate skillful predictions. Here, we find that ANN provides significant benefit in terms of predictive skill (out-of-sample accuracy from cross-validation analysis), even when the model did not particularly distinguish itself in terms of goodness of fit (in-sample accuracy). It has more reliable performance where the models that provide better fit have artificial skills as they adapt themselves to the data (Von Storch 1999; Von Storch and Zwiers 2001).

The results of this study have direct relevance to the development of operational seasonal prediction systems for the Sahel, and the model development process that we employed—applying well-known climate indices, addressing multicollinearity through PCA of these indices, comparing multiple model structures, and evaluating the physical basis of predictions through partial dependence plots and intermodel comparisons—could be applied to other regions as well. Through this process, it is possible to build confidence in the application of nonlinear statistical algorithms to problems of climate prediction, thus maximizing the value of these techniques to predictive skill and to a physically based explanation of climate variability.

*Acknowledgments.* This study was supported by the Department of Earth and Planetary Sciences, The Johns Hopkins University, and NASA Applied Sciences Grant NNX09AT61G.

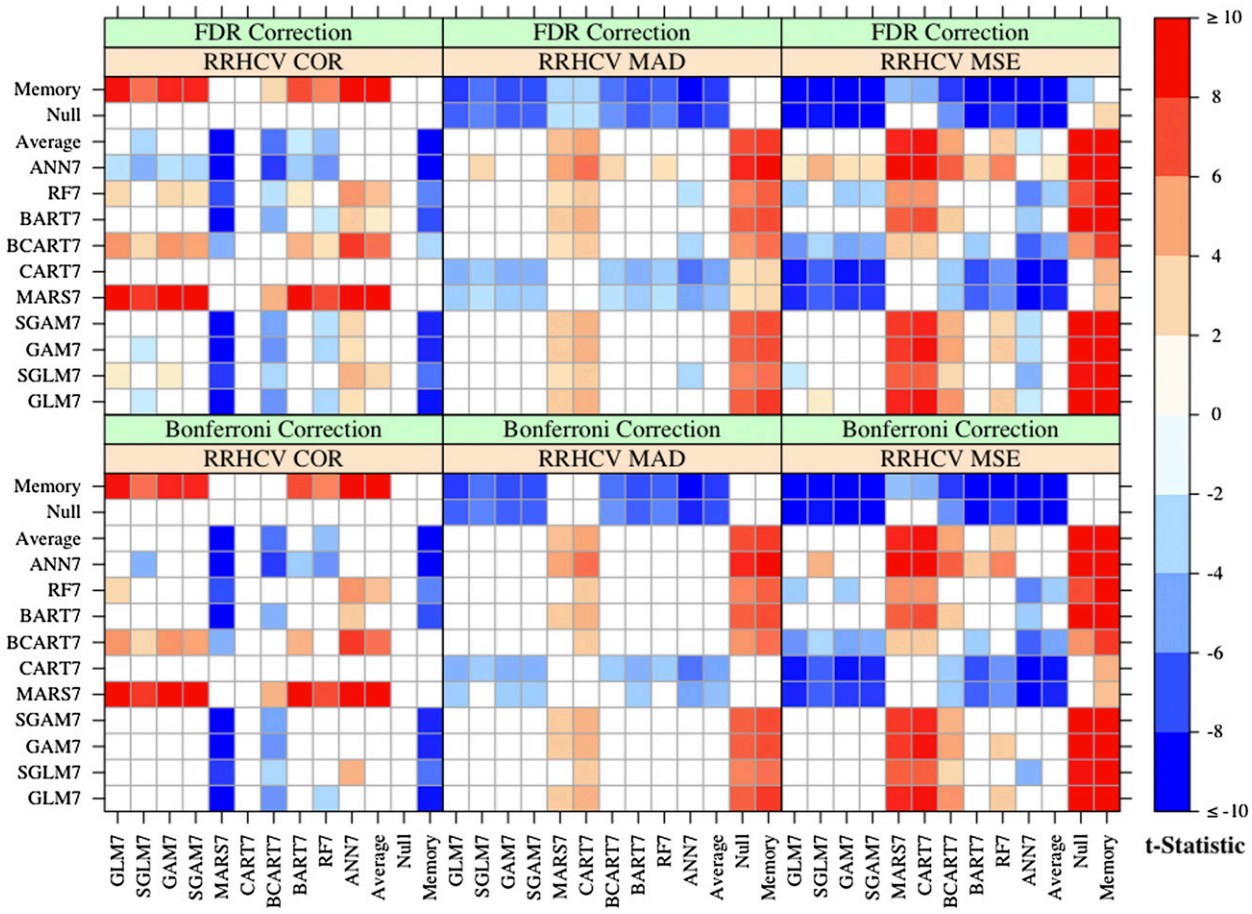


FIG. B1. Model comparisons for the best-performing models of each type, as in Fig. 5, showing results of additional cross-validation statistical tests: out-of-sample COR, MAD, and MSE measures from RRHCV.

## APPENDIX A

### Statistical Models

Ten different model approaches listed in Table A1 were applied and compared with each other and against three reference models based on the average of all models, null/climatology model, and a memory model. These models are from different categories: likelihood based (GLMs and GAMs), MARS, tree-based (CARTs, BART, and RF), and neural networks (ANNs). The last three rows are for reference average, null, and memory models. A set of R packages (Table A2) was utilized for preprocessing of the data, fitting the models, and visualization of the results. The BayesTree package (Chipman et al. 2010) was used for fitting BART model(s). The caret package (Kuhn 2008) was used for preprocessing and training. The earth package (Friedman 1991) was used for fitting MARS models. The HH package (Heiberger 2013) was used mainly for visualizations. The ipred package (Peters et al. 2002) was used for bagging by the caret package. The mgcv (Wood 2001) package was utilized in fitting GAMs. The nnet (Venables et al. 1994) package was utilized to apply a feed-forward neural network approach in ANN models. The party (Hothorn et al. 2010) and tree (Ripley 2013) packages were used for fitting the trees. The randomForest package was used to apply Breiman and Cutler's random forests for regression.

To test the robustness of our approach, large numbers of models, tuning parameters, and variable selection methods were tested. Different trials were performed by applying PCA on each set of variables and keeping only the leading PC, or choosing one by one; all trials were assessed based on both predictive skill (standard out-of-sample error measures and correlation coefficient from cross-validation analysis) and goodness of fit (in-sample correlations and residuals). One ENSO index was selected from the set of SST indices (Niño-1.2, Niño-3, Niño-3.4, or Niño-4) and the Southern Oscillation index (SOI). Similarly, one index was selected for the equatorial North Atlantic SST from AMO, NAO, SST_MDR,

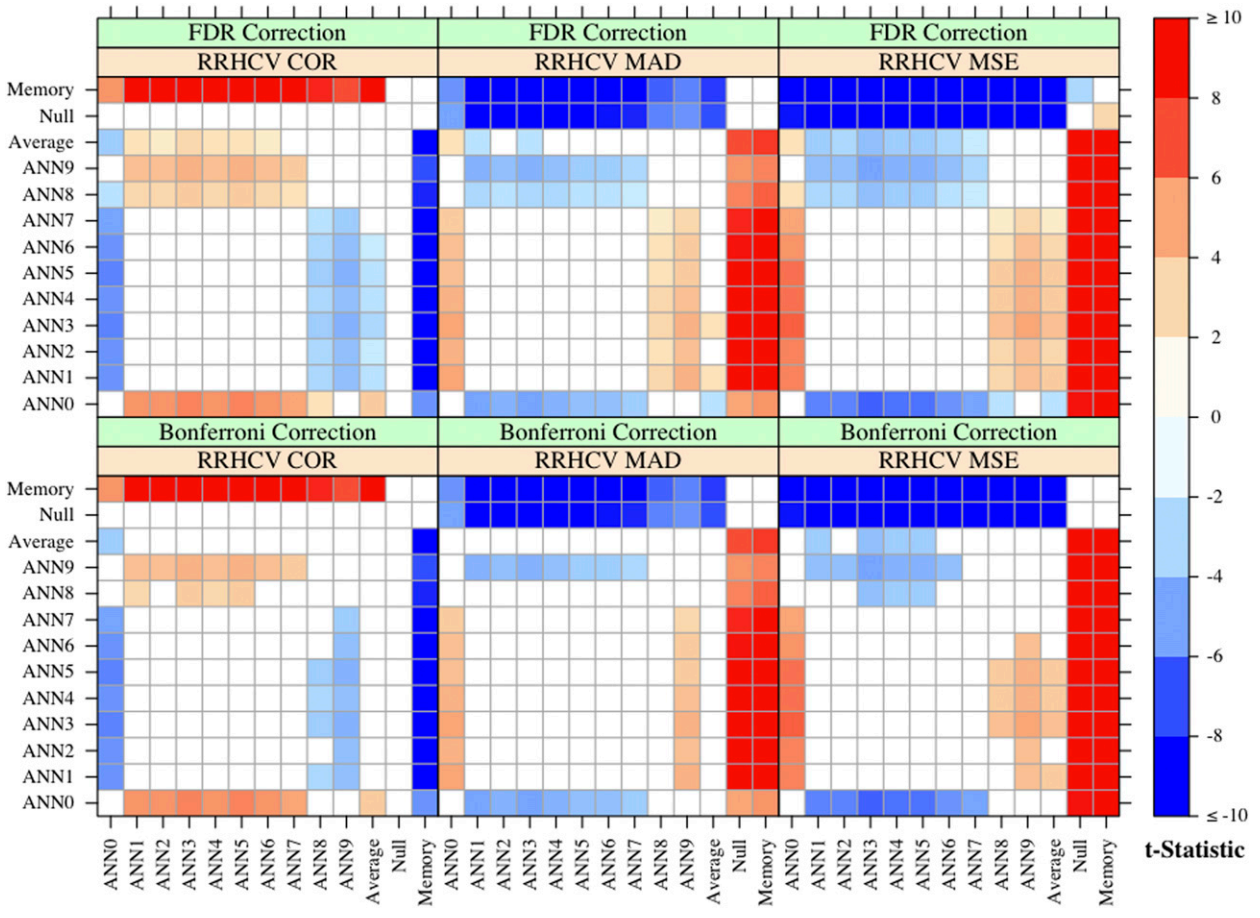


FIG. B2. Comparisons between ANN models with different sets of variables. Models included in the comparisons are the same as in Fig. 6 and performance parameters are the same as in Fig. B1.

and TNA. The subtropical Indian Ocean dipole was considered to be the difference between SIOD_E and SIOD_W instead of using the two variables. One index was tested for the Pacific Ocean from NP and PDO. Finally, one SATA index was used. Another approach of variable selection was examined based on the regression patterns between Sahelian rainfall index and global SST (Fig. 2). However, it was found that the PCA-based models outperform all models that are based on different combinations of physically selected sets of original variables in terms of their predictive skill. The best sets of covariates that were selected for the final analysis are listed in Table A3. It was somewhat easier to interpret the physical meaning of simpler models that are based on well-known climate indices, but the primary objective of this study is to improve the operational rainfall predictions for the Sahel. For this reason we have adopted the approach that provided the best predictions, even though interpretation of the PCs is not as straightforward as interpretation of models based on conventional indices.

In all cases, the PCA model presented in the paper provided better predictive skill and strong generalizability to different baseline periods relative to other possible approaches. Unfortunately, most index-based models, which are much easier to interpret, were statistically indistinguishable when compared with the null model through both leave-one-out and random-holdout cross-validation analyses, and the best-performing of these models was significantly less accurate than the ANN PCA-based model.

## APPENDIX B

### Supplementary Results

The results of hypothesis tests presented in section 4 were selected appropriately to support the core findings of this work. Here, we provide supplementary results for the other metrics for completeness. Figure B1 shows the results of *t* tests with Bonferroni and FDR corrections

for comparing COR, MAD, and MSE based on the repeated random-holdout cross-validation method. Figure B2 shows model comparisons for the ANN model with different sets of variables. It summarizes the cross-validation results in terms of out-of-sample COR and MAD and MSE measures from RRHCV. These supplementary results confirm the conclusions drawn from results presented in the main body of the paper. The final nine-PC (PC1–8 and PC13) model is the simplest best-performing ANN model. GLM using the same covariates is the next best model in performance, but it lacks the ability to capture nonlinear interactions between variables that ANN has. Moreover, ANN has better accuracy and fit when compared with linear GLMs.

## REFERENCES


Arlot, S., and A. Celisse, 2010: A survey of cross-validation procedures for model selection. *Stat. Surv.,* **4,** 40–79, doi:10.1214/09-SS054.

Benjamini, Y., and Y. Hochberg, 1995: Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.,* **57B,** 289–300.

Biasutti, M., I. Held, A. Sobel, and A. Giannini, 2008: SST forcings and Sahel rainfall variability in simulations of the twentieth and twenty-first centuries. *J. Climate,* **21,** 3471–3486, doi:10.1175/2007JCLI1896.1.

Breiman, L., 1984: *Classification and Regression Trees.* Chapman & Hall/CRC, 358 pp.

——, 2001: Random forests. *Mach. Learn.,* **45,** 5–32, doi:10.1023/A:1010933404324.

Brooks, N., 2004: Drought in the African Sahel: Long-term perspectives and future prospects. Tyndall Working Paper 61, Tyndall Climate Centre for Climate Change Research, 31 pp. [Available online at www.tyndall.ac.uk.]

Cameron, A. C., and P. K. Trivedi, 1998: *Regression Analysis of Count Data. Econometric Society Monogr.,* Vol. 30, Cambridge University Press, 436 pp.

Chipman, H. A., E. I. George, and R. E. McCulloch, 2010: BART: Bayesian additive regression trees. *Ann. Appl. Stat.,* **4,** 266–298, doi:10.1214/09-AOAS285.

Collier, R., 1994: An historical overview of natural language processing systems that learn. *Artif. Intell. Rev.,* **8,** 17–54, doi:10.1007/BF00851349.

Dezfuli, A. K., and S. E. Nicholson, 2011: A note on long-term variations of the African easterly jet. *Int. J. Climatol.,* **31,** 2049–2054, doi:10.1002/joc.2209.

Farrar, D. E., and R. R. Glauber, 1967: Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.,* **49,** 92–107, doi:10.2307/1937887.

Folland, C., J. Owen, M. N. Ward, and A. Colman, 1991: Prediction of seasonal rainfall in the Sahel region using empirical and dynamical methods. *J. Forecast.,* **10,** 21–56, doi:10.1002/for.3980100104.

Fontaine, B., and N. Philippon, 2000: Seasonal evolution of boundary layer heat content in the West African monsoon from the NCEP/NCAR reanalysis (1968–1998). *Int. J. Climatol.,* **20,** 1777–1790, doi:10.1002/1097-0088(20001130)20:14<1777::AID-JOC568>3.0.CO;2-S.

——, ——, and P. Camberlin, 1999: An improvement of June–September rainfall forecasting in the Sahel based upon region April–May moist static energy content. *Geophys. Res. Lett.,* **26,** 2041–2044, doi:10.1029/1999GL900495.

Friedman, J. H., 1991: Multivariate adaptive regression splines. *Ann. Stat.,* **19,** 1–67, doi:10.1214/aos/1176347963.

Garric, G., H. Douville, and M. Deque, 2002: Prospects for improved seasonal predictions of monsoon precipitation over Sahel. *Int. J. Climatol.,* **22,** 331–345, doi:10.1002/joc.736.

Giannini, A., R. Saravanan, and P. Chang, 2003: Oceanic forcing of Sahel rainfall on interannual to interdecadal time scales. *Science,* **302,** 1027–1030, doi:10.1126/science.1089357.

——, M. Biasutti, I. M. Held, and A. H. Sobel, 2008: A global perspective on African climate. *Climatic Change,* **90,** 359–383, doi:10.1007/s10584-008-9396-y.

Grist, J. P., and S. E. Nicholson, 2001: A study of the dynamic factors influencing the rainfall variability in the West African Sahel. *J. Climate,* **14,** 1337–1359, doi:10.1175/1520-0442(2001)014<1337:ASOTDF>2.0.CO;2.

Hastie, T., and R. Tibshirani, 1986: Generalized additive models. *Stat. Sci.,* **1,** 297–318.

——, ——, and J. Friedman, 2009: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Springer Series in Statistics, Springer-Verlag, 793 pp.

Haywood, J. M., A. Jones, N. Bellouin, and D. Stephenson, 2013: Asymmetric forcing from stratospheric aerosols impacts Sahelian rainfall. *Nat. Climate Change,* **3,** 660–665, doi:10.1038/nclimate1857.

Heiberger, R. M., cited 2013: Statistical analysis and data display: Heiberger and Holland. R package version 2.3-42, R Project for Statistical Computing. [Available online at http://www.r-project.org/.]

Herceg, D., A. H. Sobel, and L. Sun, 2007: Regional modeling of decadal rainfall variability over the Sahel. *Climate Dyn.,* **29,** 89–99, doi:10.1007/s00382-006-0218-5.

Hoerling, M., J. Hurrell, J. Eischeid, and A. Phillips, 2006: Detection and attribution of twentieth-century northern and southern African rainfall change. *J. Climate,* **19,** 3989–4008, doi:10.1175/JCLI3842.1.

Hothorn, T., K. Hornik, C. Strobl, and A. Zeileis, 2010: Party: A laboratory for recursive partytioning. R Project for Statistical Computing, 18 pp. [Available online at cran.r-project.org/web/packages/party/vignettes/party.pdf.]

Hulme, M., 2001: Climatic perspectives on Sahelian desiccation: 1973–1998. *Global Environ. Change,* **11,** 19–29, doi:10.1016/S0959-3780(00)00042-X.

Hyndman, R. J., and A. B. Koehler, 2006: Another look at measures of forecast accuracy. *Int. J. Forecast.,* **22,** 679–688, doi:10.1016/j.ijforecast.2006.03.001.

Janowiak, J. E., 1988: An investigation of interannual rainfall variability in Africa. *J. Climate,* **1,** 240–255, doi:10.1175/1520-0442(1988)001<0240:AIOIRV>2.0.CO;2.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kandji, S., L. Verchot, and J. Mackensen, 2006: Climate change and variability in southern Africa: Impacts and adaptation in the agricultural sector. UNEP and ICRAF, 36 pp. [Available online at http://www.unep.org/themes/freshwater/documents/climate_change_and_variability_in_the_southern_africa.pdf.]

Kuhn, M., 2008: Building predictive models in R using the caret package. *J. Stat. Software,* **28,** 1–26.

Lu, J., 2009: The dynamics of the Indian Ocean sea surface temperature forcing of Sahel drought. *Climate Dyn.,* **33,** 445–460.

——, and T. L. Delworth, 2005: Oceanic forcing of the late 20th century Sahel drought. *Geophys. Res. Lett.,* **32,** L22706, doi:10.1029/2005GL023316.

McIntosh, P. C., A. J. Ash, and M. S. Smith, 2005: From oceans to farms: The value of a novel statistical climate forecast for agricultural management. *J. Climate,* **18,** 4287–4302, doi:10.1175/JCLI3515.1.

Ndiaye, O., L. Goddard, and M. N. Ward, 2009: Using regional wind fields to improve general circulation model forecasts of July–September Sahel rainfall. *Int. J. Climatol.,* **29,** 1262–1275, doi:10.1002/joc.1767.

——, M. N. Ward, and W. M. Thiaw, 2011: Predictability of seasonal Sahel rainfall using GCMs and lead-time improvements through the use of a coupled model. *J. Climate,* **24,** 1931–1949, doi:10.1175/2010JCLI3557.1.

Nelder, J. A., and R. W. M. Wedderburn, 1972: Generalized linear models. *J. Roy. Stat. Soc.,* **135A,** 370–384.

Neupane, N., and K. H. Cook, 2013: A nonlinear response of Sahel rainfall to Atlantic warming. *J. Climate,* **26,** 7080–7096, doi:10.1175/JCLI-D-12-00475.1.

Nicholson, S. E., 1995: Sahel, West Africa. *Encycl. Environ. Biol.,* **3,** 261–275.

——, 2013: The West African Sahel: A review of recent studies on the rainfall regime and its interannual variability. *ISRN Meteor.,* **2013,** 453521, doi:10.1155/2013/453521.

——, and J. Selato, 2000: The influence of La Niña on African rainfall. *Int. J. Climatol.,* **20,** 1761–1776, doi:10.1002/1097-0088(20001130)20:14<1761::AID-JOC580>3.0.CO;2-W.

——, A. K. Dezfuli, and D. Klotter, 2012: A two-century precipitation dataset for the continent of Africa. *Bull. Amer. Meteor. Soc.,* **93,** 1219–1231, doi:10.1175/BAMS-D-11-00212.1.

Palmer, T., 1986: Influence of the Atlantic, Pacific and Indian Oceans on Sahel rainfall. *Nature,* **322,** 251–253, doi:10.1038/322251a0.

Peters, A., T. Hothorn, and B. Lausen, 2002: Ipred: Improved predictors. *R news,* **2,** 33–36.

Ripley, B., cited 2013: Classification and regression trees. R package version 1.0-34, R Project for Statistical Computing. [Available online at http://www.r-project.org/.]

Rowell, D. P., 2001: Teleconnections between the tropical Pacific and the Sahel. *Quart. J. Roy. Meteor. Soc.,* **127A,** 1683–1706, doi:10.1002/qj.49712757512.

——, 2003: The impact of Mediterranean SSTs on the Sahelian rainfall season. *J. Climate,* **16,** 849–862, doi:10.1175/1520-0442(2003)016<0849:TIOMSO>2.0.CO;2.

——, C. K. Folland, K. Maskell, and M. N. Ward, 1995: Variability of summer rainfall over tropical North Africa (1906-92): Observations and modelling. *Quart. J. Roy. Meteor. Soc.,* **121,** 669–704, doi:10.1002/qj.49712152311.

Smith, T. M., R. W. Reynolds, T. C. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Climate,* **21,** 2283–2296, doi:10.1175/2007JCLI2100.1.

Sutton, C. D., 2005: Classification and regression trees, bagging, and boosting. *Handbook Stat.,* **24,** 303–329.

Venables, W. N., B. D. Ripley, and W. Venables, 1994: *Modern Applied Statistics with S-PLUS.* Series on Statistics and Computing, Vol. 250, Springer-Verlag, 462 pp.

Vizy, E. K., and K. H. Cook, 2002: Development and application of a mesoscale climate model for the tropics: Influence of sea surface temperature anomalies on the West African monsoon. *J. Geophys. Res.,* **107,** 4023, doi:10.1029/2001JD000686.

Von Storch, H., 1999: Misuses of statistical analysis in climate research. *Analysis of Climate Variability,* H. Von Storch and A. Navarra, Eds., Series on Applications of Statistical Techniques, 2nd ed., Springer-Verlag, 11–26.

——, and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

Wood, S. N., 2001: Mgcv: GAMs and generalized ridge regression for R. *R news,* **1,** 20–25.