Clim Dyn
DOI 10.1007/s00382-016-2980-3

Atmospheric controls on Puerto Rico precipitation using artificial neural networks

Craig A. Ramseyer[1] · **Thomas L. Mote**[1]

Received: 14 July 2015 / Accepted: 4 January 2016

Abstract The growing need for local climate change scenarios has given rise to a wide range of empirical climate downscaling techniques. One of the most critical decisions in these methodologies is the selection of appropriate predictor variables for the downscaled surface predictand. A systematic approach to selecting predictor variables should be employed to ensure that the most important variables are utilized for the study site where the climate change scenarios are being developed. Tropical study areas have been far less examined than mid- and high-latitudes in the climate downscaling literature. As a result, studies analyzing optimal predictor variables for tropics are limited. The objectives of this study include developing artificial neural networks for six sites around Puerto Rico to develop nonlinear functions between 37 atmospheric predictor variables and local rainfall. The relative importance of each predictor is analyzed to determine the most important inputs in the network. Randomized ANNs are produced to determine the statistical significance of the relative importance of each predictor variable. Lower tropospheric moisture and winds are shown to be the most important variables at all sites. Results show inter-site variability in u- and v-wind importance depending on the unique geographic situation of the site. Lower tropospheric moisture and winds are physically linked to variability in sea surface temperatures (SSTs) and the strength and position of the North Atlantic High Pressure cell (NAHP). The changes forced by anthropogenic climate change in regional SSTs and the NAHP will impact rainfall variability in Puerto Rico.

Keywords Climate downscaling · Artificial neural networks · Daily Caribbean precipitation · Predictor variables

1 Introduction

Understanding the controls on rainfall variability is useful for creating climate change scenarios, particularly when employing empirical-dynamical or statistical downscaling techniques. The selection of predictor variables is one of the most important methodological steps in statistical downscaling studies (Winkler et al. 2011). Hewitson and Crane (1996) emphasize that the first assumption of statistical downscaling approaches is the inclusion of the most physically explanatory predictors into the transfer functions. The ideal predictor variable is responsive to climate change, is simulated well by global climate models, has a stable relationship with the predictand, and is sensitive to variability in the predictand (Giorgi et al. 2001; Wilby et al. 2004; Winkler et al. 2011).

Sea level pressure (SLP) and geopotential heights and thicknesses have been the most common predictor variables for downscaling studies analyzing local temperature and precipitation (Cavazos and Hewitson 2005). The majority of these studies examine areas in the mid-latitudes and polar regions. Methodological studies identifying appropriate predictor variables for tropical locations have been largely isolated to studies in Mexico (Cavazos and Hewitson 2005; Hewitson and Crane 1992). Atmospheric variables most affecting precipitation and temperature processes are likely to have some regional and local variability. For example, precipitation downscaling studies should consider important factors in precipitation variable including thermodynamics, circulation variables, and moisture content

✉ Craig A. Ramseyer
ramseyca@uga.edu

[1] Department of Geography, University of Georgia, Athens, GA 30602, USA

Published online: 19 January 2016

(Cavazos and Hewitson 2005; Trenberth et al. 2003). An appropriate methodological step for downscaling studies in novel study areas is to systematically determine the best predictor variables.

This study aims to produce a methodology for determining the controlling atmospheric variables on rainfall variability in six sites in Puerto Rico (PR). The results will produce a suite of appropriate predictor variables for use in Caribbean downscaling studies. Precipitation falling in northeast PR is critical to the island in multiple facets. Precipitation and streamflow out of the Luquillo Mountains are a significant source of municipal water for the San Juan metropolitan area (Crook et al. 2007; Fig. 1) The LM contain a range of climate-sensitive ecosystems, including montane cloud forest at the peaks. El Yunque National Forest is located within the LM and is an important economic driver for the island's tourism industry. Changes in precipitation variability have important consequences economically, biologically, and ecologically. The sites examined in the study are primarily located in northeast PR along the steep topographic gradient that culminates at the peaks of the LM with the other sites situated along the southern and western coasts. Analyzing six sites that are uniquely situated both geographically and topographically allows for an analysis of how predictor variables of precipitation vary across the island and up the topographic gradient.

This study employs the use of artificial neural networks (ANN) as it allows for non-linear functions to be established between the atmospheric predictor variables and rainfall at each site. The connections between the input and output in the ANN can be examined to determine the importance of each of the input variables to the predicted output. This provides an approach for determining the controlling atmospheric variables on observed meteorological fields at varying temporal and spatial scales.

2 Background

Climate variability throughout the Caribbean region is strongly associated with global scale wave patterns, tropical cyclones, orographic effects, sea breeze circulations, regional scale wind patterns (primarily the easterly trade winds), and intense solar heating (Taylor et al. 2002). Thus,

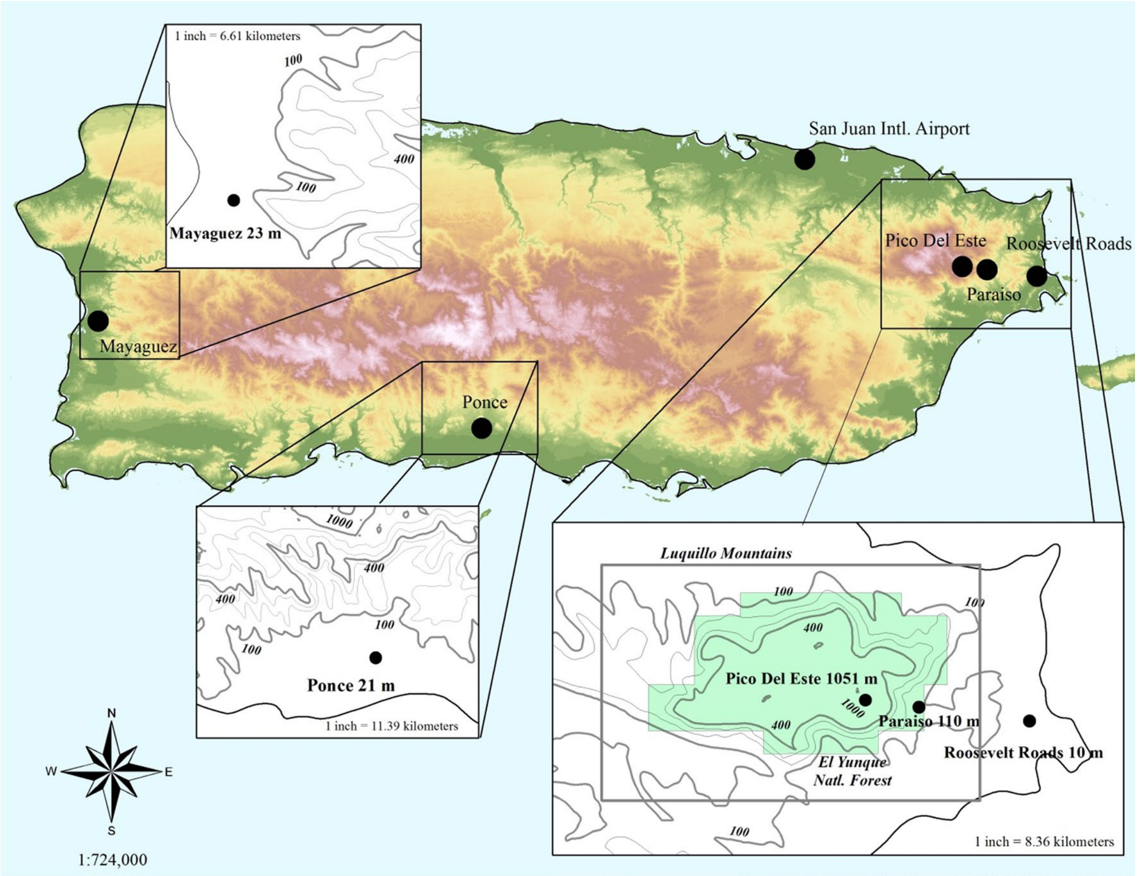

Fig. 1 Puerto Rico relief map showing the locations of the study sites. The *inset maps* show local contour maps for selected sites. The *green polygon* in the *lower-right inset* map denotes the boundary of the El Yunque National Forest

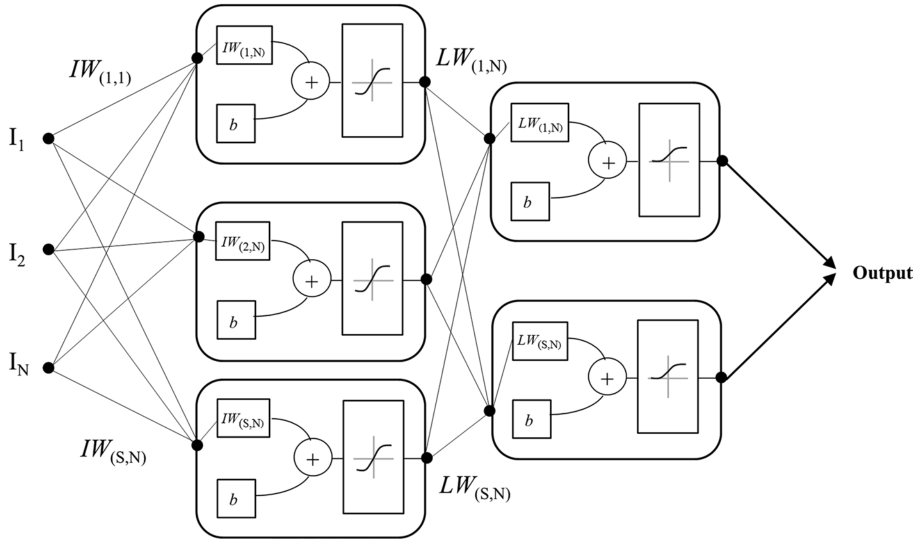

Fig. 2 Schematic of the architecture of the multilayer perceptron used in this study where the number of input weights (IW) and layer weights (LW) are a function of the number of inputs ($N = 37$) and the number of hidden nodes ($S = 20$)

local scale understanding of climate and precipitation variability depends heavily on the understanding of the regional, synoptic-scale phenomena in the Caribbean (Comarazamy and González 2008, 2011).

Precipitation over much of Puerto Rico exhibits a bimodal distribution with a dry period separating the wet seasons (Comarazamy and González 2008; Daly et al. 2003; Malmgren and Winter 1999). Spatial precipitation patterns across Puerto Rico are most strongly associated with topography, exposure and direction of the predominant winds, and proximity to the ocean (Daly et al. 2003). Much of the annual precipitation occurs in intense showers from easterly waves and tropical disturbances that occur from May to October with the balance from northerly frontal systems and localized convection (Larsen 2000).

Precipitation increases at a rate of approximately 140 % (of island average) per kilometer of elevation (Daly et al. 2003). In addition, the presence of urban areas influences precipitation patterns over Puerto Rico, but to a much lesser extent than topography (Comarazamy and González 2008). In Puerto Rico, the area of highest precipitation is in the Luquillo Mountains in northeast Puerto Rico, where steep terrain, coupled with exposure to predominant northeasterly winds and proximity to the ocean leads to frequent orographic lifting of moisture laden air (Comarazamy and González 2008; Daly et al. 2003; Jury 2009). Perturbations in predominant winds have been shown to spatially redistribute localized areas of moisture convergence which has a profound effect on precipitation patterns (Comarazamy and González 2011).

Climate downscaling involves taking coarse resolution global climate model (GCM) data and through the use of a dynamical or statistical methodology, producing local, fine-resolution climate information. Dynamical downscaling typically refers to the use of a regional climate model (RCM). These RCMs use nested grids within the GCM grid space to produce increasingly higher resolution simulations. The other group of downscaling techniques can be broadly referred to as statistical downscaling.

ANNs have been used across the physical sciences, including meteorology and climatology, as a powerful, non-linear function approximation and forecasting tool. Due to the ability to produce nonlinear functions, ANNs are well suited to be used as a problem solving tool for climate related problems (Tsonis and Elsner 1992). ANNs have been used for a wide range of applications in the atmospheric sciences including precipitation modeling (Bellerby et al. 2000; Cavazos and Hewitson 2005; Gardner and Dorling 1998; Hall et al. 1999; Haylock et al. 2006; Kuligowski and Barros 1998; Sahai et al. 2000; Schoof and Pryor 2001; Silverman and Dracup 2000; Valverde Ramírez et al. 2005). These studies use ANNs to derive nonlinear relationships between atmospheric controls and precipitation. In this way, ANNs are a tool used for statistical climate downscaling, more specifically, empirical-dynamical downscaling (Winkler et al. 2011). ANNs have been shown to out-perform other statistical techniques for studies analyzing precipitation (Kuligowski and Barros 1998; Valverde Ramírez et al. 2005).

This study uses feed-forward multilayer perceptrons (FFMLP) with backpropagation of error, one type of artificial neural network. The FFMLPs used in the study contain three layers (Fig. 2). The architecture includes an input layer containing a suite of atmospheric variables, a hidden layer, and an output layer that contains the predicted/forecasted precipitation. The input layer receives input data, with one node per input variable. The input layer is connected to the hidden layer through a matrix of weighted connections, referred to as input weights. The hidden layer is linked to the output layer via a matrix of layer weights.

Table 1 ECMWF ERA-Interim reanalysis variables used in the network input layer

Atmosphere level	Circulation	Moisture	Thickness
Surface/1000 hPa	SLP (slp) Geopotential Height (z0) U and V winds (u0, v0) Vorticity (vo0) Divergence (d0)	SH (q0) RH (rh0)	500–1000 hPa (th1)
925 hPa	Geopotential Height (z9) U and V winds (u9, v9) Divergence (d9) Vertical Velocity (w9)	SH (q9)	
850 hPa	U and V winds (u8, v8) Divergence (d8) Geopotential Height (z8) Vertical Velocity (w8)	SH (q8)	500–850 hPa (th8)
700 hPa	U and V winds (u7, v7) Divergence (d7) Geopotential Height (z7) Vertical Velocity (w7)	SH (q7) RH (rh7)	
500 hPa	U and V winds (u5, v5) Divergence (d5) Geopotential Height (z5) Vorticity (vo5)	SH (q5)	
200 hPa	Divergence (d2) Geopotential Height (z2)		

The weight matrices are initialized randomly. As training data are introduced to the model, the network "learns" via a training function (Bayesian regularization here). This algorithm seeks to reduce some error statistic (e.g. mean squared error). After each training epoch, the error between the output and the target is back-propagated through the ANN to update the weights for the next training epoch (Cavazos 1999; Cavazos and Hewitson 2005). The remaining data is used for independent testing and/or validation. For further details on FFMLP and the backpropagation of error algorithms, refer to (Hewitson and Crane 1994) and (Maier and Dandy 2000).

3 Data

3.1 Data sources

Six study sites in Puerto Rico were chosen representing a range of geographic situations and precipitation climatologies. Three of the sites are located in the vicinity of the Luquillo Mountains including Roosevelt Roads (RR), Paraiso (PA), and Pico Del Este (PE). To determine if changes in optimum predictor variables existed up the topographic gradient, each of the three sites is located at different elevations (Fig. 1). PE was selected as the high elevation site where the rain gauge is located at an elevation of 1051 m asl. RR provides a low-elevation site at 10 m asl and is situated on the eastern coast of Puerto Rico. The PA rain gauge is situated at 110 m asl and serves as a mid-elevation site, between PE and RR on the topographic gradient up the Luquillo Mountains. The fourth site is located at the San Juan International Airport (SJ) which provides an additional low-elevation site northwest of Luquillo Mountains in a highly urbanized area. The two remaining sites are relatively far removed from the other sites with Ponce (PO) on the south coast and Mayaquez (MA) on the west coast (Fig. 1).

Daily precipitation data for each site were acquired from the Global Historical Climatology Network-Daily dataset (Menne et al. 2012). All data available from 1980 to 2009 were obtained. This date range was used as it provided the least missing values for a 30 year period for the six sites. The SJ site is the only site that has a continuous record spanning the 30 year period. The other sites had less than 30 % missing data. The reanalysis data were pruned to match the availability of the precipitation data at each site.

Six-hourly ERA-Interim gridded reanalysis variables were obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) for 12Z for a 0.125° latitude by 0.125° longitude grid situated over Puerto Rico. ERA-Interim is a global atmospheric reanalysis product available from 1979 and updated continuously (Dee et al. 2011). The data assimilation scheme used is a four-dimensional variational analysis (4D-Var) with a 12-h analysis window. Cycle 31r2 of the ECMWF Integrated Forecast System (IFS) was used in the development of ERA-Interim. Reanalysis data were collected for the study period from 1980 to 2009. The grid resolution of this product allows for better representation of smaller scale atmospheric circulations and fluctuations in atmospheric variability. This study uses 37 atmospheric predictor variables which include divergence, geopotential heights, surface wind fields, mean sea-level pressure, relative humidity, specific humidity, vorticity, vertical velocities, u and v winds, and atmospheric thickness at different pressure levels (Table 1). These variables were selected following the methodology of Cavazos and Hewitson (2005) with the addition of the 925 hPa level, which was added for increased vertical resolution in the lower troposphere.

3.2 Methods

The utility of the networks presented here is to provide a predictive tool of precipitation variability, and ultimately, to determine which atmospheric controls are most important in that precipitation variability. The precipitation and reanalysis data were pre-processed prior to training the ANN. The reanalysis and precipitation data were smoothed with a 5-day equally weighted moving average filter. This allowed for the focus to be on the atmospheric modes conducive to precipitation (Hewitson and Crane 1994). Missing data, including the first and last 2 days, were removed.

The independent (reanalysis atmospheric variables) and the dependent variables (rainfall observations) were normalized prior to training the ANNs. The rainfall data were fitted to the range (0–1) in order to conform to the output provided by the ANN using a log-sigmoid output transfer function. The following algorithm is used to normalize the rainfall data to the range (0–1)

$$\mathrm{t_n} = \frac{\mathrm{Y_n} - \mathrm{Y_{min}}}{\mathrm{Y_{max}} - \mathrm{Y_{min}}} \tag{1}$$

where $\mathrm{t_n}$ is the normalized value of Y at time n, $\mathrm{Y_n}$ is the original rainfall value at time n, and $\mathrm{Y_{min}}$ and $\mathrm{Y_{max}}$ are the minimum and maximum values of the variable Y. The reanalysis variables were standardized to account for the variation in measurement scales of each. This involves producing converted reanalysis variables with zero mean and unity standard deviation.

The normalized rainfall data and standardized reanalysis data were assigned as targets and inputs, respectively, in the networks. The MATLAB Neural Network Toolbox was utilized to produce the two-layer FFMLPs in this study. The networks used in this study consisted of three components, an input layer ($\mathrm{N} = 37$), a hidden layer, and an output layer. In order to determine the optimal number of hidden neurons in the hidden layer, a suite of ANNs were constructed using a range of neurons in the hidden layer. The hidden layer size was increased at increments of five and ranged from 5 to 40 neurons. The other model parameters were held constant. The mean square error (MSE) was assessed for each ANN. There are a few considerations that need to be taken into account when deciding the number of hidden neurons. Networks with fewer hidden neurons are preferable due to their decreased computational expense and because the network has fewer connections, and can be easier to interpret. Networks with a large number of hidden neurons can have advantages when the number of input variables is high. If too many neurons are used however, the model can begin to over-fit and become less useful in generalizing. MSE decreased appreciably from the five- node model to the 20-node model. The lowest MSE occurred when having 40 neurons; however, there was a negligible difference in MSE between the 20- and 40-node models. The computational cost was significantly higher for the 40-node network. As a result, the 20-node model architecture was deemed most appropriate for this study (Fig. 2). All training networks used in the analysis performed best with 250 epochs (training runs).

Two neural transfer functions are used in the networks. The first transfer function operates on the input data to produce the hidden layer output. For our data, a hyperbolic tangent sigmoid transfer function is appropriate as it creates output in the range (−1 1), which conforms to the range of the standardized reanalysis data. The second transfer function operates on the hidden layer output to create the output layer data. Because the rainfall data only have positive values, the log-sigmoid transfer function was selected as it constrains output to the range (0 1).

For each observation site, 1000 networks were created with each starting with different initial input weights. This step is necessary because any one network may be converging to local minima rather than global minima. A network converging to local minima versus global minima will have different weight matrices. The weight matrices are used in the calculation of variable importance, thus, the local minima weight matrix would lead to incorrect assumptions about variable importance.

Each network divided the input and target data randomly between a training (85 %) and test set (15 %). The best performing network was selected by determining which network had the lowest MSE. For the optimized network, Pearson's correlation coefficient was calculated on the observed and modeled rainfall data for both the training and test set (Table 2). The MSE of the optimized network was decomposed into variance and bias:

$$\begin{aligned}
\mathrm{MSE} &= \sum_{i=1}^{n}(\mathrm{y}_i - \mathrm{x}_i)^2 \\
&= \overline{\left(\mathrm{y}_i - \mathrm{x}_i\right)^2} \\
&= \overline{(\mathrm{y}-\mathrm{x})^2} - \overline{\left[(\mathrm{y}-\mathrm{x})\right]}^2 - \overline{(\mathrm{y}-\mathrm{x})}^2 \\
&= s^2_{\mathrm{y-x}} + (\bar{\mathrm{y}} - \bar{\mathrm{x}})^2 \\
&= \text{Variance} + \text{Bias}^2
\end{aligned}$$

Table 2 Performance statistics from the best performing networks for each site from 1980 to 2009 including mean-squared error (MSE), root mean-squred error (RMSE), variance (VAR), bias, mean absolute error (MAE), Pearson's Coefficient for the test set (r_{test}) and for all observations (validation and test sets; r_{all}), and the coefficient of determination for the test set (R^2_{test}) and for all observations (R^2_{all})

Location	RMSE (mm)	MSE (mm^2)	MAE (mm)	VAR (mm^2)	BIAS (mm)	r_{test}	r_{all}	R^2_{test}	R^2_{all}
RR	2.50	6.24	1.70	6.18	-0.25	0.80	0.93	0.71	0.86
PE	5.14	26.40	3.86	26.39	-0.11	0.81	0.90	0.70	0.80
PA	4.49	20.14	3.25	20.12	-0.18	0.76	0.84	0.59	0.70
SJ	2.64	6.98	1.82	6.96	-0.13	0.77	0.86	0.61	0.73
MA	3.21	10.33	2.18	10.31	-0.16	0.78	0.86	0.58	0.74
PO	2.34	5.45	1.35	5.34	-0.34	0.81	0.93	0.68	0.86

MSE, MAE, RMSE, VAR, and BIAS are calculated on the full time series

The variance term in this decomposition is the variance of the errors, and the bias term is the average deviation of the model forecasted rainfall from the observed rainfall. The model forecasts can be improved by eliminating the bias term. Thus, models with a large bias will be improved more than models with a small bias term. The forecasts can be shifted (±) by the constant bias term, but leaves the variance unaffected.

The bias term above was discussed as a constant value. However, it can be assumed that bias may be larger for certain portions of the rainfall distribution. In other words, the bias may be dependent on rainfall amount. Assume E(x|y) is the conditional mean (or expected value) of an observed rainfall amount (x), for a given forecasted rainfall amount (y). In order to calculate E(x|y) for our continuous variable, the forecasted rainfall amounts are binned and the average of the observed rainfall is computed for each bin. The resulting means can be plotted as a function of the forecasts and assess the reliability and conditional bias of the model. To account for the sampling variation of the conditional mean, 1000 bootstrapping trials were performed on the conditional mean of the observed rainfall. The sampling variations of the conditional mean can be used to create a reliability plot which visualizes uncertainty at each bin (Marzban 2009).

The input and layer weight matrices of the optimized networks are used to compute the variable importance. Variable importance was determined by following Garson's approach (Garson 1991; Goh 1995). The networks consist of connection weights between the input-hidden layers (IW) and the hidden-output layers (LW). The Garson method involves calculating the product of the IW and LW connection weights (IHO) for each input and hidden neuron (e.g. $IHO_{A1} = IW_{1,A} \times LW_{1,A}$). With 37 input neurons and 20 hidden neurons, the networks in this study have 740 IHO connection weights.

Next, the absolute value of the IHO weights are summed across each input variable (CW) resulting in 37 overall connection weights (e.g. $CW_1 = |IHO_{A1}| + |IHO_{B1}| \ldots$) where IHO_{A1} and IHO_{B1} refer to the IHO value for the first input neuron and the first and second hidden neuron, respectively. The overall connection weights are summed and the relative importance (RI) of each input variable is calculated (e.g.$RI\ (\%) = \frac{CW_1}{\sum_{i=1}^{n} CW_i} \times 100$). Because the absolute value is used in the calculation of the IHO values, the variable importance calculated here is a magnitude, and does not contain information on the direction of the input–output interaction (Olden and Jackson 2002).

A randomization test is performed on each sites' optimized network in an effort to quantitatively determine which IHO connections are statistically significant (Olden and Jackson 2002). Additionally, the statistical significance of the relative importance can be assessed and is represented using p-values. After the optimized network is established and the IHO weights ($IHO_{observed}$) and relative importance ($RI_{observed}$) have been calculated, the rainfall observations are randomly permuted and new neural networks are constructed using the same initial connection weights for the optimized network. It is critical to use the same initial input and layer weights for each randomization. Without this step, important differences between the observed and random weights cannot be separated from the differences that arise from different initial conditions of the weight matrices (Olden and Jackson 2002). In this study, 999 randomized networks are developed. After each network is trained, the IHO weights (IHO_{random}), relative importance (RI_{random}), and W_{random} are calculated and recorded. The statistical significance can be calculated as the proportion of randomized values whose value is equal to or more extreme than the observed values (e.g. $RI_{random} \geq RI_{observed}$). The p-values from the randomization tests can help determine which IHO, CW, and RI values are statistically different than what would be expected by chance alone. Further details on the randomization test for ANNs can be found in Olden and Jackson (2002).

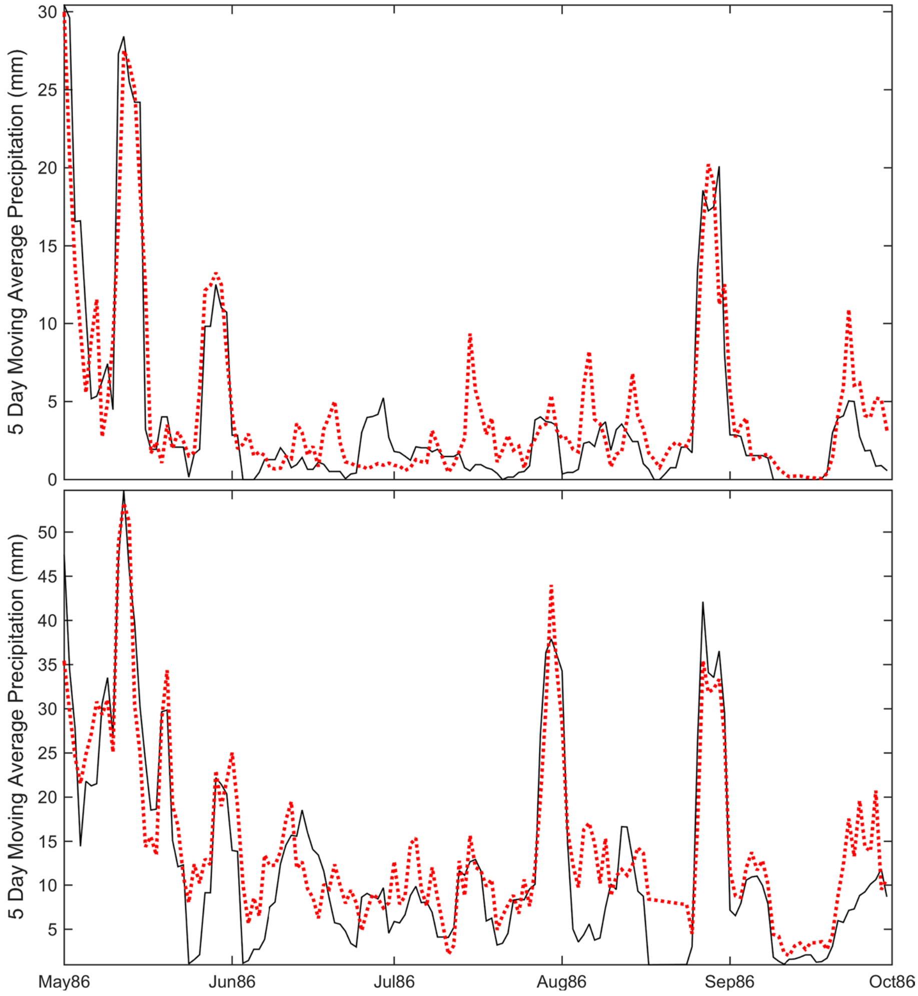

Fig. 3 Time series of predicted (*red*) and observed (*black*) smoothed rainfall for SJ (*top*) and PE (*bottom*) during the wet months from May 1986 to October 1986

4 Results

4.1 Network performance

As discussed in the methodology, for every site 1000 networks were created with different network weight initializations for each network. The best performing ANNs were selected based on a range of performance metrics. Table 2 shows the summary statistics for each sites' best network. According to the mean squared error, mean absolute error, and the root mean squared performance metrics, the RR and SJ ANNs performed better than the ANNs for PE and PA (Table 2). This was expected due to the wider precipitation distribution and greater number of extreme events observed at PE and PA. Smoothed rainfall values of over 20 mm were recorded at PE in 16 % of all observations while PA surpassed this threshold in 7 % of the observations. SJ, MA, and RR surpassed the 20 mm threshold in 2 % or fewer of the observations. These data were included in the training and testing of the network because the authors wanted to ensure that the model was capable of reproducing the variability in the observed precipitation record. Removing the outlier data may result in a model with better performance. However, keeping the data outliers ensures that the variable importance calculated is based on a model that is capable of predicting rainfall throughout the distribution. The networks show considerable ability at capturing the overall trends in the smoothed rainfall data (Fig. 3).

Figure 4 shows the PE reliability plot of the modeled versus measured rainfall. The graphic allows for a visual assessment of the uncertainty and reliability of the predicted rainfall. For reference, a trendline is displayed showing a perfect prediction. The boxplots provide visually represents the sampling distribution of the conditional mean of the observed rainfall, given the predicted rainfall. The horizontal line inside the boxplot represents the median of the conditional mean while the upper and lower edges of the boxes represent the 25th and 75th percentile. A skewed distribution is represented by a horizontal line that is not centrally located in the boxplot. The size of the boxplots represents certainty where small boxes are representative

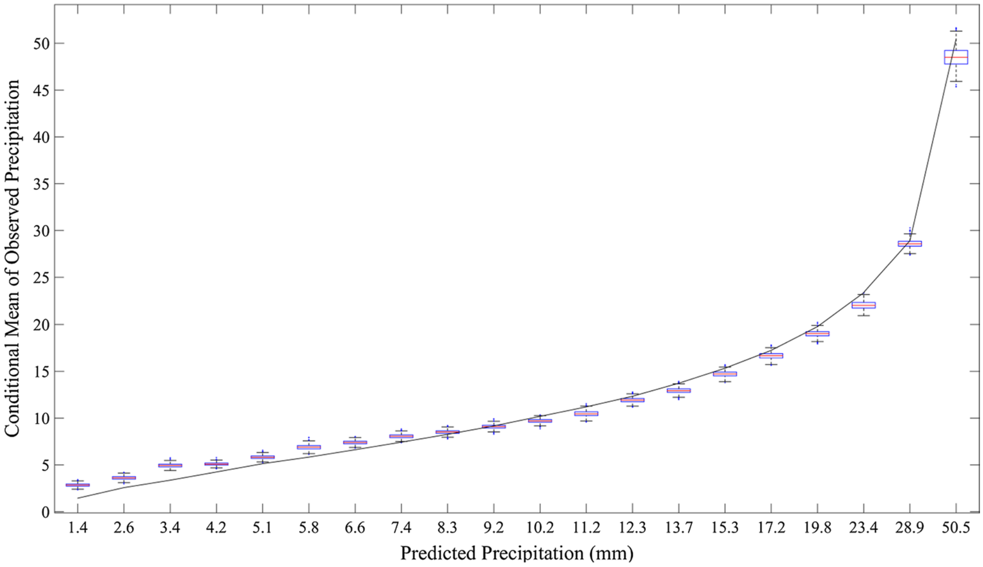

Fig. 4 Reliability plot for PE showing the conditional mean of the observed precipitation and the predicted precipitation. A *reference line* is added showing an optimal 1:1 relationship

Table 3 Percentage of daily smoothed rainfall values where the difference between observed and predicted rainfall was within 1 and 5 mm

	Predicted ± 1 mm (%)	Predicted ± 5 mm (%)
RR	45.6	94.1
PE	18.8	72.5
PA	24.7	78.9
SJ	43.3	93.6
PO	60.5	95.2
MA	40.8	89.2

of more certain predictions. The reliability of the predictions can be visually assessed by the proximity of the boxplots to the diagonal. The whiskers represent the extreme data not including the outliers while the outliers are plotted individually. The size of the boxplot is indicative of the uncertainty.

Using PE as an example, moderate predicted rainfall (e.g. 9.2 mm) is accompanied by a narrow range of observed rainfall. This narrow range indicates that there is a high degree of certainty in the predictions. Comparing this to the heaviest predicted rainfall boxplot (50 mm), a larger variation in corresponding observed rainfall is noted. The median of the heavy rainfall boxplot intersects the trendline and represents that the model is predicting rainfall reliably in these amounts. However, due to the spread in the observed rainfall (represented by the size of the boxplot), these predictions are less certain. The reliability plot also shows that moderate rainfall predictions are reliable and that there is a high degree of certainty about those predictions. For low daily rainfall, the network slightly overpredicts the observed rainfall, and there is a high degree of certainty. Under-estimation of high rainfall and overestimation of low rainfall is common in different methods and has been documented in similar studies (e.g. Cavazos and Hewitson 2005).

The correlation coefficient between all of the predicted and observed 5-day smoothed daily rainfall data points (r_{all}) was 0.84 or greater for all six networks. This indicates that at least 70 % of the local rainfall can be explained by the atmospheric variables, while 30 % of the local rainfall is a function of effects not included as input variables. Additionally, neural networks are specifically trained in order to generalize the target data. Neural network training involves using a sub-sample of a continuous function to find a generalized form of that function (Hewitson and Crane 1994). Thus, the models ability to forecast extreme events can be affected. Despite these limitations, the models showed relatively good skill at predicting the onset of a heavy rainfall period. At the PE, there were 1211 rainfall values between 20 and 40 mm, 223 values between 40 and 60 mm, and 72 values over 60 mm making it the wettest of the six sites. The mean difference between the observed and predicted precipitation was 6.0, 6.8, and 7.7 mm (0.30 in.) for each bin respectively. The ability for the model to correctly predict the onset of a heavy rainfall event (despite under- or over-predicting the magnitude) ensures that the input variable contribution to the net is relevant for the full precipitation distribution for the site. This is important because the atmospheric controls on precipitation seem to be similar for dry and wet days.

Another metric of assessing the networks performance is to calculate the number of rainfall observations that were predicted to within one mm and five mm of rainfall (Table 3). PE had the lowest percentage of predictions within these thresholds with 78.5 % of prediction within five mm of the observed rainfall. The four lower elevation sites with less variable rainfall and fewer extreme events had a higher percentage of predictions within these ranges with 89 % or more of predictions within five mm of the observed rainfall.

4.2 Variable importance in networks

To calculate the relative importance of each of the input variables, Garson's algorithm was employed. The top six contributors to the best networks are listed in Table 4. The best networks were randomized using the approach discussed in (Olden and Jackson 2002). After 999 bootstrapping runs, the statistical significance of the importance of the input variables in the original (best) network was assessed. The italicized variables in the table are statistically significant based on χ = 0.05. This indicates that there is a low probability of getting a value that larger or larger than the relative importance of that variable in the original network. The results indicate that q0 is one of the two leading contributors at each of the three sites in or directly adjacent to the Luquillo Mountains (RR, PE, and PA). At PE, q0 is also statistically significant.

The low-level u winds (u0, u9, u8) are also important at each of the four northeast sites (Fig. 5). U0 is the most important predictor variable at PA and SJ and it also statistically significant. U9 is one of the leading six contributors at each of the four northeast sites. U9 are statistically significant and the most important variable in the MA network. The MA and PO networks are more controlled by low-level v-winds than the northeast sites.

Seasonal models were also produced to evaluate variability in predictor variables between the dry and wet seasons. Low-tropospheric moisture and u-winds were the most common predictors in both seasons. Thus, the vertical profile of u-wind and specific humidity through the lower troposphere from the surface to 700 hPa is driving rainfall in the study area regardless of season. As expected, the dry season models perform better than the yearly models due to decreased precipitation variability and lack of extreme precipitation events during the dry season. The wet season models have slightly decreased performance compared to the dry season or yearly models due to the increase in precipitation variability and extreme events (e.g. easterly waves). Despite a slight decrease in the wet season model performance, these models still capture the variability in the time series. However, the extreme events are under-predicted and cause an increase in the error metrics.

The seasonal models tend to be slightly noisier compared to the yearly models in terms of the top predictors. The seasonal models tend to have less agreement on which levels in the lower troposphere are most important. The primary difference between the seasonal and yearly models was the increased prevalence of z5 as a top predictor in the wet and dry season models. This could be attributed mid-tropospheric intrusions from extratropical frontal systems during the dry season and easterly waves during the wet season.

Table 4 The most important input variables for the best performing network as calculated using Garson's algorithm for determining variable relative importance (RI)

RR	PE	PA	SJ	MA	PO
q0	*q0*	*u0*	*u7*	*u9*	*u9*
u8	*q7*	q0	z5	v9	q7
u9	u8	w9	u9	w9	rh7
q7	v9	q9	q7	w8	v9
u7	*rh7*	v9	q0	u0	u8
w8	u9	u9	u8	v8	q0

Italicized variables represent those RI values that are statistically significant based on χ = 0.05 when compared to the randomized networks

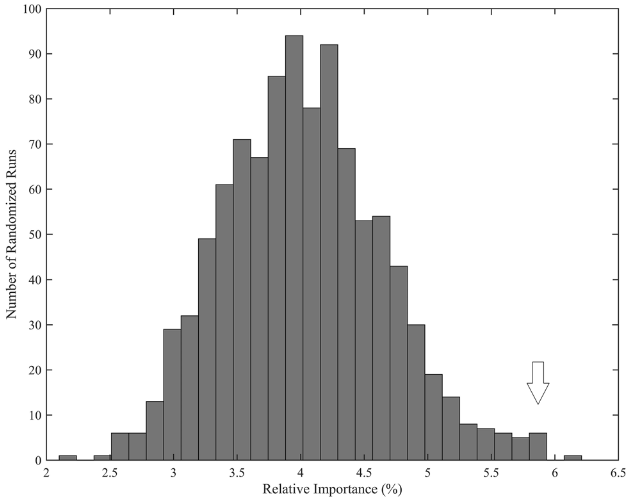

Fig. 5 The relative importance (%) of q0 for all randomized networks for PE. The *arrow indicates* the relative importance for the observed network, which in this example is statistically significant

5 Discussion

The most important atmospheric predictor variables for rainfall in Puerto Rico are low tropospheric specific humidity and low tropospheric u-winds. For the southern and western site, low tropospheric v-winds are also an important control. More specifically, q0 and u0 are the top predictors at the northeast Puerto Rico sites. The PO and MA networks top predictor was u9. Additionally, v9 are an important predictor in both networks outside of northeast Puerto Rico.

Q0 is the most important variable in the RR and PE networks and is statistically significant in the PE network. The PA, SJ, and PO networks also have q0 as one of the top six contributors. It is important to note that PE is located above the 1000 hPa pressure level. The q0 control on precipitation at PE is likely due to moisture advection up the topographic gradient into the site. Air at 1000 hPa is transported by the easterly trade winds and encounters the LM. Forced ascent up the windward side of the LM provides an additional lifting mechanism and may allow condensation processes to occur even with a lack of another lifting mechanism (e.g. instability). At a more broad scale, high specific humidity likely has two roles in increased precipitation in the LM. High humidity at low-levels leads to decreased atmospheric stability. Additionally, the high moisture content of the air decreases the lifted condensation level (LCL) and allows for initiation of precipitation processes at lower levels.

Other than specific humidity, the low-level u-winds are the other top predictor in the networks. U0 is the top predictor at the PA and SJ sites and is statistically significant. Additionally, u9 is the top predictor at the MA (statistically significant) and PO sites. U9 is one of the top six predictors at the PE and RR sites. The prevalence of the low-level u-winds and low-level specific humidity as the two most important predictors of precipitation in PR in the networks is validated in the literature.

Due to strong ocean–atmosphere coupling in the tropics, low level specific humidity is closely related to SSTs and the low level wind field. Previous research suggests tropical North Atlantic SSTs strongly influences early season Caribbean rainfall variability while equatorial Pacific and Atlantic SSTs influence late season rainfall variability (Taylor et al. 2002). Q0 is linked to the magnitude of the near-surface trade winds in the tropical Atlantic. The North Atlantic Subtropical High Pressure cell (NAHP) is a large scale driver of these feedbacks as it controls trade wind strength in the region. The intensification of the NAHP translates in stronger trade winds and lower SSTs (Gamble and Curtis 2008; Giannini et al. 2000; Granger 1985; Hastenrath 1976).

Variations in low tropospheric and surface trade wind intensity are the primary forcing mechanism in sea-surface temperatures (SST) over the tropical Atlantic (Nobre and Srukla 1996). As winds increase, wind stress and turbulent mixing increase leading to cooler SSTs. This also leads to increased wind shear (assuming no mid-upper tropospheric disturbance) which increases atmospheric stability and caps deep convection (Enfield and Alfaro 1999). Decreases in trade winds leads to increased SSTs, enhanced surface evaporation, and lower tropospheric moisture convergence (Wu and Kirtman 2005, 2011). This also leads to increased atmospheric instability, allowing for greater deep convection and rainfall in the Caribbean (Enfield and Alfaro 1999).

After q0, u0, and u9, the other common variable amongst most of the networks is 700 hPa specific humidity (q7). It is the second most important input variable, and is statistically significant in the PE network. In three other networks q7 is one of the four most important variables. Over this portion of the Caribbean, the trade wind inversion is located at approximately 2300 m (Gutnick 1958). Cumulus and stratocumulus convection is capped at this level, due to subsidence aloft (Schubert et al. 1995). The closest pressure level included in the network input layer was 700 hPa. The importance of q7 in the networks is indicating that the moisture content in the layer from the surface to the trade wind inversion is an important predictor of PR rainfall.

6 Conclusions

This study presents a methodology for selecting appropriate predictor variables to be used in climate downscaling studies. FFMLP neural networks were tested and randomized for six sites in Puerto Rico. The results of this study indicate that 1000 hPa specific humidity, 10-m and 925 hPa u-winds are the largest controls on precipitation in Puerto Rico. Increased low-tropospheric moisture destabilizes the lower troposphere and leads to lower lifted condensation levels. Decreased low-tropospheric and surface u-winds enhance evaporation off the Caribbean Sea and tropical North Atlantic (increasing 1000 hPa specific humidity) and increase low-tropospheric moisture convergence. Decreases in low-tropospheric winds leads to low wind shear environments which allows for effective trade wind cumulus convection. The notable difference between the northeast sites and the PO and MA sites was the slight increase in importance of the low-level v-component of the wind. This is due to the orientation of the sites to nearby topography. Our results indicate that site elevation had little influence on the controlling atmospheric variables in the networks. The highest elevation site (PE) in northeast PR had similar controls as the nearby low elevation sites.

Although not explicitly resolved in this study, the results of this study highlight the importance of SSTs on precipitation variability in the study area. Due to the strong ocean–atmosphere coupling in the tropics, SST variability is linked to low tropospheric moisture and wind. These changes are coupled with variability in NAHP strength and position, which drive trade wind strength. Future changes in Caribbean and tropical North Atlantic SSTs as well as the strength and position due to anthropogenic climate change will force changes in precipitation variability in Puerto Rico. The results from this study suggest that future climate scenarios of precipitation variability in PR and the Caribbean using statistical downscaling methodologies should use low tropospheric specific humidity and winds as predictor variables. This research was supported by grant DEB-1239764 from NSF to the Institute for Tropical Ecosystem Studies, University of Puerto Rico, and to the International Institute of Tropical Forestry USDA Forest Service, as part of the Luquillo Long-Term Ecological Research Program.

References

Bellerby T, Todd M, Kniveton D, Kidd C (2000) Rainfall estimation from a combination of TRMM precipitation radar and GOES multispectral satellite imagery through the use of an artificial neural network. J Appl Meteorol 39:2115–2128

Cavazos T (1999) Large-scale circulation anomalies conducive to extreme precipitation events and derivation of daily rainfall in northeastern Mexico and southeastern Texas. J Clim 12:1506–1523. doi:10.1175/1520-0442(1999)012<1506:LSCACT>2.0.CO;2

Cavazos T, Hewitson BC (2005) Performance of NCEP-NCAR reanalysis variables in statistical downscaling of daily precipitation. Clim Res 28:95–107. doi:10.3354/cr028095

Comarazamy DE, González JE (2008) On the validation of the simulation of early season precipitation on the Island of Puerto Rico using a mesoscale atmospheric model. J Hydrometeorol 9:507–520. doi:10.1175/2007jhm804.1

Comarazamy DE, González JE (2011) Regional long-term climate change (1950–2000) in the midtropical Atlantic and its impacts on the hydrological cycle of Puerto Rico. J Geophys Res. doi:10.1029/2010jd015414

Crook K, Scatena FN, Pringle CM (2007) Water withdrawn from the Luquillo experimental forest, 2004. USDA forest service general technical report IITF-GTR-34

Daly C, Helmer EH, Quiñones M (2003) Mapping the climate of Puerto Rico, Vieques and Culebra. Int J Clim 23:1359–1381. doi:10.1002/joc.937

Dee D et al (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorolog Soc 137:553–597

Enfield DB, Alfaro EJ (1999) The dependence of Caribbean rainfall on the interaction of the tropical Atlantic and Pacific Oceans. J Clim 12:2093–2103. doi:10.1175/1520-0442(1999)012<2093:TDOCRO>2.0.CO;2

Gamble DW, Curtis S (2008) Caribbean recipitation: review, model and prospect. Prog Phys Geogr 32:265–276. doi:10.1177/0309133308096027

Gardner MW, Dorling SR (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 32:2627–2636. doi:10.1016/S1352-2310(97)00447-0

Garson DG (1991) Interpreting neural-network connection weights. Artif Intell Expert 6:47–51

Giannini A, Kushnir Y, Cane MA (2000) Interannual variability of Caribbean rainfall, ENSO, and the Atlantic Ocean. J Clim 13:297–311. doi:10.1175/1520-0442(2000)013<0297:IVOCRE>2.0.CO;2

Giorgi F et al (2001) Emerging patterns of simulated regional climatic changes for the 21st century due to anthropogenic forcings. Geophys Res Lett 28:3317–3320

Goh A (1995) Back-propagation neural networks for modeling complex systems. Artif Intell Eng 9:143–151

Granger OE (1985) Caribbean climates. Prog Phys Geogr 9:16–43

Gutnick M (1958) Climatology of the trade-wind inversion in the Caribbean. Bull Am Meteorol Soc 39:410–420

Hall T, Brooks HE, Doswell CA III (1999) Precipitation forecasting using a neural network. Weather Forecast 14:338–345

Hastenrath S (1976) Variations in low-latitude circulation and extreme climatic events in the tropical Americas. J Atmos Sci 33:202–215

Haylock MR, Cawley GC, Harpham C, Wilby RL, Goodess CM (2006) Downscaling heavy precipitation over the United Kingdom: a comparison of dynamical and statistical methods and their future scenarios. Int J Clim 26:1397–1415. doi:10.1002/joc.1318

Hewitson BC, Crane RG (1992) Large-Scale atmospheric controls on local precipitation in tropical Mexico. Geophys Res Lett 19:1835–1838

Hewitson BC, Crane RG (1994) Neural nets: applications in geography: applications for geography, vol 29. Springer, Berlin

Hewitson B, Crane R (1996) Climate downscaling: techniques and application. Clim Res 7:85–95

Jury MR (2009) An intercomparison of observational, reanalysis, satellite, and coupled model data on mean rainfall in the Caribbean. J Hydrometeorol 10:413–430

Kuligowski RJ, Barros AP (1998) Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. Weather Forecast 13:1194–1204. doi:10.1175/1520-0434(1998)013<1194:LPFFAN>2.0.CO;2

Larsen MC (2000) Analysis of 20th century rainfall and streamflow to characterize drought and water resources in Puerto Rico. Phys Geogr 21:494–521

Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ Modell Softw 15:101–124

Malmgren BA, Winter A (1999) Climate zonation in puerto rico based on principal components analysis and an artificial neural network. J Clim 12:977–985. doi:10.1175/1520-0442(1999)012<0977:CZIPRB>2.0.CO;2

Marzban C (2009) Performance measures and uncertainty. In: Haupt S, Pasini A, Marzban C (eds) Artificial intelligence methods in the environmental sciences. Springer, Netherlands, pp 49–75. doi:10.1007/978-1-4020-9119-3_3

Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG (2012) An overview of the global historical climatology network-daily database. J Atmos Ocean Technol 29:897–910

Nobre P, Srukla J (1996) Variations of sea surface temperature, wind stress, and rainfall over the tropical Atlantic and South America. J Clim 9:2464–2479

Olden JD, Jackson DA (2002) Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecol Model 154:135–150. doi:10.1016/S0304-3800(02)00064-9

Sahai A, Soman M, Satyan V (2000) All India summer monsoon rainfall prediction using an artificial neural network. Clim Dyn 16:291–302

Schoof JT, Pryor SC (2001) Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks. Int J Clim 21:773–790. doi:10.1002/joc.655

Schubert WH, Ciesielski PE, Lu C, Johnson RH (1995) Dynamical adjustment of the trade wind inversion layer. J Atmos Sci 52:2941–2952

Silverman D, Dracup JA (2000) Artificial neural networks and long-range precipitation prediction in California. J Appl Meteorol 39:57–66. doi:10.1175/1520-0450(2000)039<0057:ANNALR>2.0.CO;2

Taylor MA, Enfield DB, Chen AA (2002) Influence of the tropical Atlantic versus the tropical Pacific on Caribbean rainfall. J Geophys Res: Ocean 107:10-11-10-14

Trenberth KE, Dai A, Rasmussen RM, Parsons DB (2003) The changing character of precipitation. Bull Am Meteorol Soc 84:1205–1217

Tsonis AA, Elsner JB (1992) Nonlinear Prediction as a way of distinguishing chaos from random fractal sequences. Nature 358:217–220

Valverde Ramírez MC, de Campos Velho HF, Ferreira NJ (2005) Artificial neural network technique for rainfall forecasting applied to the São Paulo region. J Hydrol 301:146–162. doi:10.1016/j.jhydrol.2004.06.028

Wilby R, Charles S, Zorita E, Timbal B, Whetton P, Mearns L (2004) Guidelines for use of climate scenarios developed from statistical downscaling methods. Supporting material of the intergovernmental panel on climate change, DDC of IPCC TGCIA

Winkler JA et al (2011) Climate scenario development and applications for local/regional climate change impact assessments: an overview for the non-climate scientist. Geogr Compass 5:275–300. doi:10.1111/j.1749-8198.2011.00425.x

Wu R, Kirtman BP (2005) Roles of Indian and Pacific Ocean air–sea coupling in tropical atmospheric variability. Clim Dyn 25:155–170

Wu R, Kirtman BP (2011) Caribbean Sea rainfall variability during the rainy season and relationship to the equatorial Pacific and tropical Atlantic SST. Clim Dyn 37:1533–1550