GiST, A MODEL FOR GENERATING SPATIAL-TEMPORAL DAILY RAINFALL DATA

Guillermo A. Baigorria[1*] and James W. Jones[1]

ABSTRACT

Weather generators are tools developed to create synthetic daily weather data over long periods of time. These tools have also been used for downscaling from monthly to seasonal forecasts produced by global and regional circulation models to daily values in order to provide inputs for crop and other environmental models. A major limitation of weather generators is that they do not take into account the spatial structure of weather and climate for a given region or watershed. This spatial correlation is important when one aggregates variables, for example, simulated crop yields or water resources, across a watershed or region. A method was developed to generate realizations of daily rainfall for multiple sites in an area while preserving the spatial and temporal patterns among sites. A two-step method generates rainfall events followed by rainfall amounts at sites where a generated rainfall event occurs. Generation of rainfall events was based on a two-state orthogonal Markov chain for discrete distributions. For generating rainfall amounts, a vector of random numbers ($r_{norm} \sim N[0,1]$) of order equal to the number of locations with rainfall events that were generated to occur in a specific day was matrix-multiplied by the corresponding reduced function of the correlation matrix to create correlated random numbers. To generate the final rainfall amounts, elements from the resulting vector of spatially-correlated random numbers were retransformed to a gamma distribution using cumulative probability functions calculated individually for each location. Values were next rescaled to rainfall amounts. Seven weather stations in North-Central Florida were selected, and a thousand replications of daily rainfall data were generated for this study. Rainfall events and amounts from the new method were compared to those from the WGEN point-based weather generator. The spatial structure in generated daily rainfall events and amounts closely matched the observed ones among all pairs of weather stations and other monthly rainfall statistics for each weather station. Correlation coefficient between observed and generated ($\rho_{o\text{-}g}$) joint probabilities that station pairs are both with rainfall was 0.996 and for both without rainfall was 0.991. The $\rho_{o\text{-}g}$ correlation among weather stations was 0.983 for rainfall amounts and was significant at the 0.01 probability level. Root mean square errors of correlation values ranged from 0.04 to 0.08 for rainfall events and from 0.01 to 0.09 for amounts.

Key words: Correlated random numbers, correlation matrix, Eigen decomposition, Markov chain, multi-site analysis, rainfall, stochastic downscaling, Toeplitz-Cholesky matrix, weather generator

[1] Department of Agricultural and Biological Engineering, University of Florida, Gainesville, FL
* Corresponding author e-mail: gbaigorr@ufl.edu

INTRODUCTION

Modeling rainfall in space and time is necessary to better understand soil erosion, runoff, and pollutant transport processes at the watershed level (Baigorria and Romero, 2007; Keener et al., 2007; Romero et al., 2007). It is also needed to translate categorical forecasts of El Niño (Podestá et al., 2002) or to downscale monthly rainfall forecasts from global or regional circulation models to daily rainfall inputs for crop simulation models (Baigorria, 2007; Hansen and Ines, 2005). Simulation of dry spell distribution (Baigorria et al., 2007b) and irrigation requirements for large areas (Romero et al., 2009) also have spatial implications as potentially applied to the establishment of water use limits among different sectors, such as, cities, agricultural lands, production of energy, industry, and conservation areas. Therefore, a simple and flexible approach was developed to generate synthetic rainfall data while taking into account the spatial and temporal correlation structure observed in the historical records.

Rainfall events in most point-based weather generators, such as WGEN (Richardson and Wright, 1984), LARS–WG (Racsko et al., 1991; Semenov et al. 1998) and CLIGEN (Nicks et al. 1995), are based on generating a random number from a uniform distribution that is then used with a two-state, first- or second-order Markov chain to create each day's rainfall event state (rain or no-rain). Weather generators also generate a random number from the probability distribution function of rainfall amounts, which is then rescaled according to the statistical parameters of specific weather stations (Racsko et al., 1991; Richardson and Wright, 1984).

Other spatial weather generators have been developed. Wilks (1998) used empirical relationships between pair-wise correlations of series of random numbers and its counterpart following two-state first-order Markov chains for generating rainfall events; whereas for rainfall amounts used a mixed exponential distribution for nonzero amounts using spatially correlated random numbers from a multivariate normal distribution. Based on Moran's *I* autocorrelation index (Moran, 1950), Khalili et al. (2006) developed another method using a spatial moving average process to generate spatially auto-correlated random numbers to feed a stochastic daily weather generator. Apipattanavis et al. (2007) used Markov chains to capture the spell statistics, while a *k*-nearest neighbor bootstrap resampling method captured the distributional and lag-dependence statistics of rainfall.

Other spatial weather generators were developed based on daily or monthly atmospheric circulation patterns. Some approaches were based on Monte-Carlo simulations or nearest-neighbor methods combined with resampling techniques from daily historical records of rainfall for multiple-site observations classified by season, circulation weather type, transition probabilities, and storm-rain cell properties, such as the relationships among storm duration, radius, and intensity (Burton et al., 2008; Cannon, 2008; Fowler et al., 2005; Wilby et al., 2003). Another approach used atmospheric circulation patterns to calibrate parameters of the probability distributions (Quian et al., 2002).

In general, these weather generators are either site-specific or conceptually complex and consequently difficult to implement. Another disadvantage is that they are based on pair-wise weather station relationships and do not take into consideration the spatial-temporal correlation among all weather stations in an area.

This research was conducted to help answer several questions related to the generation of synthetic rainfall events and amounts for the purposes of stochastic downscaling and modeling. Is it possible to reproduce the inter-annual variability of rainfall events and amounts? Is it possible to generate synthetic rainfall data that reproduce the temporal and spatial structure of daily observed rainfall? Is it possible to recreate the observed monthly spatial correlation patterns by generating daily spatially correlated rainfall events? The objective of this study is to design a

simple rainfall event and rainfall amount data generator capable of reproducing both the daily spatial correlations among weather stations as well as the monthly statistics of each individual weather station.

METHODS

One difficulty in randomly generating correlated occurrences of daily rainfall events over space and time is that this variable is not Gaussian, so the normal distribution and its associated methods cannot be used (Appendix).

The methodology presented here is based on the assumption of spatial-temporal covariance stationarity, which implies that the mean and autocovariance functions of a data series, as well as the spatial correlations among the data series, do not change through time for the period under consideration. From a temporal point of view, this allows one to characterize a time series of a given variable, such as rainfall, as a probability distribution at each weather station. From the spatial point of view, this allows one to characterize the spatial relationship between pairs of weather stations with a Pearson's correlation and among all the weather stations in a selected area with Pearson's correlation matrix.

As in point-specific weather generators, only daily historical records of rainfall are needed as input for the approach described in this paper because the model calibrates itself according to the observed statistics within and among locations. An orthogonal Markov chain for discrete distributions is used to generate rainfall events; an extension of the Toeplitz-Cholesky (Benoît 1924;Taussky and Todd, 2006) factorization and Eigen decomposition matrices is presented for the generation of rainfall amount data. Rainfall events are generated location by location; therefore, the Euclidean N-correlation distance is defined in order to establish an order for the locations in which to be generated.

It is important to point out that the proposed method will generate rainfall events and amounts for weather station locations rather than a continuous interpolation over space; however, spatially and temporally generated data across a given region should provide better inputs for interpolating downscaled scenarios using geostatistical techniques.

Generation of rainfall events: Two-state orthogonal Markov chain for discrete distributions

There are two main steps in generating rainfall events. The first one is to calculate all parameters and initial conditions. This step includes (a) the calculation of the Pearson's correlation matrix, (b) the calculation of the Euclidean N-correlation distance, (c) the calculation of the two-state orthogonal Markov transitional probabilities, and (d) the generation of the spatially correlated total number of monthly rainfall events at each location. This last entry constitutes target values used in the rainfall event generation only at the initial conditions. The second step, the spatially-temporally correlated rainfall event generation, includes two processes: (a) resampling and iteratively ordering the total block of daily generated values in a month for the two most associated location (closest Euclidean N-correlation distance), and (b) the use of the two-state orthogonal Markov transitional probabilities to generate rainfall events for the remaining locations. All these calculations are made independently for each month.

Parameterization and initial conditions

a. Pearson's correlation (ρ_{ij}): Calculations use the following equation (see Table 1 for variable definitions):

$$\rho_{ij} = \frac{1}{\eta} \frac{\sum_{t=1}^{\eta} (\chi_{it} - \mu_i)(\chi_{jt} - \mu_j)}{\sigma_i \sigma_j}. \quad [1]$$

b. Euclidean N-correlation distance (G_i): Spatially-temporally correlated rainfall events for several locations are calculated one location at a time after the first two most associated locations. To find the best order for the generation process, the Euclidean N-correlation distance (G_i) is then defined. Based on correlation measurements, G_i is the degree of association of a given variable from a given location with the same or other variables from all remaining locations under study:

$$G_i = \sum_{j=1}^{n-1} \left(1 - \left|\rho_{i,j}\right|\right)^2 \quad [2]$$

By ranking in ascending order the G_i values obtained from all locations, an ordered list of locations is generated in which the first location is the one most correlated with all remaining locations, whereas the last location in the list is the most independent. Because the spatial correlation patterns of rainfall events change monthly according to the season (for instance, from a convective rainfall season to a frontal rainfall season), G_i must be calculated for homogenous periods of time, for example, by month or season. The rank list can differ across these periods.

***c. Two-state orthogonal Markov's transitional probabilities*:** These transition probabilities are conditional probabilities for the state of a location i at time t, e.g., whether rainfall will occur in a given day, given: a) the state at time $t-1$, that is, whether or not rainfall occurred the previous day at the same location i; and b) the state at time t, that is, whether or not rainfall occurred in the same given day in another two locations (j,k) that are to some degree correlated:

$$\begin{cases} P_{1|0,0|0} = \Pr\left[\chi_{it} = 1 \middle| \chi_{jt} = 0 \wedge \chi_{kt} = 0 \middle| \chi_{i(t-1)} = 0\right] \\ P_{1|0,0|1} = \Pr\left[\chi_{it} = 1 \middle| \chi_{jt} = 0 \wedge \chi_{kt} = 0 \middle| \chi_{i(t-1)} = 1\right] \\ P_{1|0,1|0} = \Pr\left[\chi_{it} = 1 \middle| \chi_{jt} = 0 \wedge \chi_{kt} = 1 \middle| \chi_{i(t-1)} = 0\right] \\ P_{1|0,1|1} = \Pr\left[\chi_{it} = 1 \middle| \chi_{jt} = 0 \wedge \chi_{kt} = 1 \middle| \chi_{i(t-1)} = 1\right] \\ P_{1|1,0|0} = \Pr\left[\chi_{it} = 1 \middle| \chi_{jt} = 1 \wedge \chi_{kt} = 0 \middle| \chi_{i(t-1)} = 0\right] \\ P_{1|1,0|1} = \Pr\left[\chi_{it} = 1 \middle| \chi_{jt} = 1 \wedge \chi_{kt} = 0 \middle| \chi_{i(t-1)} = 1\right] \\ P_{1|1,1|0} = \Pr\left[\chi_{it} = 1 \middle| \chi_{jt} = 1 \wedge \chi_{kt} = 1 \middle| \chi_{i(t-1)} = 0\right] \\ P_{1|1,1|1} = \Pr\left[\chi_{it} = 1 \middle| \chi_{jt} = 1 \wedge \chi_{kt} = 1 \middle| \chi_{i(t-1)} = 1\right] \end{cases} \quad [3]$$

The two-state orthogonal Markov chain for discrete distributions in the case of rainfall events can have only two states at a given time, t, that is, rain or non-rain; therefore, the transitional probabilities of having a non-rainfall occurrence and the transitional probabilities of having a rainfall event in Equation [3] must sum to 1 (for instance in the case of equation 3.a: $P_{1|0,0|0} + P_{0|0,0|0} = 1$).

d. Total monthly number of rainfall events: The multi-annual variability of monthly number of rainy days tends to follow a Gaussian distribution $\sim N\left[\overline{T}_m, \sigma_m\right]$ in non-arid regions.

Therefore, Equation [4] was used to generate these data at each location (variables are defined in Table 1).

$$\hat{T}_m = \overline{T}_m \pm \sigma_m r^*_{norm} \quad [4]$$

Table 1: Variable descriptions.

Symbol	Definition	Units
i, j, k, l	Locations	unitless
α, β	Shape and Scale parameters of the gamma function (Γ)	unitless
e	Number of locations with generated rainfall events ($e \le n$)	unitless
$[Eig]$	Eigen decomposition matrix	unitless
G_i	Euclidean N-correlation distance	unitless
n	Total number of locations	unitless
η	Total number of pair-wise daily observations	unitless
P_{01}, P_{11}	Markov transitional probabilities of dry-wet and wet-wet days	fraction
$P_{1\|0,0\|0}$	Two-state orthogonal Markov transitional probabilities	fraction
$\vec{r}^*_{gam}$	Vector of spatially correlated random numbers following a Gamma distribution $\Gamma[\alpha]$	unitless
$\vec{r}_{norm}$	Vector of random numbers following a Gaussian distribution $N[0,1]$	unitless
$\vec{r}^*_{norm}$	Vector of spatially correlated random numbers following a Gaussian distribution $N[0,1]$	unitless
$\vec{r}_{unif}$	Vector of random numbers following a Uniform distribution $U[-1,1]$	unitless
$\vec{r}^{\,\psi}$	Temporal vector from matrix-multiplying a vector of random numbers by $[TC]$ or $[Eig]$	unitless
$\hat{R}_m$	Expected monthly rainfall amount	mm
$\rho_{i,j}$	Pearson's correlation between two locations	unitless
$[\rho_T]$	Correlation matrix of monthly number of rainfall events	unitless
$[\rho_R]$	Correlation matrix of daily rainfall amounts	unitless
σ_i, σ_j	Standard deviation of daily observations	fraction, mm
σ_m	Multi-annual standard deviation of monthly number of rainy days in month *m*	days
$[TC]$	Toeplitz-Cholesky factorization matrix	unitless
$\hat{T}_m$	Expected total number of rainy days in month *m*	days
$\overline{T}_m$	Multi-annual average of monthly number of rainy days in month *m*	days
μ_i, μ_j	Mean of daily values for locations i, j	fraction, mm
χ_{it}, χ_{jt}	Pair-wise observations on day t for locations i, j	Boolean, mm

Taking into consideration the spatial correlation among locations, the vector of random numbers ($\vec{r}_{norm}$) must be transformed to a vector containing correlated random numbers r^*_{norm} that follows the observed correlation values. To obtain r^*_{norm}, a vector of random numbers, $r_{norm} \sim N[0,1]$, with a number of elements equal to the number of locations is matrix-multiplied by the reduced function of the Pearson's correlation matrix. The function was obtained by the matrix summation between the scalar multiplication of 0.5 by the Pearson's correlation matrix of the monthly number of rainfall events ($[\rho_T]$) and the Toeplitz-Cholesky factorization matrix ($[TC]$; Appendix A) obtained from the same Pearson's correlation matrix. The correlation matrices are calculated monthly using multi-annual variability.

$$\vec{r}_{norm}\left(\frac{[\rho_T]}{2}+[TC]\right)=\vec{r}^{\psi} \qquad [5]$$

Afterward, $\vec{r}^{\psi}$ is range-rescaled according to the percentage of the population included (i.e., from -1 to 1 for 68% or -1.96 to 1.96 for 95%) to finally obtain $\vec{r}^*_{norm}$.

Data generation

To apply the two-state orthogonal Markov transitional probabilities, at least two locations with spatial-temporal rainfall events that were previously generated are needed. Afterward, the two-state orthogonal Markov transitional probabilities can be applied to the remaining weather stations.

***a. Generating rainfall events for the two core locations*:** To generate spatially-temporally correlated rainfall events for the first two locations, a property of the Pearson's correlation for discrete distributions is applied. The two first locations are selected after ascend-ranking the Euclidean N-correlation distance (G_i) values calculated with Equation [2]. In the case of rainfall events for two locations, there are only four discrete possible combinations (i.e., spatial rainfall event patterns):

$$\chi_{it}=0 \wedge \chi_{jt}=0$$
$$\chi_{it}=0 \wedge \chi_{jt}=1$$
$$\chi_{it}=1 \wedge \chi_{jt}=1$$
$$\chi_{it}=1 \wedge \chi_{jt}=0$$

Therefore, Pearson's correlation becomes a function of the fraction of times each combination is observed in the historical record.

$$\rho_{ij}=f\left(\frac{N_{\substack{\chi_{it}=0\\ \chi_{jt}=0}}}{\eta},\frac{N_{\substack{\chi_{it}=1\\ \chi_{jt}=0}}}{\eta},\frac{N_{\substack{\chi_{it}=0\\ \chi_{jt}=1}}}{\eta},\frac{N_{\substack{\chi_{it}=1\\ \chi_{jt}=1}}}{\eta}\right) \qquad [6]$$

Because the spatially correlated number of rainy days per month for these two locations is already estimated with equations [4] and [5], the problem becomes a two-step resampling problem. Resampling is based first on the number of rainfall events at each location in order to best fit the percentages of spatial rainfall event patterns. This process will produce just four blocks of time (i.e., numbers of days) that equal the percentage of spatial rainfall event patterns for Equation [6]. This will reproduce the observed spatial correlations between both locations for the specific month. Second, resampling is completed by sorting the previously generated daily spatial rainfall event patterns in order to best fit the Markov transitional probability patterns. Iteratively, the completely random sorting process selects the best temporal order of the spatial

rainfall event patterns by reducing the root mean square error between the generated and observed Markov transitional probabilities for both locations at the same time (P_{01}, P_{11} ; see Table 1 for variable descriptions).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{2}\left[\left(\hat{P}_{01} - P_{01}\right)_i^2 + \left(\hat{P}_{11} - P_{11}\right)_i^2\right]}{4}}$$

***b. Nearest Euclidean 3-correlation neighbor assimilation*:** As opposed to the previous step, where the daily values for both locations are produced by resampling and iteratively ordering the total block of daily generated values within a month, in this step each day is generated individually based on the two-state orthogonal Markov transitional probabilities. It is possible to extend the resampling and iterative ordering process for three locations at the same time; however, the process requires more time and more computational resources than using the two-state orthogonal Markov transitional probabilities.

The order for generating rainfall events in the remaining locations followed the ranked G_i list. For each new location, two other previously generated locations are selected for having shown the highest Pearson's correlation values with the location being generated based on historical records. These two locations are used as a reference to apply the two-state orthogonal Markov transitional probabilities calculated using Equation [3]. Then, in order to assign a rainfall event for each new generated day, a random number from the Uniform distribution ($r_{unif} \sim U[0,1]$) is compared to the corresponding two-state orthogonal Markov transitional probability of a rainfall event.

Generation of rainfall amounts

***a. Pearson's correlation (ρ_{ij})*:** Calculations are performed using Equation [1] after transforming all daily rainfall amounts from Gamma to Gaussian distribution. This process is performed to avoid the problems associated from using Pearson's correlations in non-Gaussian distributions. Daily pair-wise observations without rainfall events, that is, registering zero at both locations, are removed from the calculations to avoid overestimation of ρ_{ij} . Previous results can be found in Baigorria et al. (2007a) for the same study area regarding different spatial correlation patterns of daily rainfall events and amounts as well as monthly total rainfall amounts and number of rainy days per months.

***b. Parameterization*:** Previous studies have determined that rainfall in the region tends to follow a 2-parameter gamma distribution (Baigorria et al., 2007b). A gamma distribution was implemented here to generate rainfall amounts; however, the proposed spatial rainfall data generation method is flexible enough to allow for the implementation of other probability distributions functions. For the application of the 2-parameter gamma distribution, both shape (α) and scale (β) parameters are calculated for each month and weather station based on historical records.

c. Rainfall amount data generation: The temporal structure of rainfall amounts was already generated with rainfall events; therefore, the generation of spatially-structured rainfall amounts is applied only to locations where a rainfall event was generated to occur. Rainfall amounts for each location are generated using:

$$\hat{R}_m = r_{gam}^{*}\beta_m \ln\left\{\Gamma\left(\alpha_m\right)\right\}$$

The vector of correlated random numbers from a gamma distribution $\left(\vec{r}_{gam}^{*} \sim \Gamma[\alpha]\right)$ is obtained from a vector with elements randomly generated from the Gaussian distribution,

$(\vec{r}_{norm} \sim N[0,1])$, which was in turn matrix-multiplied by the reduced function of the correlated matrix (Equation 5). In this case, the Pearson's correlation matrix was calculated from the transformed daily rainfall amounts $[\rho_R]$. The size of $([\rho_R]/2+[TC])$ and the random vector's element number (e) are modified daily according to the number of locations where rainfall events were generated to occur. The resulting temporary correlated vector, which follows a Gaussian distribution, is next transformed into a Gamma distribution with values ranging from 0 to 1 by using cumulative probability functions.

Case study evaluation

Study area

The case study was conducted in the North Central region of the state of Florida located in the southeastern USA between 30°16' N, 83°10' W and 28°37' N, 81°28' W (Figure 1). Annual rainfall climatology ranges from 1230 to 1460 mm among the seven selected weather stations. Rainfall occurs throughout the year and is caused by three different atmospheric processes. During most spring and summer months, rainfall occurs mainly by convective processes and tropical storms. During the convective rainy season, spatial correlations of rainfall

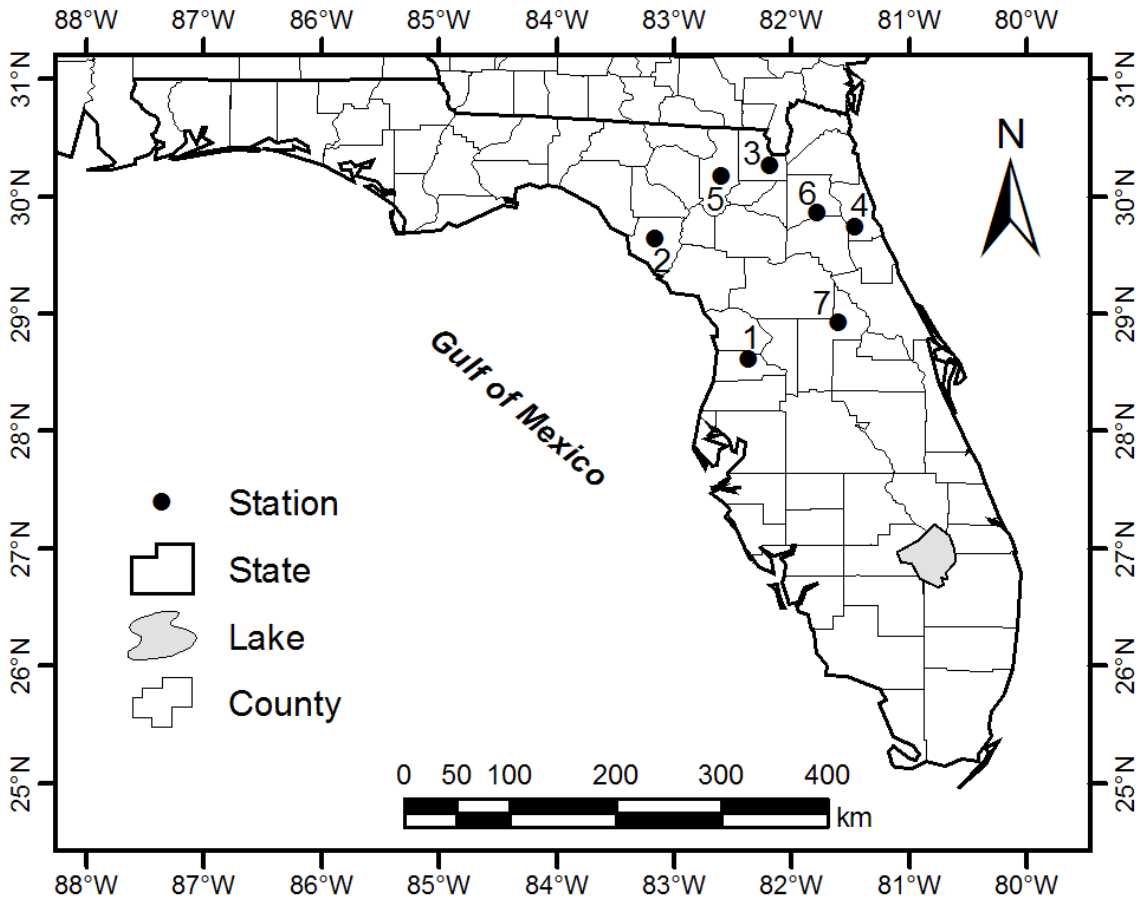

Fig. 1. Map of the study area and location of the weather stations used in the study: 1) 1046 = Brooksville, Hernando Co.; 2) 2008 = Cross City, Dixie Co.: 3) 3470 = Glen, Baker Co.; 4) 3874 = Hastings, St. Johns Co.; 5) 4731 = Lake City, Columbia Co.; 6) 5076 = Lisbon, Lake Co.; and 7) 5973 = Mountain Lake, Polk Co.

among weather stations are characterized by concentric patterns in which correlations decrease rapidly over short distances (Baigorria et al., 2007a). During most fall and winter months, rainfall occurs mainly by frontal storms coming through the northwestern USA. During this frontal rainy season, spatial correlations among rainfall are characterized by a wide pattern in a northeast-southwest direction, parallel to the usual weather fronts (Baigorria et al., 2007a).

Daily rainfall data from the seven weather stations from 1974 to 2004 were obtained from the NOAA/NWS/National Climate Data Center [http://www.ncdc.noaa.gov]. A 30-year period was selected to avoid the effects of longer term temporal climatic shifts detected in the study area (Baigorria et al., 2007a). Monthly rainfall climatology values for these weather stations are shown in Table 2, whereas monthly climatic Pearson's correlations calculated using daily observations of rainfall events and amounts among the weather stations are shown in Table 3.

Validation

Two methods were compared: a) the two-state first-order Markov chain approach, in which only the temporal structure is taken into account (this method is used by most point-specific weather generators); and b) a new method, in which both spatial and temporal structures are taken into account simultaneously.

For validation, one-thousand replications of synthetic daily rainfall events and amounts were generated for the seven locations using the standard point-site weather generator WGEN and the proposed multi-site Geo-spatial temporal weather generator (GiST). From the daily generated values, Pearson's correlation matrices for each month were calculated by both methods. Statistical and graphical comparisons were perform by comparing generated correlations with the corresponding Pearson's correlations matrices calculated using observed rainfall event and amounts from the historical record. Root mean square errors (RMSEs) for each month were also calculated by comparing the generated and observed Pearson's correlations of daily rainfall events and amounts for both methods. Two-state first-order Markov transitional probabilities were also calculated for the one-thousand replications of generated rainfall events and amounts. Results from both methods were compared with their corresponding transitional probabilities calculated based on the historical record. A region-wide analysis of days without rainfall at any station was performed for observed and generated rainfall events.

RESULTS AND DISCUSSION

Generation of rainfall events

Comparisons between observed and generated Pearson's correlations of daily rainfall events among all pairs of locations for each month are shown in Figure 2. As expected, because WGEN was not designed to consider spatial correlations, the point-based weather generator did not reproduce any observed spatial correlation of rainfall events between pairs of locations ($\rho = -0.0820^{ns}$). In comparison, the GiST rainfall generator reproduced the monthly observed correlations with $\rho = 0.940$, which is statistically significant at the 0.01 probability level. It is possible to increase the number of locations used in the orthogonal Markov chains, but this possibility is constrained by the number of parallel observations in all the locations. For our seven locations, for example, the regional percentage of missing values ranged from 46% to 68%.

Figure 3 shows the comparison between observed and generated full joint distributions of simultaneous rainfall events across all weather stations using WGEN and GiST. The correlation coefficients between observed and generated ($\rho_{o\text{-}g}$) joint probabilities that stations

pairs are both with rainfall were 0.939 for WGEN and and 0.996 for GiST. The $\rho_{o\text{-}g}$ joint probabilities that stations pairs are for both without rainfall were 0.808 for WGEN and 0.991 for GiST.

Figure 4 shows the monthly root mean square (RMSE) values for both methods. WGEN produced the highest RMSE in comparison with GiST, especially during the frontal rainy season from November through April. As a percentage of the observed monthly mean correlations, RMSE ranged from 98 to 104% for WGEN and from 7 to 26% for GiST.

Table 2: Monthly climatology from 1974 through 2004 for: a) number of rainy days, and b) total rainfall amounts for 7 Florida weather stations: 1046 = Brooksville, Hernando Co.; 2008 = Cross City, Dixie Co.; 3470 = Glen, Baker Co.; 3874 = Hastings, St. Johns Co.; 4731 = Lake City, Columbia Co.; 5076 = Lisbon, Lake Co.; and 5973 = Mountain Lake, Polk Co..

a. Number of rainy days

Month	**Weather station**						
	1046	**2008**	**3470**	**3874**	**4731**	**5076**	**5973**
Jan	7.8	8.4	8.4	8.1	10.6	8.1	6.4
Feb	6.5	7.4	6.8	8.1	8.4	7.1	6.4
Mar	7.4	7.6	7.3	8.0	9.1	7.3	6.7
Apr	5.0	6.0	5.3	5.2	6.3	5.4	5.4
May	6.8	6.0	6.4	6.5	7.7	7.7	7.3
Jun	13.4	12.1	12.3	14.0	13.3	13.7	14.3
Jul	15.8	15.4	13.1	14.2	15.3	15.7	15.5
Aug	16.0	16.0	13.3	15.8	15.4	15.8	15.2
Sep	12.0	11.2	10.7	13.4	11.2	13.5	12.3
Oct	6.4	5.3	5.6	10.2	6.8	7.8	7.5
Nov	5.4	5.5	5.4	7.7	7.1	7.8	5.8
Dec	7.2	7.2	6.2	7.7	9.2	8.3	6.7

b. Rainfall amounts, mm

Month	**Weather station**						
	1046	**2008**	**3470**	**3874**	**4731**	**5076**	**5973**
Jan	81	105	103	77	112	81	59
Feb	78	78	82	91	90	71	58
Mar	99	113	110	114	131	105	81
Apr	58	88	75	73	74	67	47
May	77	71	75	77	90	99	92
Jun	187	168	166	194	172	172	202
Jul	192	233	164	157	180	144	177
Aug	205	244	174	190	187	161	171
Sep	171	152	151	229	143	165	165
Oct	63	74	85	99	74	62	64
Nov	53	57	56	67	59	60	57
Dec	62	78	64	74	71	66	65

Table 3: Pearson's correlations of daily rainfall events (shaded) and rainfall amounts transformed from gamma to Gaussian distributions (unshaded) among weather stations for the months of January and July. 1046 = Brooksville, Hernando Co.; 2008 = Cross City, Dixie Co.; 3470 = Glen, Baker Co.; 3874 = Hastings, St. Johns Co.; 4731 = Lake City, Columbia Co.; 5076 = Lisbon, Lake Co.; and 5973 = Mountain Lake, Polk Co.. Pearson's correlations were calculated using pair-wise data with at least one weather station registering rainfall.

January	**1046**	**2008**	**3470**	**3874**	**4731**	**5076**	**5973**
1046	1	0.507	0.540	0.568	0.505	0.505	0.543
2008	0.529	1	0.727	0.654	0.698	0.535	0.484
3470	0.428	0.740	1	0.703	0.717	0.577	0.504
3874	0.504	0.618	0.618	1	0.624	0.607	0.503
4731	0.485	0.824	0.806	0.587	1	0.493	0.438
5076	0.645	0.391	0.361	0.550	0.407	1	0.544
5973	0.582	0.460	0.348	0.397	0.369	0.539	1

July	**1046**	**2008**	**3470**	**3874**	**4731**	**5076**	**5973**
1046	1	0.228	0.190	0.193	0.238	0.277	0.184
2008	0.044	1	0.310	0.254	0.372	0.254	0.158
3470	0.145	0.137	1	0.277	0.360	0.234	0.143
3874	0.003	0.111	0.016	1	0.314	0.311	0.190
4731	0.126	0.227	0.212	0.079	1	0.244	0.163
5076	0.125	0.144	0.071	-0.011	0.086	1	0.275
5973	0.070	0.015	0.071	0.050	0.096	0.047	1

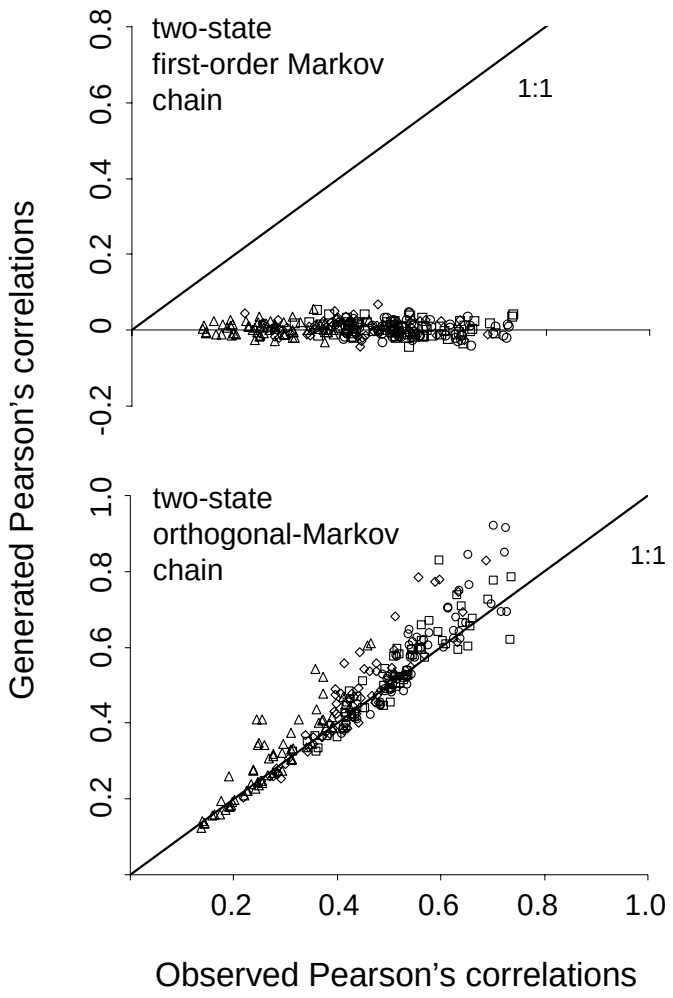

Fig. 2. Comparison between observed and generated daily Pearson's correlations of rainfall events for each month among all weather stations. (o) December-January-February, (□) March-April-May, (Δ) June-July-August, and (◊) September-October-November.

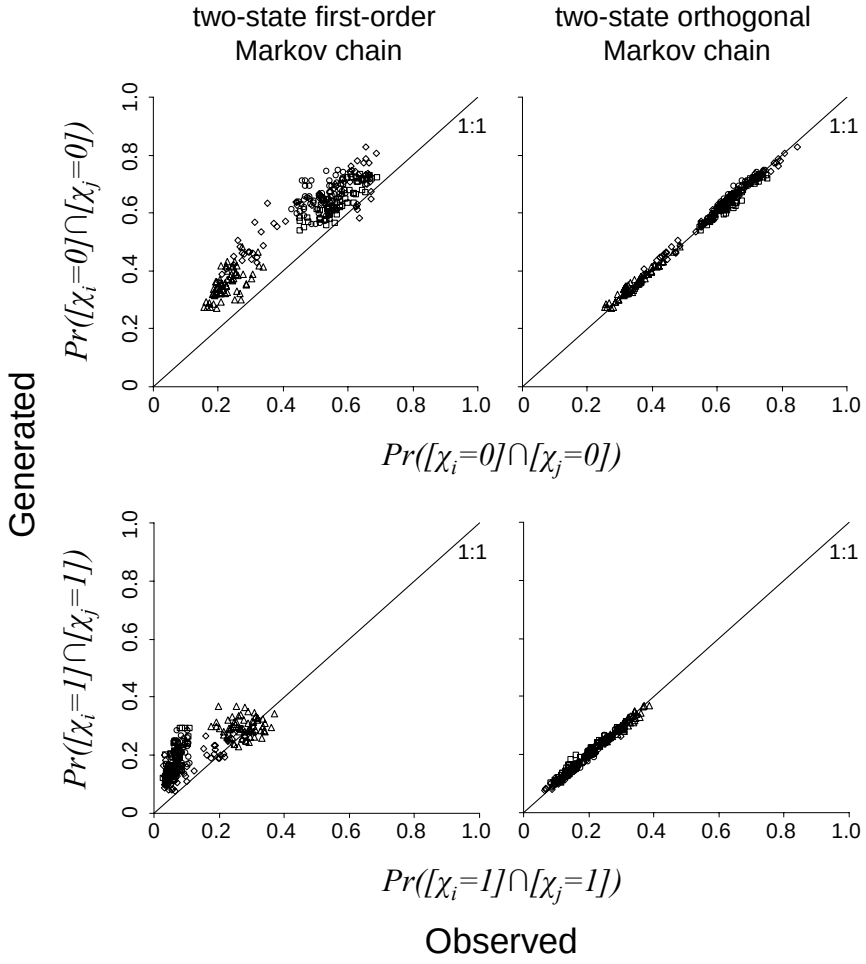

Fig. 3. Observed versus generated joint probabilities that station pairs are both with rainfall (top graphs) and without rainfall (bottom graphs) occurrences on a given day for all combinations of station pairs and all 12 months. (o) December-January-February, (□) March-April-May, (Δ) June-July-August, and (◊) September-October-November.

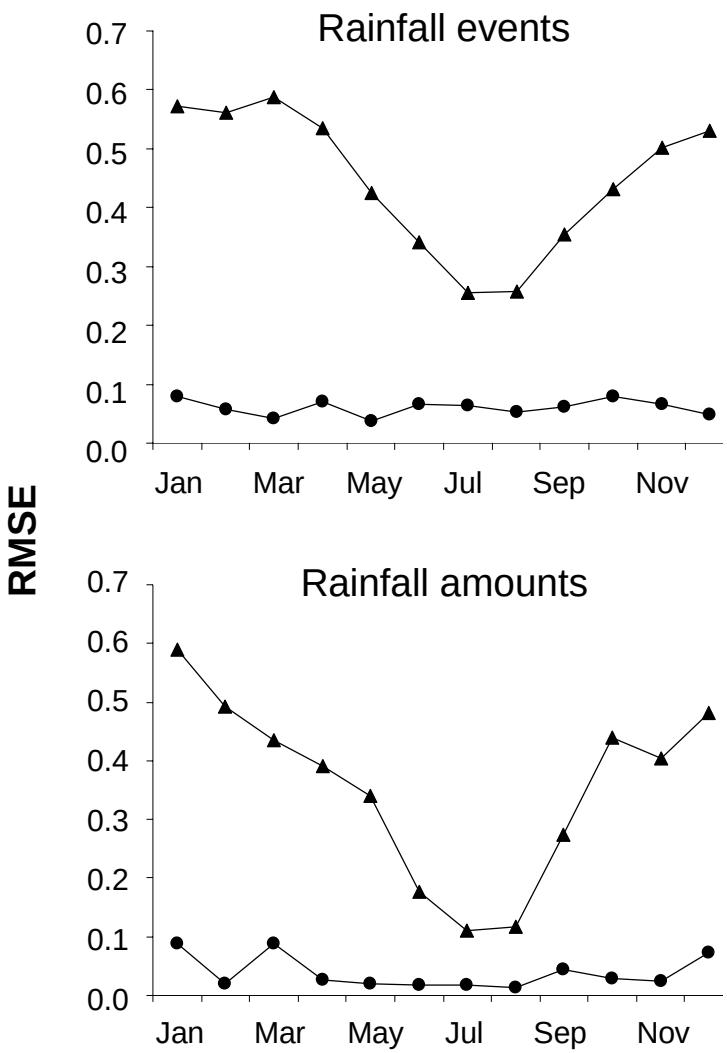

Fig. 4. Comparison of the monthly variation of the root mean square error (RMSE) between weather generators.

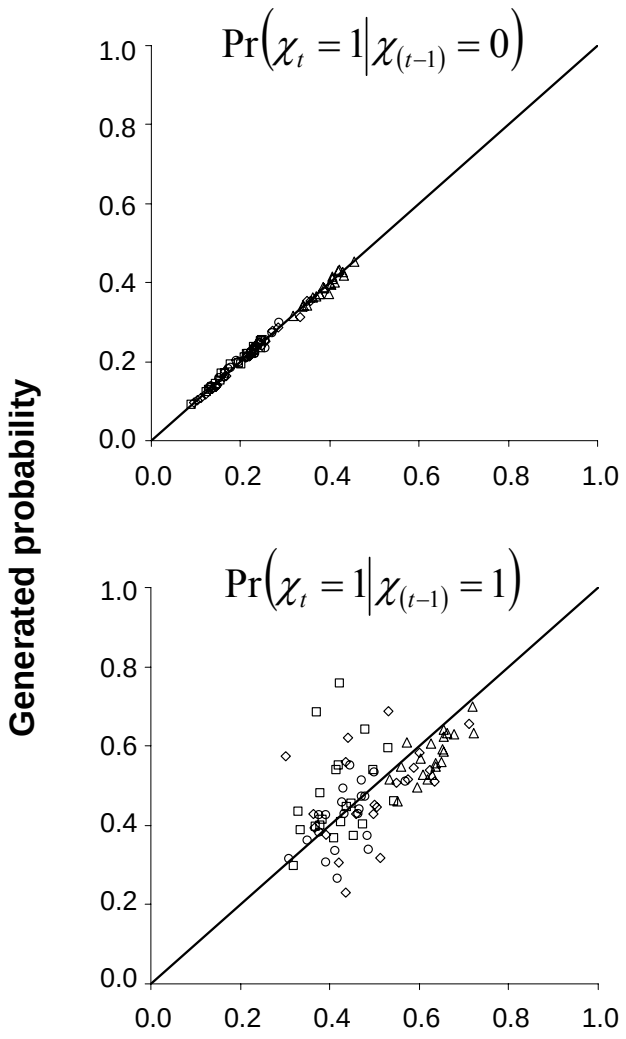

Fig. 5. Comparison between observed and generated Markov transitional probabilities for a rainy day following a non-rainy day (top) or for a rainy day following a rainy day (bottom).

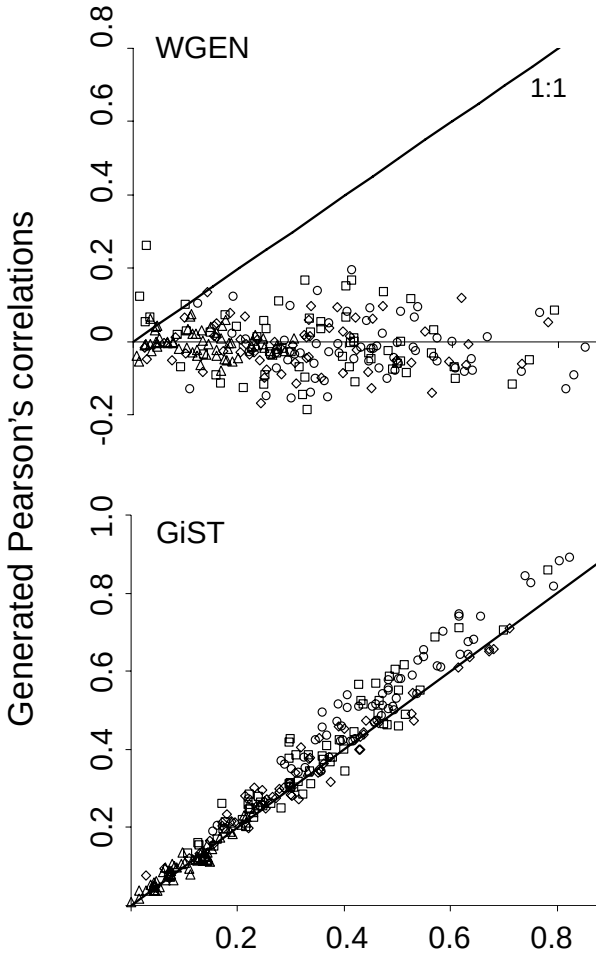

Fig. 6. Comparison between observed and generated daily Pearson's correlations of transformed from gamma to Gaussian distributions for rainfall amounts for each month among all weather stations. (o) December-January-February, (□) March-April-May, (Δ) June-July-August, and (◊) September-October-November.

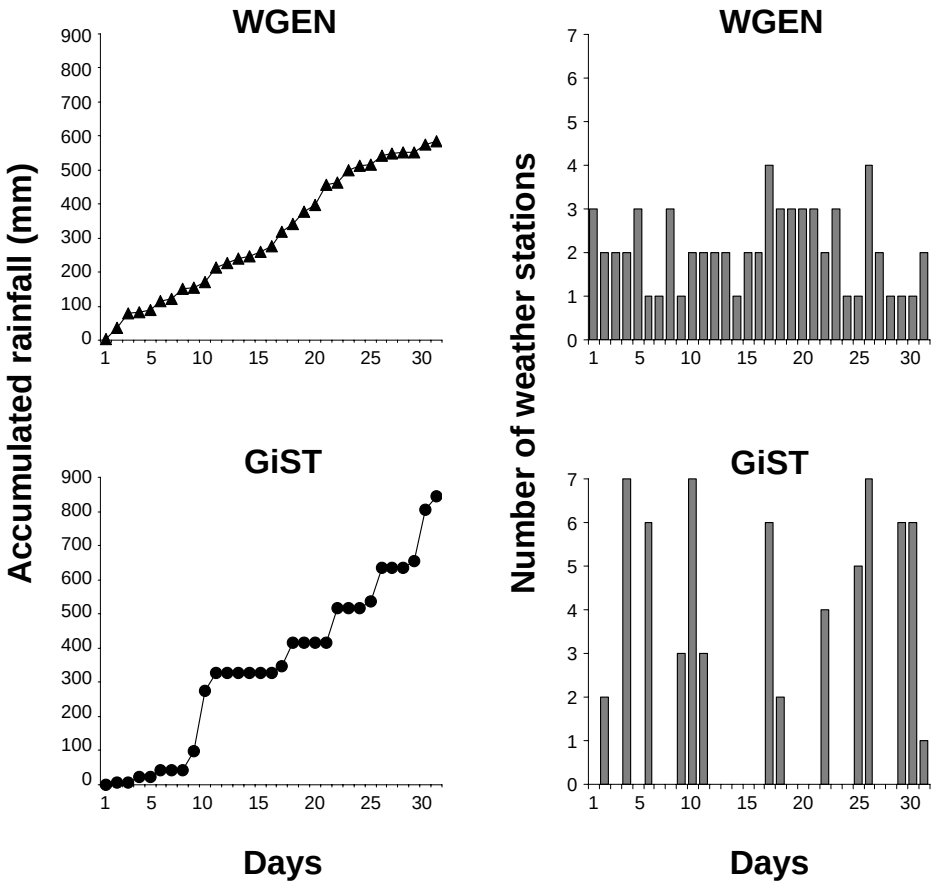

Fig. 7. The aggregate effect of using point-specific and spatial-temporal weather generators over dry and wet spell distributions. Figures in the left side are 31-day accumulations of generated rainfall across the 7 weather stations. Figures in the right side are the daily number of weather stations registering rainfall events.

Comparisons of the two-state first-order Markov transitional probabilities from observed and generated daily rainfall events are shown in Figure 5. Calculated using WGEN, P_{01} was 0.979, whereas for GiST, this value was 0.997. Meanwhile, P_{11} calculated from WGEN was 0.883, whereas it was 0.596 based on GiST. Transitional probabilities of wet-wet days showed higher degradation than the transitional probabilities of dry-wet days (Figure 5). This degradation was observed for both methods.

Generation of rainfall amounts

Figure 6 compares observed and generated Pearson's correlations of transformed (from Gamma to Gaussian distributions) daily rainfall amounts among all pairs of locations. Spatial correlation (ρ = -0.1499) levels for WGEN were small, as expected. As a percentage of the observed monthly mean correlations, RMSE ranged from 103 to 132% using the point-based weather generator, whereas RMSE using the GiST generator ranged from 4% to 21%. GiST did not show RMSE seasonal variability in comparison with the point-based method. A correlation of 0.983 was obtained between observed and generated correlations of transformed rainfall amounts between pairs of locations; this correlation was significant at the 0.01 probability level.

Implications of reproducing the region-wide number of days without rainfall at any location are shown in Figure 7, which shows the first 31 days from a randomly selected year for all seven locations. The region-wide accumulated rainfall generated by the point-based weather generator showed a linear trend, because every day is rainy in at least one location at the time. Using the GiST weather generator, the region-wide accumulated synthetic rainfall showed a stair-shaped trend, because dry and wet spells are well-defined in the daily sequence. Hydrologists studying floods and soil scientists studying soil erosion, for example, should be interested in the shifts in watershed accumulated rainfalls; alternatively, agronomists studying the regional effects of droughts over crops would be interested in the regional dry spell lengths.

CONCLUSIONS

The proposed algorithms reproduce the main statistics of the observed historical record of each individual weather station as well as the spatial correlation between pairs of weather stations. Pearson's correlations between observed and generated joint probabilities were 0.996 for pair-wise weather stations with rainfall events and 0.991without rainfall events. Pearson's correlation between observed and generated pairs of weather stations was 0.983 for rainfall amounts, and was statistically significant at the 0.01 probability level. The proposed methodology reproduced the two-state first-order Markov transitional probabilities with statistical significance as well as the regional-wide number of days without rainfall at any location.

Due to its simplicity and flexibility, the proposed methodology can be applied in part or in whole beyond rainfall data and even beyond meteorological applications to any kind of information flow that follows a spatial and temporal structure. Further studies incorporating cross-correlations with other variable, such as incoming solar radiation and maximum and minimum temperatures, are needed. Inclusion of cross-correlations with other variables will allow for stochastic downscaling of global and regional circulation model forecasts; this should also enhance the application of this method to crop and environmental modeling at watershed and regional levels.

Possible applications of these methodologies may involve the estimation of rainfall values missing from historical records, the downscaling of GCM/RCM's with or without including bias corrections, regional analyses of crop production using crop modeling techniques, hydrological analyses from watershed to regional levels, soil erosion from watershed to regional scales, the creation of spatial and temporal structured downscaling climatic change scenarios, and regional risk management.

REFERENCES

Apipattanavis S, Podestá G, Rajagopalan B, Katz R. 2007. A semiparametric multivariate and multisite weather generator. Water Resources Research 43:1-19.

Baigorria GA. 2007. Assessing the use of seasonal-climate forecasts to support farmers in the Andean Highlands. p. 99-110. In: M.V.K. Sivakumar and J.W. Hansen. (Eds). Climate prediction and agriculture: Advances and Challenges. Springer Berlin/Heidelberg.

Baigorria GA, Jones JW, O'Brien JJ. 2007a. Understanding rainfall spatial variability in the southeast USA at different time scales. International Journal of Climatology 27:749-760.

Baigorria GA, Jones JW, O'Brien JJ. 2008. Potential predictability of crop yields using an ensemble climate forecast by a regional circulation model. Agricultural and Forest Meteorology 148:1353-1361.

Baigorria GA, Jones JW, Shin DW, Mishra A, O'Brien JJ. 2007b. Assessing uncertainties in crop model simulations using daily bias-corrected Regional Circulation Model outputs. Climate Research 34:211-222.

Baigorria GA, Romero CC. 2007. Assessment of erosion hotspot in a watershed: integrating the WEPP model and GIS in a case study in the Peruvian Andes. Environmental Modeling and Software 22:1175-1183.

Benoît C. (Procédé du Commandant A L Cholesky) 1924. Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéaires en nombre inférieur à la résolution d'un système défini d'équations linéaires. Bulletin Géodésique 2:67-77.

Brezinski C. 2006. The life and work of André Cholesky. Numerical Algorithms 41:197-202.

Burton A, Kilsby CG, Fowler HJ, Cowpertwait PSP, O'Connell PE. 2008. RainSim: A spatial-temporal stochastic rainfall modelling system. Environmental Modelling & Software 23:1356-1369.

Cannon A. 2008. Probabilistic multisite precipitation downscaling by an expanded Bernoulli-Gamma density network. Journal of Hydrometeorology 9:1248-1300.

Fowler HJ, Kilsby CG, O'Connell PE, Burton A. 2005. A weather-type conditioned multi-site stochastic rainfall model for the generation of scenarios of climatic variability and change. Journal of Hydrology 308:50-66.

Hansen JW, Ines AVM. 2005. Stochastic disaggregation of monthly rainfall data for crop simulation studies. Agricultural and Forest Meteorology 131:233-246.

Hilbert D. 1904. Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen. Nachrichten von der Königliche Gesellschaft der Wissenschaften zu Göttingen, Mathematisch – Physikalische Klasse, pp 49-91.

Hotelling H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24:417-441.

Iman RL, Conover WJ. 1982. A distribution-free approach to inducing rank correlation among input variables. Communications in Statistics Simulation and Computation 11:311-334.

Keener VW, Ingram KT, Jacobson B, Jones JW. 2007. Effects of El Niño/Southern Oscillation on simulated phosphorus loadings in south Florida. Transactions of the ASABE 50:2081-2089.

Khalili M, Leconte R, Brissette F. 2006. Stochastic multisite generation of daily precipitation data using spatial autocorrelation. Journal of Hydrometeorology 8:396-412.

Moran PAP.1950. Notes on continuous stochastic phenomena. Biometrika 37:17-23.

Nicks AD, Lane LJ, Gander GA. 1995. Weather Generator. In: Flanagan DC and Nearing MA (Eds). USDA-Water Erosion Prediction Project (WEPP). WEPP users summary. NSERL Report No. 10. USDA-ARS National Soil Erosion Research Laboratory, West Lafayette, IN.

Podestá G, Letson D, Messina C, Royce F, Ferreyra RA, Jones JW, Hansen JW, Llovet I, Grondona M, O'Brien JJ. 2002. Use of ENSO-related climate information in agricultural decision making in Argentina: a pilot experience. Agricultural Systems 74:371-392.

Qian B, Corte-Real J, Xu H. 2002. Multisite stochastic weather models for impact studies. International Journal of Climatology 22:1377-1397.

Racsko P, Szeidl L, Semenov MA. 1991. A serial approach to local stochastic weather models. Ecological Modelling 57:27-41.

Richardson CW, Wright DA, 1984. WGEN, A model for generating daily weather variables, USDA ARS Bulletin No. ARS-8. Washington, DC, USA: Government Printing Office.

Romero CC, Baigorria GA, Stroosnijder L. 2007. Changes of erosive rainfall for El Niño and La Niña years in the northern Andean Highlands of Peru: The case of La Encañada watershed. Climatic Change 85:343-356.

Romero CC, Dukes MD, Baigorria GA, Cohen R. 2009. Comparing theoretical irrigation requirement and actual irrigation for citrus in Florida. Agricultural Water Management 96:473-483.

Scheuer EM, Stoller DS. 1962. On the generation of normal random vectors. Technometrics 4:278-281.

Semenov MA, Brooks RJ, Barrow EM, Richardson CW. 1998. Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Climate Research 10:95-107.

Taussky O, Todd J. 2006. Cholesky, Toeplitz and the triangular factorization of symmetric matrices. Numerical Algorithms 41:197-202.

Toeplitz O. 1907. Die Jacobische Transformation der quadratischen Formen von unendlich vielen Veränderlichen, Nachrichten von der Akademie der Wissenschaften zu Göttingen, Mathematisch – Physikalische Klasse, pp. 101-110.

Wilby RL, Tomlinson OJ, Dawson CW. 2003. Multi-site simulation of precipitation by conditional resampling. Climate Research 23:183-194.

Wilks DS. 1998. Multisite generalization of a daily stochastic precipitation generation model. Journal of Hydrology 210:178-191.

Wilks DS. 2006. Statistical methods in the atmospheric sciences. International Geophysics Series. 2nd edn. Elsevier Academic Press Publications, CA.

APPENDIX

Generation of correlated random numbers

To transform a vector ($\vec{r}_{norm}$) with elements (r_i) randomly generated from a Gaussian distribution and independently generated from one each other into a vector ($\vec{r}^{*}_{norm}$) with elements (r_i^{Ψ}) randomly generated following a Gaussian distribution but pair-wise correlated with each other ($\rho_{i,j}$), each new element from $\vec{r}^{*}_{norm}$ must be based on a weighted linear combination of $\vec{r}_{norm}$ elements:

$$\begin{cases} r_1^{\Psi} = r_1 \times C_{1,1} + ... + r_n \times C_{1,n} \\ ... \\ r_n^{\Psi} = r_1 \times C_{n,1} + ... + r_n \times C_{n,n} \end{cases}$$

Therefore, $\vec{r}_{norm}$ must be matrix-multiplied by a square matrix containing the weighting values (C_{ij}), which are calculated based on the pair-wise Pearson's correlation values that form the correlation matrix $[\rho]$ (Scheuer and Stoller, 1962). The proposed multiplicative matrix is a reduced function of the summation between $[\rho]/2$ and a factored $[\rho]$. The most common factorization matrices are the Toeplitz-Cholesky factorization matrix $[TC]$ (Benoît, 1924; Brezinski, 2006; Toeplitz, 1907; Taussky and Todd, 2006; see Equation A.1), and the Eigen decomposition matrix $[Eig]$ (Hilbert, 1904; Hotelling, 1933; Equation A.2).

$$[TC] = [U][Diag(\sqrt{U})] \qquad [A.1]$$

where $[U]$ is an upper triangular matrix with positive diagonal entries generated from a special case of the symmetric LU decomposition of the correlation matrix with $[L] = [U]^T$.

$$[Eig] = [\vec{E}_i[Diag(\sqrt{\lambda_i})]]^T \qquad [A.2]$$

where $\vec{E}_i$ are eigenvectors and λ_i are Eigenvalues from the Eigen decomposition of the correlation matrix.

In our experience, both methods perform equally well; however, as pointed out by Iman and Conover (1982), this approach only works for Gaussian distributions. Because rainfall amounts usually follow a Gamma distribution in this region, these methods cannot be applied without data transformation. However, number of rainy days per month tends toward a Gaussian distribution in non-arid zones, and thus, they can be used to generate a random number of rainy days in each month.