A semiparametric multivariate and multisite weather generator

Somkiat Apipattanavis,[1,2] Guillermo Podestá,[3] Balaji Rajagopalan,[1,4] and Richard W. Katz[5]

Received 4 November 2006; revised 16 July 2007; accepted 1 August 2007; published 1 November 2007.

[1] We propose a semiparametric multivariate weather generator with greater ability to reproduce the historical statistics, especially the wet and dry spells. The proposed approach has two steps: (1) a Markov Chain for generating the precipitation state (i.e., no rain, rain, or heavy rain), and (2) a *k*-nearest neighbor (*k*-NN) bootstrap resampler for generating the multivariate weather variables. The Markov Chain captures the spell statistics while the *k*-NN bootstrap captures the distributional and lag-dependence statistics of the weather variables. Traditional *k*-NN generators tend to under-simulate the wet and dry spells that are keys to watershed and agricultural modeling for water planning and management; hence the motivation for this research. We demonstrate the utility of the proposed approach and its improvement over the traditional *k*-NN approach through an application to daily weather data from Pergamino in the Pampas region of Argentina. We show the applicability of the proposed framework in simulating weather scenarios conditional on the seasonal climate forecast and also at multiple sites in the Pampas region.

Citation: Apipattanavis, S., G. Podestá, B. Rajagopalan, and R. W. Katz (2007), A semiparametric multivariate and multisite weather generator, *Water Resour. Res.*, *43*, W11401, doi:10.1029/2006WR005714.

1. Introduction

[2] Stochastic weather generators are routinely used in water, agricultural, and erosion control management [*Skidmore and Tatarko*, 1990; *Wilks*, 1997; *Dubrovsky et al.*, 2000]. *Hutchinson* [1987] and *Wilks and Wilby* [1999] reviewed methods for generating synthetic daily weather series. Traditional generators, also known as "parametric" weather generators, typically use precipitation as the driving variable [*Jones et al.*, 1972; *Nicks and Harp*, 1980; *Richardson*, 1981]. In most available models the generation of daily precipitation involves two components: (1) the occurrence process (i.e., the sequence of "dry" or "wet" days), and (2) the intensity process (i.e., the sequence of precipitation amounts on wet days). A simple model of the occurrence process is a two-state, first-order Markov chain. This model has been used in many situations with remarkable success [*Richardson*, 1981]. There is extensive literature on precipitation generation using a Markov chain or Poisson process framework; to list a few, *Katz* [1977], *Foufoula-Georgiou and Georgakakos* [1991], *Woolhiser* [1992], *Lettenmaier* [1995], and *Katz and Parlange* [1995]. An alternative to Markov chain models for simulating precipitation occurrence is the use of spell-length models, in which probability distributions are fitted to observed relative frequencies of wet- and dry-spell lengths [*Wilks and Wilby*, 1999]. The rainfall amount on a wet day is generated from a probability density function (PDF) fitted to historical data; examples include the lognormal, cubic root normal, exponential, mixed exponential, kappa, gamma, and Weibull distributions [*Richardson*, 1981; *Hutchinson*, 1987; *Woolhiser*, 1992]. Other variables such as maximum and minimum temperatures are generated by fitting a lag-1 multivariate autoregressive model (MAR-1) with exogenous precipitation input [*Richardson*, 1981]; that is, the MAR-1 model is fitted to the variables whose standardizations depend on whether the day is dry or wet. Furthermore, to simulate precipitation throughout the year, the seasonal nature of these parameters may be described by using Fourier series. *Parlange and Katz* [2000] extended this framework to include other variables such as daily mean wind speed and dew point.

[3] The parametric approaches are easy to implement and have a rich background, but they suffer from the following main shortcomings. (1) The MAR framework requires normality of the data. If the data are not normally distributed, they have to be transformed to normality. With several variables and seasons (e.g., months), this transformation task can be quite difficult. Furthermore, good performance of the model in the transformed space does not guarantee the same in the original space. (2) For the rainfall amounts, potential nonnormal features such as bimodality, if exhibited by the data, cannot be captured by the limited suite of PDFs.

[1]Department of Civil, Environmental and Architectural Engineering, University of Colorado, Boulder, Colorado, USA.
[2]Now at Office of Research and Development, Royal Irrigation Department, Nonthaburi, Thailand.
[3]Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, Florida, USA.
[4]Also at Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, Colorado, USA.
[5]National Center for Atmospheric Research, Boulder, Colorado, USA.

0043-1397/07/2006WR005714$09.00

[4] Nonparametric methods provide an attractive alternative to the traditional parametric weather generators described above. They do not assume a particular theoretical PDF for the variables [*Wilks and Wilby*, 1999]; rather, they are data driven. Use of empirical distributions of wet and dry spells and precipitation amounts are the simplest nonparametric approach, e.g., the LARS-WG by *Semenov and Porter* [1995]. Other methods include simulated annealing for the generation of precipitation series [*Bárdossy*, 1998] and neural networks for generation of temperature series [*Trigo and Palutikof*, 1999]. Multivariate kernel density estimators based on nonparametric weather generators were developed by *Rajagopalan et al.* [1997]. The kernel-based methods involve fitting a multivariate PDF to the suite of weather variables [*Rajagopalan et al.*, 1997]. This approach has been applied successfully for monthly streamflow generation [*Sharma et al.*, 1997] and also for daily weather generation [*Rajagopalan et al.*, 1997], but it can be cumbersome for higher-dimension problems (i.e., more variables). Resampling (i.e., bootstrap) historical observations provides a simpler and effective alternative to kernel methods [*Young*, 1994]. This alternative was improved upon through the *k*-nearest neighbor (*k*-NN) time series bootstrap approach by *Lall and Sharma* [1996], subsequently extended for stochastic weather generation at a single site [*Rajagopalan and Lall*, 1999] and at multiple sites [*Buishand and Brandsma*, 2001; *Yates et al.*, 2003].

[5] Classical bootstrap techniques [*Efron*, 1979; *Hardle and Bowman*, 1988; *Zucchini and Adamson*, 1989; *Yakowitz*, 1993] were developed quite a while ago in the statistics literature for estimating standard errors, and then have been used largely for providing confidence intervals for estimates from models. Recently, the bootstrap has been modified for time series simulation [*Young*, 1994; *Lall and Sharma*, 1996]. In the weather generation context, *Young* [1994] employed a multiple discriminant function to identify *k*-nearest neighbors (where *k* was 3–5 days) of the current day's weather, and one of these neighbors was then randomly selected and used as the next day's weather. Young's model preserves the cross correlation between the variables for the most part, but biases were noticed (e.g., reduced persistence and underestimation of the fraction of dry months) in the generated series.

[6] *Rajagopalan and Lall* [1999] extended the *k*-NN bootstrap method developed by *Lall and Sharma* [1996] for univariate time series resampling to multivariate data (i.e., daily weather). Rajagopalan and Lall's algorithm is very similar to Young's method but has two main differences: (1) A Euclidean space is used instead of the multiple discriminant space function, and (2) *k*-nearest neighbors to the current day's weather are obtained and then one of these neighbors is selected based on a probability metric that depends on the closeness of the neighbor. Their approach preserved the persistence (i.e., lag-1 correlations) and also seasonal statistics. *Buishand and Brandsma* [2001] and *Yates et al.* [2003] extended the *k*-NN bootstrap weather generator to multisite generation with good success. Subsequently, *Gangopadhyay et al.* [2005] used the same model for ensemble weather forecasts at multiple sites. Furthermore, the *k*-NN approach was modified for conditional resampling on atmospheric indices [*Beersma and Buishand*, 2003] and hydrologic time series [*Mehrotra and Sharma*, 2006].

[7] The *k*-NN generators tend to undersimulate the lengths of wet and dry spells, much as was observed with the *Young* [1994] approach, more so in situations with short spell lengths. This underestimation results because of the intermittent nature of precipitation and the fact that the nearest neighbors are based on all the weather variables, even though the others are not intermittent (e.g., maximum and minimum temperatures, and solar radiation). To alleviate this problem, we propose a modification to the *k*-NN weather generator. The proposed ''hybrid'' generator involves two steps combining parametric and nonparametric approaches. In an initial (parametric) step, a Markov chain is used to generate the precipitation state of the day (i.e., wet or dry). In a second (nonparametric) step, a *k*-NN method is used to generate the suite of weather variables conditioned on the simulated precipitation state. This approach is quite flexible and has the ability to generate scenarios consistent with seasonal climate forecasts, such as those operationally issued by various organizations around the world. The simulated conditional scenarios are enormously useful for water resources planning and management. The proposed framework and the implementation algorithm are described in section 2.

[8] The objectives of the work presented here are (1) to improve the ability of the traditional *k*-NN model of *Lall and Sharma* [1996], *Rajagopalan and Lall* [1999], and *Buishand and Brandsma* [2001] to capture the historical spell statistics by modifying the original algorithm to incorporate an additional Markov Chain model; and (2) to add to the modified model the capability of generating weather scenarios conditioned on seasonal climate forecasts currently issued operationally by many agencies around the world. In this study, probabilistic forecasts from the International Research Institute for Climate and Society (IRI, www.iri.columbia.edu) are used. The modified framework as semiparametric approach is first presented. We demonstrate its application to daily weather data for Pergamino, Argentina, located in one of the main agricultural production regions in the world. Conditional weather generation based on IRI forecasts is also discussed. Comparisons of the simulations from the proposed modified *k*-NN bootstrap with those from the traditional *k*-NN simulation model are also provided.

2. Proposed Semiparametric Weather Generator

[9] The traditional *k*-NN model of daily weather generation can be expressed as simulating from the conditional PDF $f(x_t \mid x_{t-1})$ where x_t and x_{t-1} are the weather states on days t and $t-1$, respectively. This conditional PDF is approximated locally via *k*-NN of x_{t-1}. As mentioned above, in their current form *k*-NN generators tend to under-simulate the lengths of wet and dry spells. This underestimation can have a significant impact on crop modeling, where the sequences of wet and dry days are critical for plant growth and, consequently, simulations of crop yields. Impacts on streamflow simulations also are apparent. We modify the *k*–NN weather generator into a semiparametric approach involving an additional step to simulate precipitation occur-

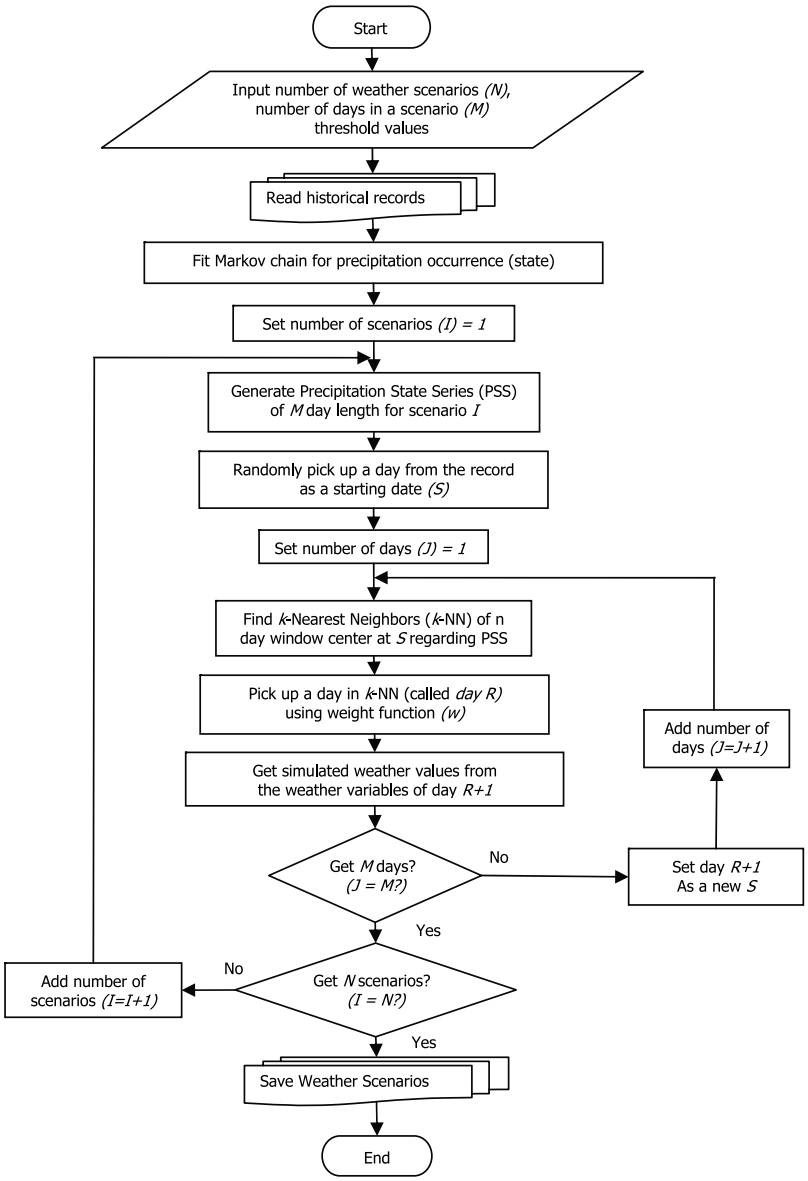

Figure 1. Schematic of the semiparametric weather generator.

rence (see Figure 1). This can be expressed as simulating from the conditional PDF

$$f(x_t|x_{t-1}, S_t, S_{t-1}), \tag{1}$$

where S_t and S_{t-1} are the precipitation state on day t and $t-1$, respectively. The precipitation state is modeled first by a Markov chain, and subsequently the k-NN method is used to locally approximate the above conditional PDF and simulate from it. A detailed description and implementation steps are described below.

2.1. Markov Chain for Precipitation State

[10] A three-state, first-order Markov chain is fitted to the daily precipitation data to simulate the precipitation state. In this approach, the precipitation states are categorized into three states: dry (daily precipitation <0.3 mm), extreme wet (precipitation greater than the 80th percentile of daily amounts for the simulated month), and wet otherwise. The thresholds between categories can be user specified. The transition probabilities p_{ij} for transitions from state i at the previous time step to j at the current time step are estimated by using maximum likelihood. The transition probabilities can be estimated separately for each month or any other desired period (weekly, biweekly, a moving time window, etc.) to capture the seasonality. In this study, monthly statistics are applied. A first-order Markov chain has been successfully used to model precipitation state [*Gabriel and Neuman*, 1962; *Caskey*, 1963; *Haan et al.*, 1976]. A second-order, two-state model also may be used, but in many cases it performs no better than the simpler first-order model (and requires estimation of more parameters). A first order, many-states model can capture a high fraction of the seasonal variability, because the use of many states improves the model's representation of spells of heavy precipitation which appears to have a considerable influence on the seasonal variance [*Gregory et al.*, 1993]. Therefore, in this application we used a first-order, three-state model

Table 1. Unconditional Probabilities of Dry, Wet, and Extreme Wet Days

	Probability
Dry day	0.850 (p_d)
Wet day	0.120 (p_w)
Extreme wet day	0.030 (p_e)

instead of a more common two-state model. With three states, larger precipitation amounts were guaranteed to be well simulated and hence better simulation of annual precipitation totals, which were substantially undersimulated with a two-state Markov chain. Tables 1 and 2 below show an example of the three-state unconditional probabilities and the three-state transition probabilities, respectively.

[11] The simulation proceeds as follows:

[12] 1. A series of uniform random numbers U_i, $i = 1,\ldots, N$, on the interval [0,1] is generated.

[13] 2. The precipitation state of day 1 ($\boldsymbol{S}_1$) is dry if $\boldsymbol{U}_1 < 0.850$ (p_d) or wet if $0.850 \leq \boldsymbol{U}_1 < 0.970$ or extreme wet otherwise (see Table 1).

[14] 3. The precipitation state of day 2 ($\boldsymbol{S}_2$) is generated using $\boldsymbol{S}_1$, $\boldsymbol{U}_2$, and Table 2. For example, if $\boldsymbol{S}_1$ is a dry day, then the transition probabilities (p_{dd}, p_{dw}, and p_{de}) in the first row of the table are considered, and if $\boldsymbol{U}_2 < 0.894$ (p_{dd}) then $\boldsymbol{S}_2$ is a dry day, or if $0.894 \leq \boldsymbol{U}_2 < 0.980$ then $\boldsymbol{S}_2$ is a wet day, else extreme wet. Similarly, if $\boldsymbol{S}_1$ is a wet day, then the transition probabilities (p_{wd}, p_{ww}, and p_{we}) in the second row of the table are considered, and if $\boldsymbol{U}_2 < 0.612$ (p_{wd}) then $\boldsymbol{S}_2$ is a dry day, or if $0.612 \leq \boldsymbol{U}_2 < 0.911$, then $\boldsymbol{S}_2$ is a wet day, else extreme wet.

[15] 4. Repeat step 3 to generate a sequence of days with the appropriate precipitation state.

[16] When crossing the monthly boundaries, the appropriate transition probabilities are used in the simulation.

2.2. The *k*-NN Algorithm for the Weather Variables

[17] The *k*-NN algorithm of lag-1 to generate the weather variables is as follows:

[18] 1. Suppose the vector of daily weather on simulated day 1 (e.g., 1 January) is $\boldsymbol{x}_1$. Also suppose from the Markov chain simulation the state on day 1 is wet and that of day 2 is dry.

[19] 2. We place a 7-day (can be user defined) window centered on 1 January (i.e., 29 December to 4 January) and select all the historical day pairs inside the window that involve a wet day followed by a dry day (the sequence generated by the Markov chain in step 1 above).

[20] 3. Compute the distance between the vector $\boldsymbol{x}_i$ and all the wet days of the sequence pairs selected from step 2. The weighted Euclidean distance between the vector $\boldsymbol{x}_i$ and a historical vector $\boldsymbol{x}_m$ is calculated as

$$r_{im} = \sqrt{\left[\sum_{j=1}^{d} w_j \left(x_{ij} - x_{mj}\right)^2\right]}, \tag{2}$$

where $x_{()j}$ is the jth component (or variable of the vector of weather variables) and the w_j are weights. The weights w_j may thus be specified a priori, as is done here with equal weights to each variable, or they may be chosen to provide the best forecast for a particular successor in a least squares sense [*Yakowitz and Karlsson*, 1987]. The latter would be the desirable method, but it adds a substantial computational burden. Multiple discriminant analysis, as used by *Young* [1994], would be another choice for neighbor identification.

[21] 4. Once the distances are computed from step 3, they are ordered from the nearest to the farthest. Denote the ordered set of nearest-neighbor indices by $J_{i,k}$. An element $j(i)$ of this set records the time t associated with the jth closest neighbor of $\boldsymbol{x}_1$. The first k neighbors are selected and each is assigned a weight defined by a discrete decreasing kernel $K[j(i)]$ as follows:

$$K[j(i)] = \frac{1/j}{\sum_{j=1}^{k} 1/j}, \tag{3}$$

where j is the ordered number, 1,2,..,k and $K[j(i)]$ is the weight function. *Lall and Sharma* [1996] evaluated a suite of weight functions and found that this weight function worked well and was easy to implement. The nearest neighbor gets a higher weight and the kth neighbor gets the least. The weights are normalized so that they sum to unity, thus forming a probability metric. The heuristic scheme of $k = \sqrt{N}$ where N equals the sample size, as proposed by [*Lall and Sharma*, 1996], is used. Note that there is a trade-off in selecting k in that a large k leads to increased variability but distorts the reproduction of distributional properties and vice versa for a small k. However, this heuristic scheme has been shown to perform well in a variety of applications [*Lall and Sharma*, 1996; *Rajagopalan and Lall*, 1999; *Yates et al.*, 2003].

[22] 5. One of the k-nearest neighbors (i.e., one of the historical days within the 7-day window) is selected based on the weights calculated in the previous step. The daily weather of the successive day to the selected neighbor (or day) is the simulated weather for day 2.

[23] Steps 1–5 are repeated to generate synthetic weather for successive days.

2.3. Multisite Weather Generation

[24] A major limitation of conventional weather generator models is that they simulate weather for single sites in isolation. That is, while they can be used to simulate weather at more than one site, the resulting series for the different locations will be independent of one another and will not reflect the strong spatial correlation exhibited by real weather [*Wilks*, 1998]. In some applications, it is important to preserve the spatial correlation when simulated series are reused as input to process models. For example, in hydrological applications the spatial distribution of precipitation may have considerable effects on discharge of a river or the formation of floods [*Qian et al.*, 2002]. For this reason, various approaches have been developed for the

Table 2. Three-State Transition Probabilities

	Transition Probability (p_{ij})		
	Dry Day	Wet Day	Extreme Wet Day
Dry day	0.894 (p_{dd})	0.086 (p_{dw})	0.020 (p_{de})
Wet day	0.612 (p_{wd})	0.299 (p_{ww})	0.089 (p_{we})
Extreme wet day	0.547 (p_{ed})	0.391 (p_{ew})	0.063 (p_{ee})

simultaneous generation of synthetic weather at multiple sites [*Wilks*, 1998, 1999a, 1999b; *Buishand and Brandsma*, 2001; *Qian et al.*, 2002; *Beersma and Buishand*, 2003; *Kottegoda et al.*, 2003; *Wilby et al.*, 2003].

[25] The semiparametric approach proposed above can be extended easily for weather generation at multiple sites. First, a spatially averaged daily weather time series is computed from the historical daily weather data at multiple locations. Then, the same threshold criteria (i.e., 0.3 mm and 80th percentile of spatially averaged precipitation time series) for categorizing precipitation states (i.e., dry, wet, and extreme wet) are applied to the spatially averaged daily time series. The weather generator described in the previous section is fitted to the "spatially averaged" daily weather time series. When one of the *k*-nearest neighbors, i.e., one of the historical days is selected during step 5 of the algorithm, the corresponding daily weather at all the sites for this historical day is selected as the simulated weather. That is, daily weather is simulated for all the sites simultaneously. This approach is similar to the multisite simulations proposed by *Yates et al.* [2003] and *Buishand and Brandsma* [2001]. An implicit assumption is that the "spatially averaged" weather time series is "representative" of the entire region; thus this approach works well if the sites are homogenous in terms of their weather.

2.4. Conditional Weather Generation

[26] Exciting scientific and technological advances have resulted in seasonal climate forecasts that anticipate characteristics of a period (e.g., a wetter than normal spring) with a lead time of 6–9 months [*Goddard et al.*, 2003]. Forecasts are operationally provided by several agencies throughout the world, e.g., the International Research Institute for Climate and Society (http://iri.columbia.edu/climate/forecast/net_asmt/). The IRI seasonal climate forecasts are provided for the entire world with a lead time of up to 6 months in 3-month moving windows. Of course, forecast skill varies with season and location. The IRI precipitation forecasts are provided as percentage likelihood of A:N:B format, where A denotes percent chance of above-normal rainfall, N denotes percent chance of near-normal rainfall, and B denotes percent chance of below-normal rainfall. These three categories are defined by the terciles of the historical distribution of rainfall (A:N:B = 33:33:33) for the location and period in question. For example, a forecast of A:N:B = 40:35:25 for an area means that there is a 40% chance of rainfall being above normal, 35% chance of rainfall being near normal, and 25% chance of below normal precipitation.

[27] Seasonal climate forecasts, typically probabilistic in nature and covering a large spatial region, can be translated into distributions of site-specific outcomes via the generation of synthetic series conditioned on the forecast. In order to generate weather scenarios conditioned on the above categorical forecast, *Yates et al.* [2003] and *Clark et al.* [2004] suggested a weighted resampling of historical years based on the seasonal forecast. Here we adapt this approach and propose a simpler version of the same. The proposed version is similar to the version of *Briggs and Wilks* [1996]; however, we have extended it to include a suite of weather variables.

[28] First, the historical years are classified into three categories (wet, normal, and dry) based on the terciles of the historical seasonal precipitation. Then, we sample with replacement the years or seasons for which the forecast applies. The sampling is weighted according to the probabilities predicted for each tercile. Follow the previous example of a 40:35:25 forecast, we would select 40 years from the wet category, 35 from the normal category, and 25 years from the dry category. Of course, since the sampling is with replacement, some of the years may be selected more than once. The 100 years of sampled data serve as input to the weather generator described in the previous section. The resulting synthetic series should have similar percentage likelihood as the seasonal predictions.

3. Model Evaluation

[29] The performance of the semiparametric approach is evaluated by applying it to daily weather data, consisting of three variables (precipitation, maximum temperature, and minimum temperature (the variables required, together with daily radiation, for the simulation of crop yields)), at Pergamino, Argentina, for the period 1931–2003. *Furrer and Katz* [2007] fitted a parametric weather generator to this same data set. Monthly statistics of precipitation totals as well as daily maximum and minimum temperatures are shown in Figure 2 as box plots. It can be seen that the wet season is during October–March and the dry season is during April–September.

[30] Daily weather data for two other locations in the same region, Junín and Nueve de Julio, are used for the evaluation of multisite weather generation. These two sites are approximately 86 and 183 km south of Pergamino, respectively. The available daily records for Junín encompass the period 1950–2001; the corresponding data for Nueve de Julio are for 1950–1996. For multisite weather generation the common period of data for all the three stations, 1950–1996, is used.

[31] The evaluation of the generator involved three separate simulations. First, we generated 100 ensembles for Pergamino, each 73 years long (the length of the historical record), using the semiparametric approach presented earlier, also referred to as "unconditional simulation."

[32] For "conditional simulation" we generated 100 ensembles of daily weather sequences for October–November–December (OND) and January–February–March (JFM) quarters based on the large-scale probabilistic seasonal precipitation forecast. To demonstrate, we chose two wetter than normal (OND 2002 and JFM 2003) quarters and two drier than normal (OND 2003 and JFM 2004) quarters. The IRI forecasts are for three equally likely categories based on the tercile boundaries; the selected quarters had their seasonal rainfall in the upper or lower tercile to qualify as wet and dry year quarters, respectively. So two sets of OND quarter simulations, 100 ensembles each of 92 days, conditioned on the IRI forecasts for the OND 2003 (drier) and the OND 2002 (wetter), were generated, and similarly, for the JFM 2004 (drier) and the JFM 2003 (wetter) quarters.

[33] The third simulation involved the simultaneous generation of daily weather at three sites (Pergamino, Junín, and Nueve de Julio) using the multisite approach described above. In this case, 100 ensembles, each 47 years long, were generated. For all simulations, a suite of diagnostic statistics was computed by month: mean, median, interquartile range, lag-1 autocorrelation, precipitation state transition probabil-

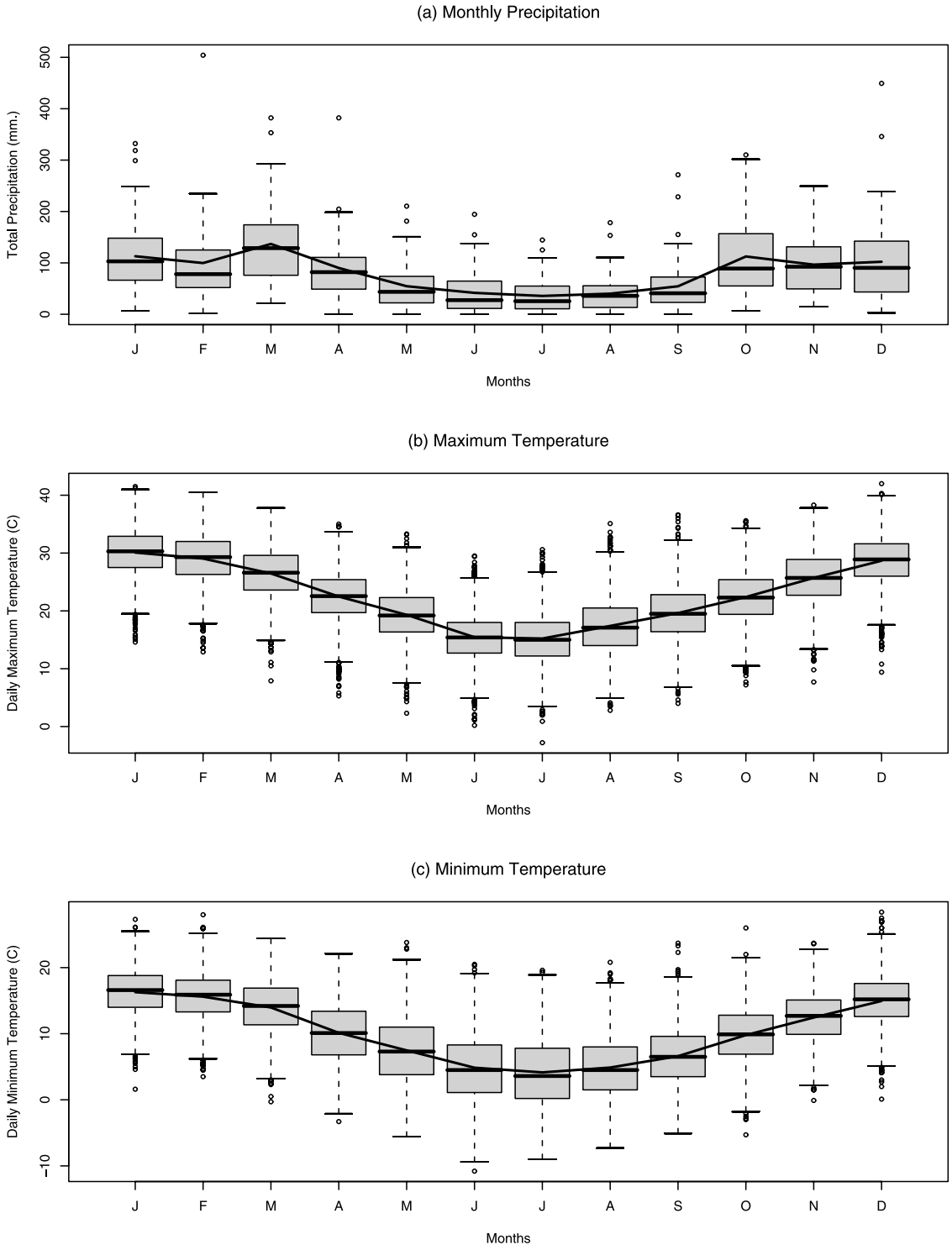

Figure 2. Box plots of weather variables throughout the year at Pergamino, Argentina (latitude 33°53′S, longitude 60°35′W, years 1931–2003). The height of the box represents the interquartile range, the horizontal line inside the box is the median, whiskers extend to the 5th and 95th percentiles, and the solid lines across boxes indicate the average monthly values.

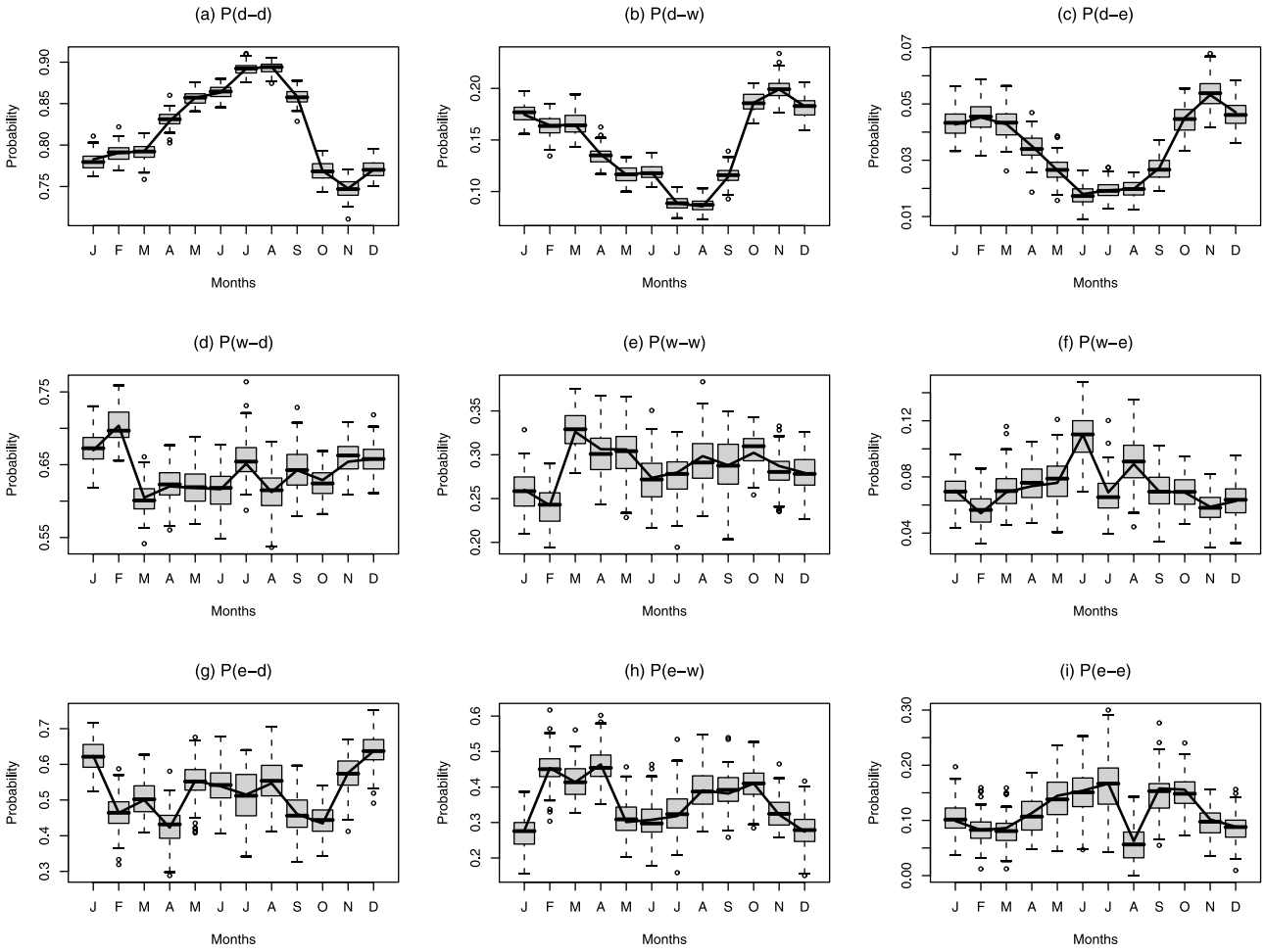

Figure 3. Same as Figure 2 but for transition probabilities from the semiparametric model. A solid line across boxes represents the average monthly historic values.

ities, and wet- and dry-spell lengths. These diagnostics were used to compare the historical and simulated series.

4. Results

[34] The statistics from the different simulations are shown by means of box plots with the corresponding historical value overlaid. Performance for a given statistic is judged as good when the historical value falls within the interquartile range of the box plots.

4.1. Unconditional Simulation

[35] As mentioned above, 100 simulations each of 73 years in length of daily weather sequences were generated. First, we show monthly box plots of nine transition probabilities between precipitation states in Figure 3. The historical value is captured within the box (which encompasses the central 50% of a distribution) for all the statistics and for all months. This is to be expected because the precipitation state is generated from a Markov chain fitted to the historical data. For comparison, the same plots (Figure 4) are created for simulations generated from a traditional *k*-NN weather generator [*Rajagopalan and Lall*, 1999]. These diagnostics show that some of the transition probabilities, in particular P_{wd} and P_{ww}, are not well captured at all by the traditional approach. The average number of dry and wet days, average wet- and dry-spell lengths, and maximum wet- and dry-spell lengths in a 73-year period are illustrated in Figure 5. All of these statistics are well captured by the semiparametric generator. However, in some months the maximum dry- and wet-spell lengths are not quite well reproduced, either because in the Markov chain process we use the maximum likelihood transition probabilities that smooth out the observed maximum and minimum dry- and wet-spell lengths or because of the assumption of first-order dependence. In contrast, spell statistics from the traditional *k*-NN weather generator, especially the average spell lengths, are substantially undersimulated, as shown in Figure 6. So the use of a Markov chain enables the generator to capture the statistics of wet and dry spells quite well compared with the traditional approach.

[36] Monthly and annual distributional statistics of the climate variables (i.e., mean, median, coefficient of skewness, etc.) are all well captured by both the traditional and semiparametric methods, except in some months for which the standard deviation and/or interquartile range of both maximum and minimum temperatures are underestimated; a problem commonly arises with weather generators [*Furrer and Katz*, 2007], since the *k*-NN resampling part is the same in both. The results from the semiparametric model are shown in Figures 7 and 8 for the months of January (wet

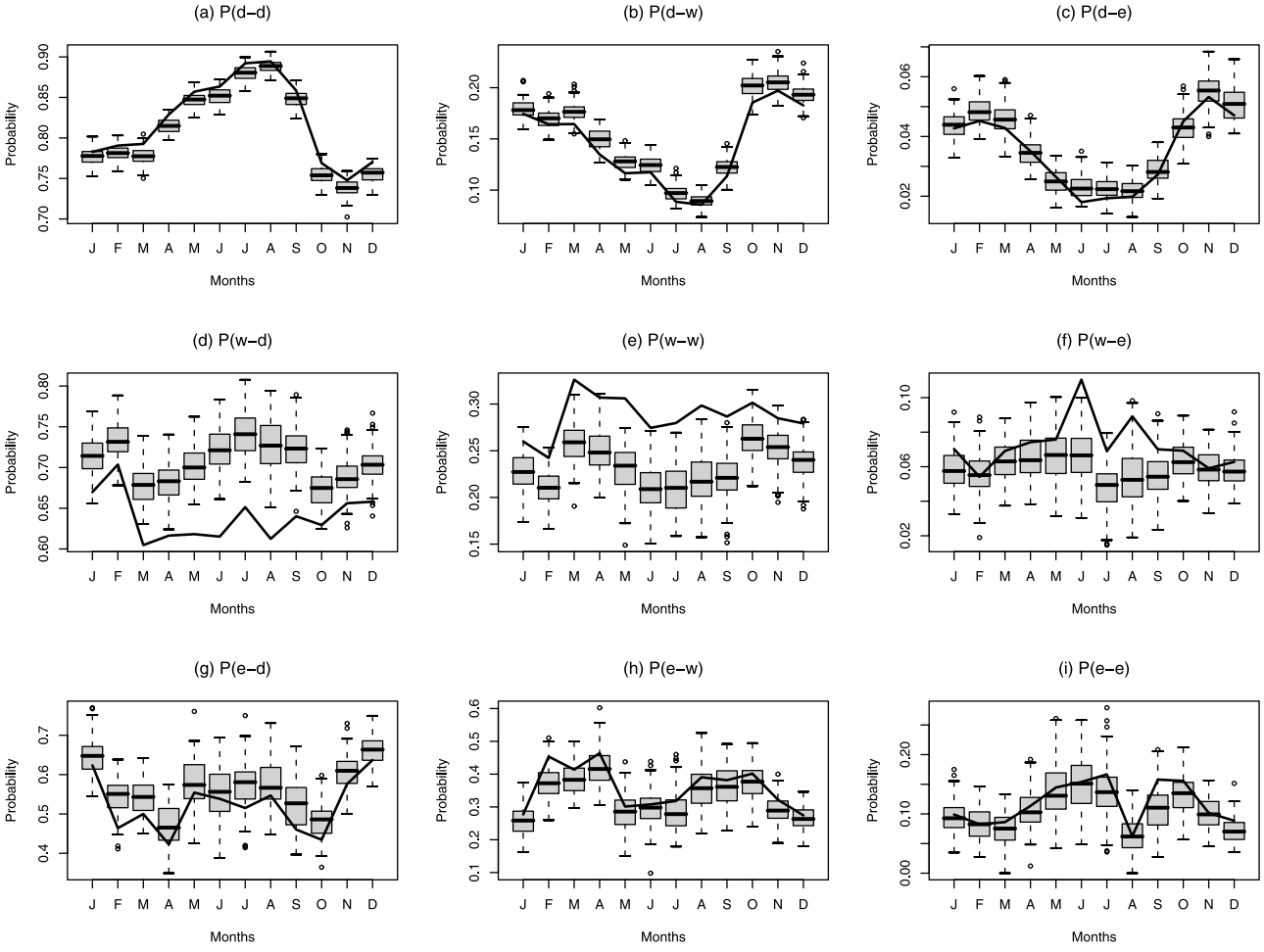

Figure 4. Same as Figure 3 but from the traditional *k*-NN model.

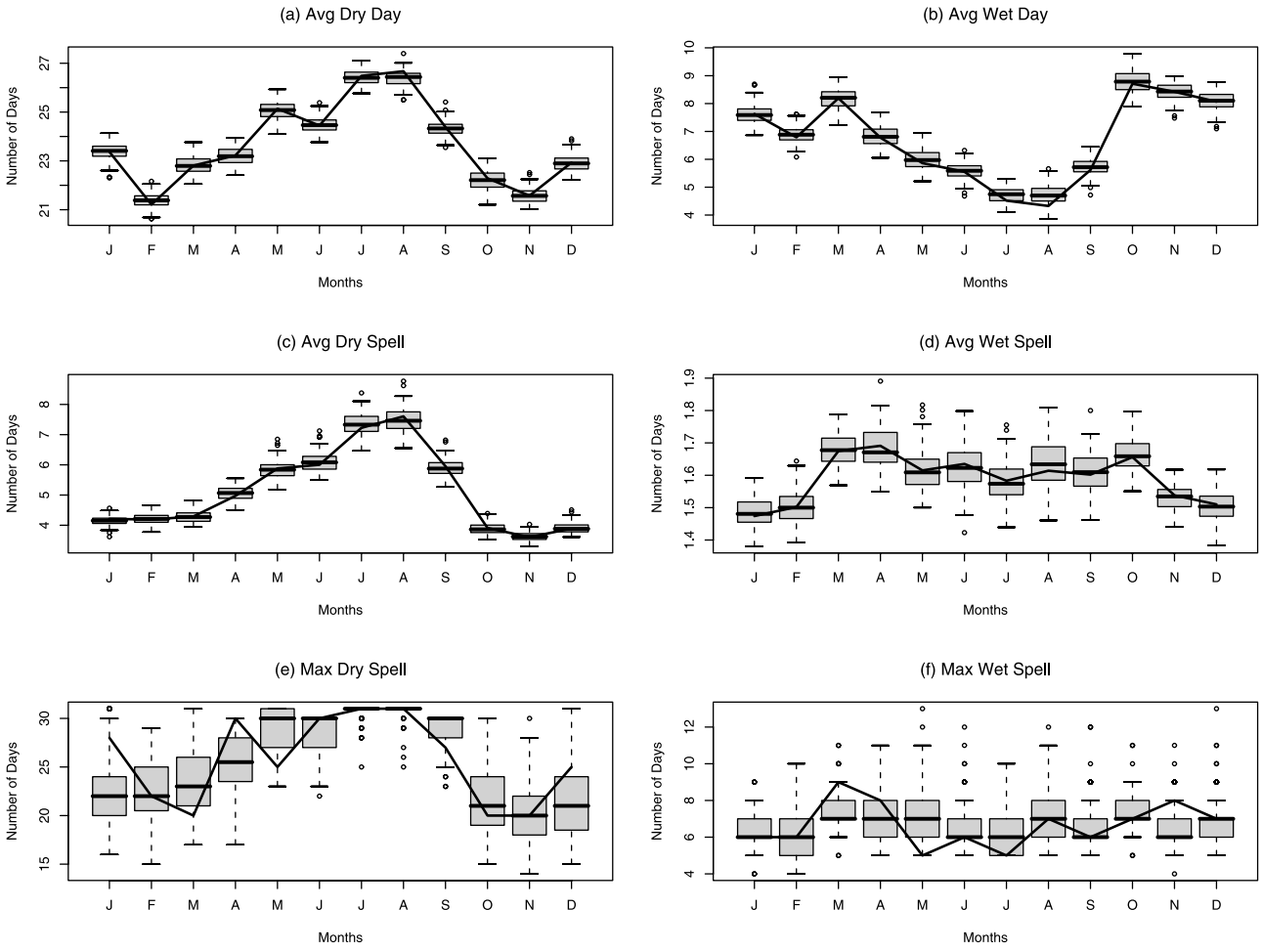

Figure 5. Spell statistics from the semiparametric model. Same format as in Figure 3.

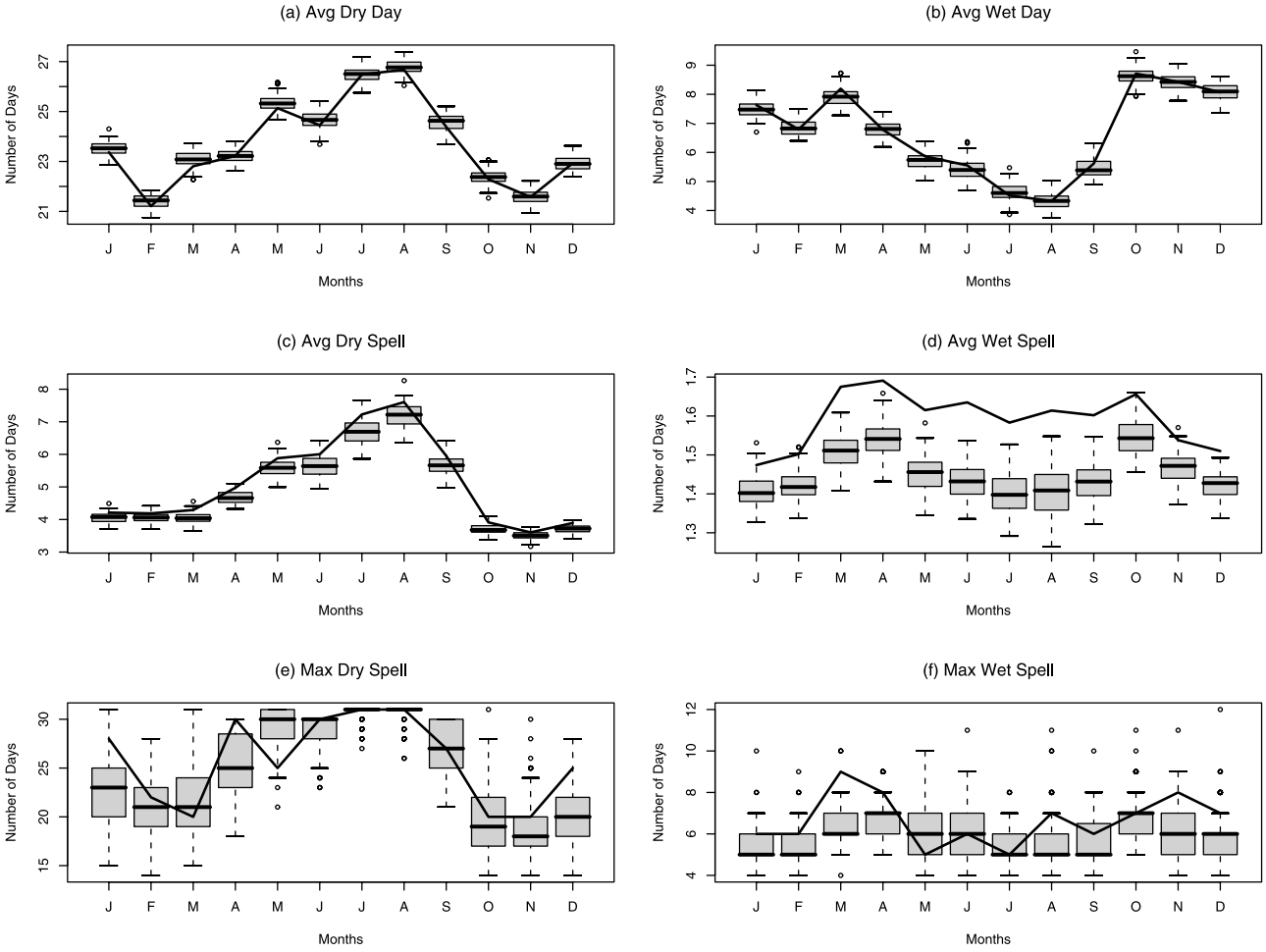

Figure 6. Same as Figure 5 but from the traditional *k*-NN model.

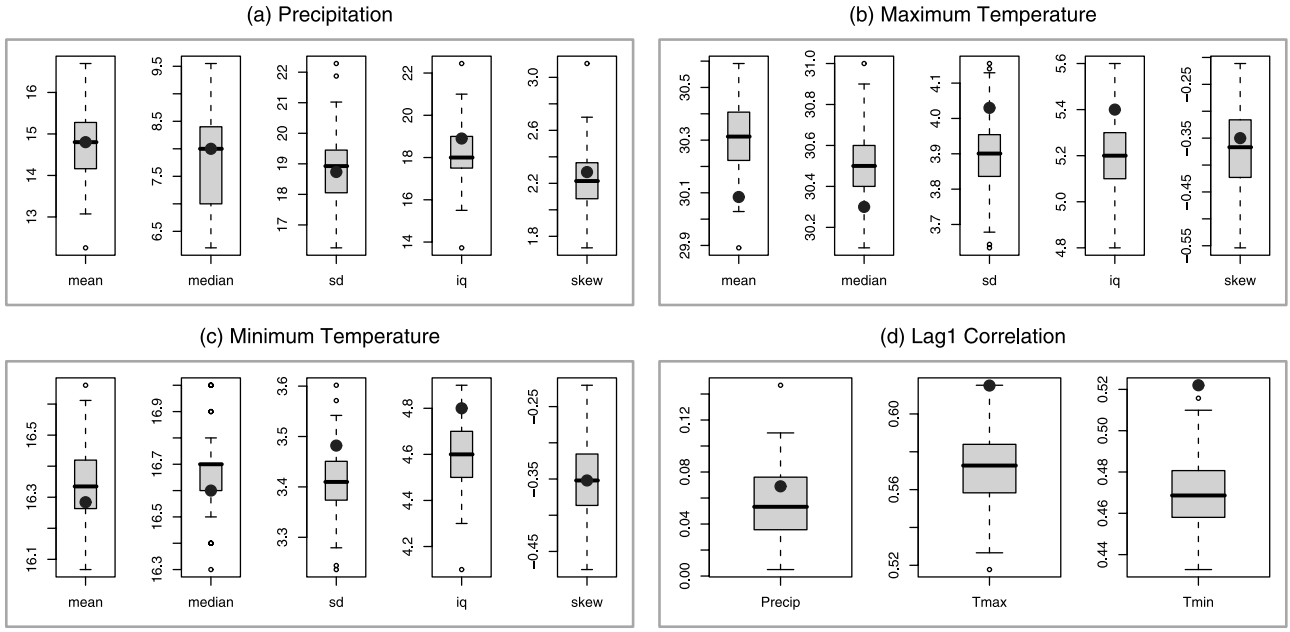

Figure 7. Basic distributional statistics for January from the semiparametric model for Pergamino. (a) Mean, median, standard deviation, interquantile range, and coefficient of skew of daily precipitation. (b) Same as Figure 7a but of daily maximum temperature. (c) Same as Figure 7a but of daily minimum temperature. (d) Lag 1 correlations of daily precipitation, maximum, and minimum temperatures. Solid dots represent the observed statistics.

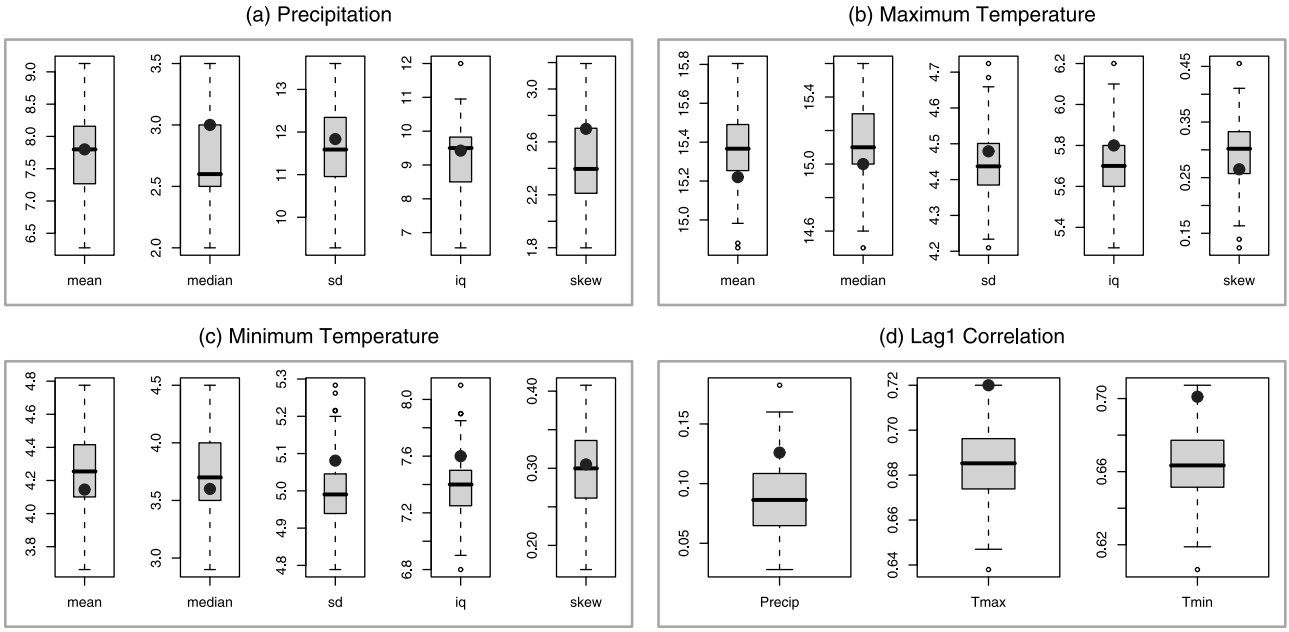

Figure 8. Same as Figure 7 but for July.

month) and July (dry month), respectively. There is a slight underestimation of the lag-1 correlations of the temperatures. This is because the neighbors in the resampling procedure are based on the precipitation state. However, it can be concluded that the abilities of both semiparametric and traditional *k*-NN models in capturing the distributional statistics are similar. Note that the results of distributional statistics from the traditional model are not shown due to the space limitation.

4.2. Conditional Simulation

[37] Conditional weather simulations were performed for two different periods: (1) dry OND quarter and (2) wet JFM quarter. For the first period, two conditional weather simulations were performed, respectively, based on IRI forecasts for drier OND 2003 of 20:35:45 and wetter OND 2002 of 45:35:20. The IRI forecasts were issued on September for the following OND season. We generated 100 ensembles each of 92 days long for each of the two OND periods, following the conditional weather generation approach described in the previous section. Similarly, for the second period, two conditional simulations were performed based on IRI forecasts for drier JFM 2004 of 25:35:40 and wetter JFM 2003 of 40:35:25.

[38] The PDF of the seasonal precipitation from the ensembles, along with the PDF of the historical data, are shown in Figure 9. It can be seen that the conditional weather generation shifts the PDF appropriately, i.e., to the right in wet years and to the left in the dry years. Figure 9a indicates that there is a 20% chance, from the climatological PDF, that seasonal OND precipitation is below 192.5 mm. However, the simulated PDF conditional on a dry forecast indicates that the probability of seasonal OND precipitation being less than 192.5 mm is 22%. Similarly, Figure 9b shows the conditional wet simulated precipitation PDF. The climatological PDF indicates a 20% chance that OND precipitation is above 411 mm. In contrast, the conditional wet PDF gives a probability of 26.1% that the OND precipitation is above 411 mm. Similar shifts are found for conditional JFM simulations. See Tables 3 and 4 for more details. The distribution of the simulated seasonal precipitation compares well with the IRI predictions (see Tables 3 and 4). The utility of this approach is in driving process models such as crop yield and watershed models to provide ensemble forecasts of decision variables.

4.3. Multisite Weather Generation

[39] We extended the weather generator developed here to multisite weather simulation as described earlier. The monthly spell statistics for the station I (Pergamino) are shown in Figure 10. It can be seen the spell statistics are captured fairly well. On the other hand, the same diagnostic from the traditional *k*-NN resampling generator (Figure 11) does not reproduce these statistics well, consistent with the findings for the single-site model. Similar results are found for the other stations (figures are not shown here). However, the simulated spell statistics from multisite generation are poorer than those from single-site generation due to the use of transition probabilities that were derived from the spatially averaged time series (see Figures 5 and 10). The basic distributional properties, mean, standard deviation, and coefficient of skew, for all the months for all stations are also quite well captured. Only the plot for station II (Junín) is shown in Figure 12.

[40] The regional spatial correlation of daily precipitation, daily maximum and minimum temperatures, and monthly mean dry- and wet-spell lengths are shown by means of monthly box plots in Figure 13. Note that regional spatial correlation is the average of the correlations among pairs of stations in the study area. It is clearly seen that the regional spatial correlations of each variable are well captured. However, the correlations of spell lengths among stations, especially dry spells during wet season (Figure 13d) and wet spells during dry season (Figure 13e), are not so well reproduced. This may be resulting from the failure to

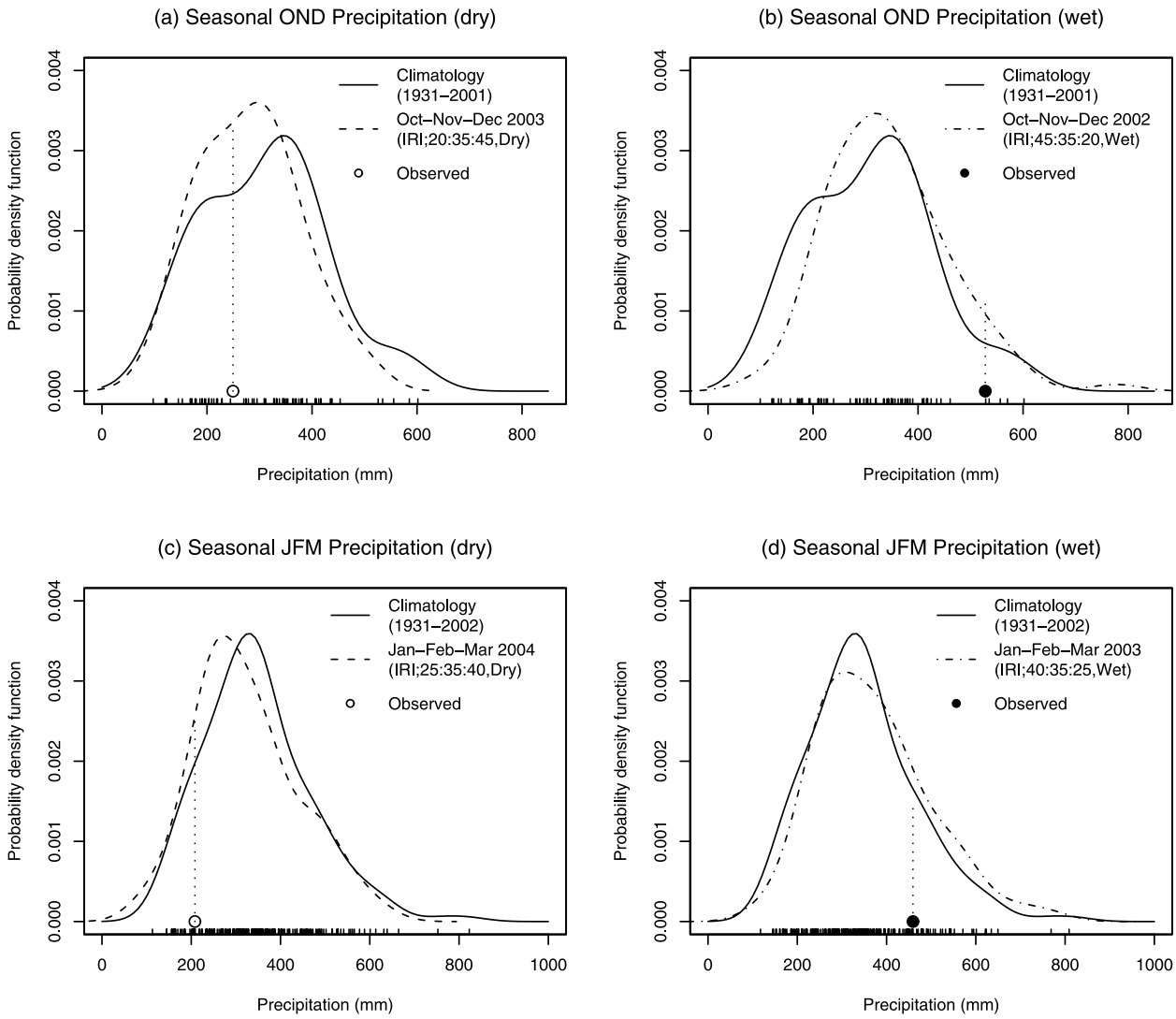

Figure 9. PDF of conditional simulation for Pergamino. The quarterly precipitation actually observed for the forecast season is shown as a vertical line.

reproduce the maximum and minimum spell lengths that show similar results (see Figures 5e and 5f).

[41] The log-odds ratio [*Stephenson*, 2000], which is a measure used to compare the hits and false alarms in weather forecast, is applied here to compare the performance of spatial occurrence of precipitation from the simulation. For a given pair of stations, the probabilities of the four wet and dry combinations (instead of hits and false alarms) at the two stations are computed for the observations for each month, and the log-odds ratio is then calculated, similarly from the simulations. The box plots of this ratio for all pairs of stations are shown in Figure 14. It can be seen that the simulations capture the spatial dependence of the occurrence process very well. As the region is quite homogeneous, the daily averaged time series used as the representative time series for resampling is able to capture the spatial dependence well. However, this may not be the case with a more heterogeneous region.

5. Summary and Conclusion

[42] We developed a semiparametric two-step *k*-NN weather generator. This modifies the traditional *k*-NN resampling approach by the inclusion of a three-state,

Table 3. Conditional Probability Distribution of Seasonal OND Precipitation[a]

Seasonal Precipitation	IRI Prediction	Simulation	P (prcp < 0.2 p)	P (prcp > 0.8 p)
Climate OND	1/3:1/3:1/3	...	20%	20%
OND 2003	20:35:45 (dry)	15:39:46	22%	11.4%
OND 2002	45:35:20 (wet)	37:38:25	9.1%	26.1%

[a]Here p denotes percentile and prcp means precipitation.

Table 4. Conditional Probability Distribution of Seasonal JFM Precipitation[a]

Seasonal Precipitation	IRI Prediction	Simulation	P (prcp < 0.2 p)	P (prcp > 0.8 p)
Climate JFM	1/3:1/3:1/3	...	20%	20%
JFM 2004	25:35:40 (dry)	32:27:41	25.7%	17.4%
JFM 2003	40:35:25 (wet)	43:25:32	16.2%	26.2%

[a]Here p denotes percentile and prcp means precipitation.

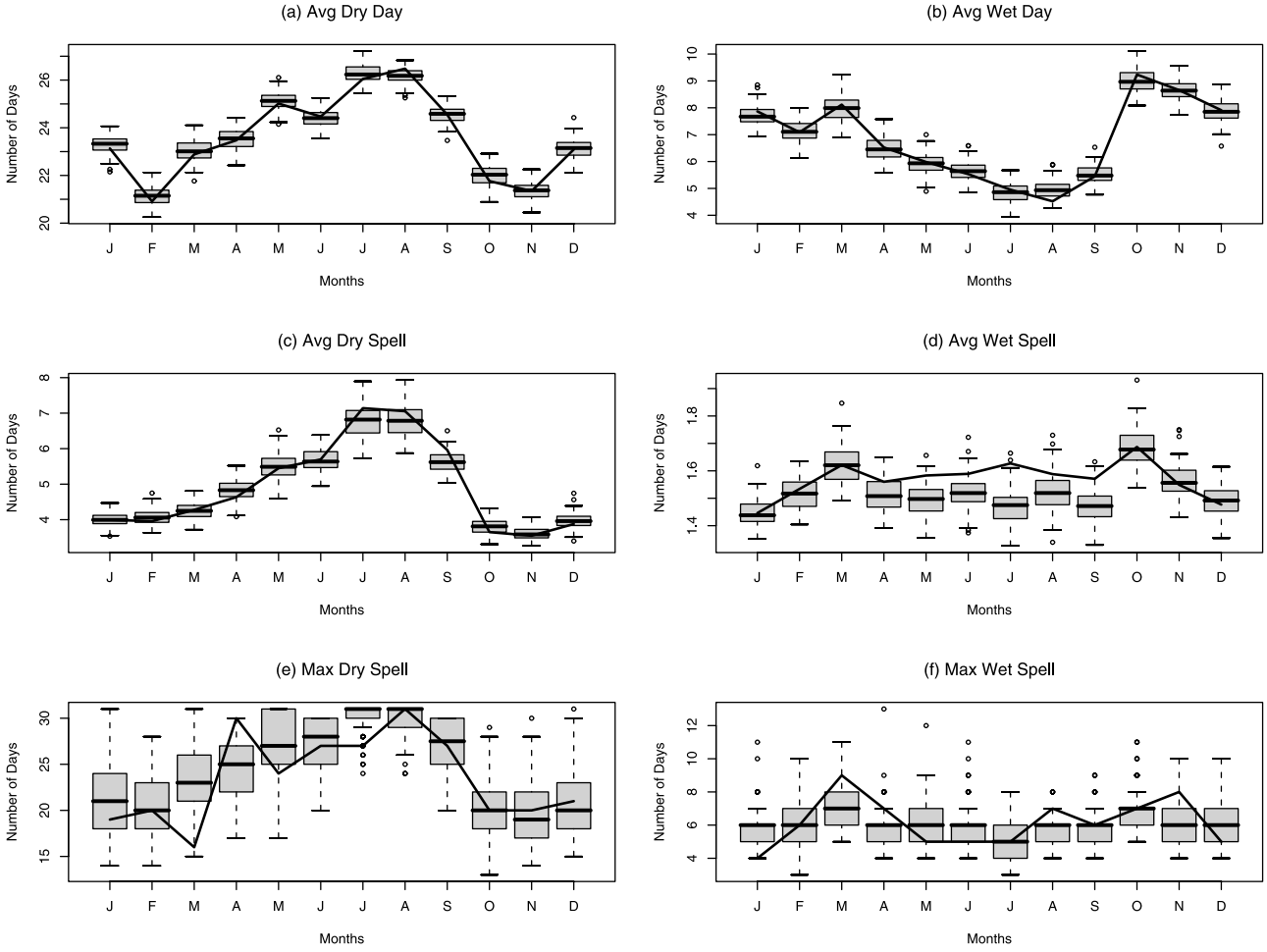

Figure 10. Spell statistics of station I (Pergamino) simulated from multisite generator using the semiparametric model. Same format as in Figure 5.

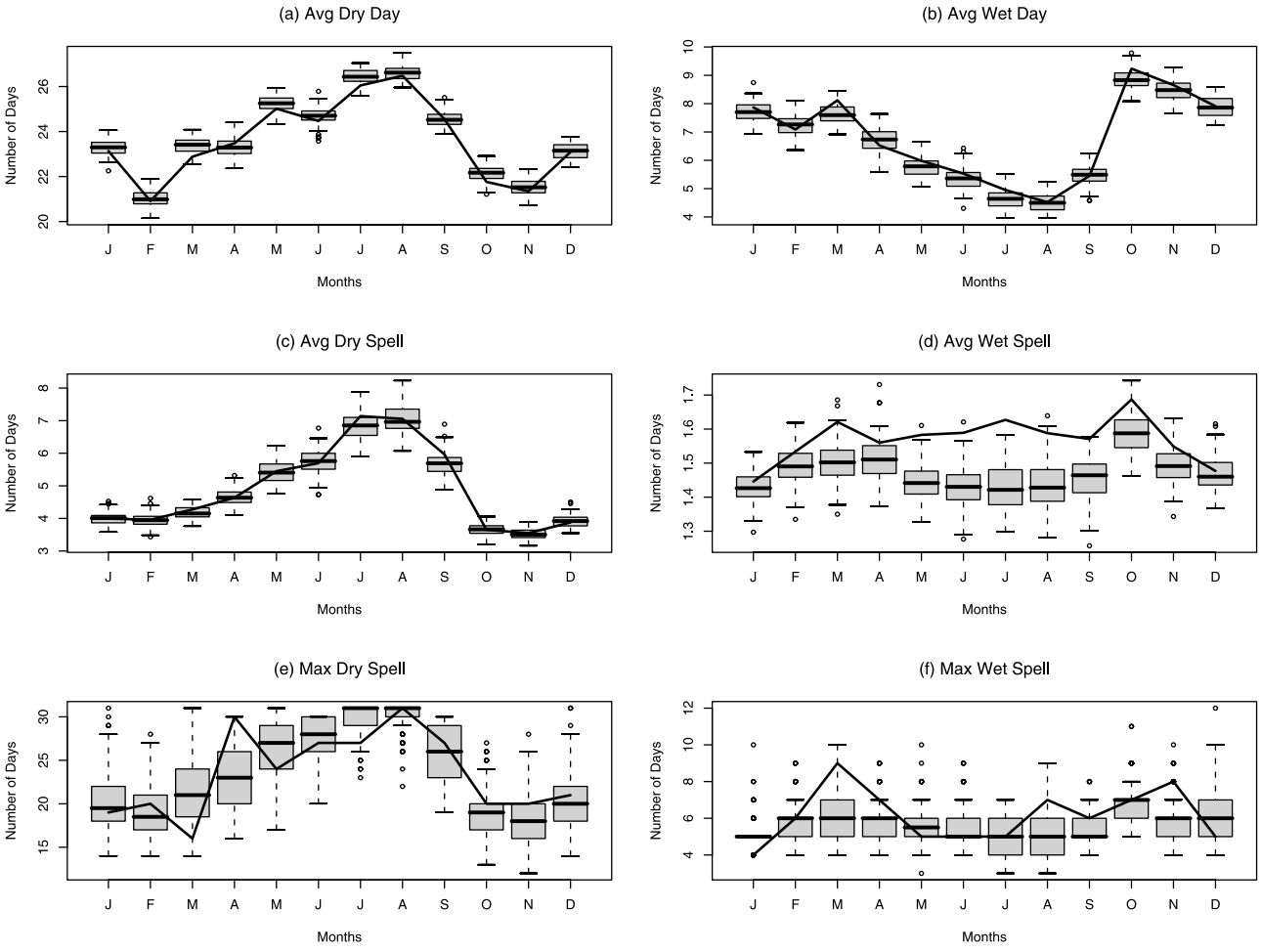

Figure 11. Same as Figure 10 but using the traditional *k*-NN model.

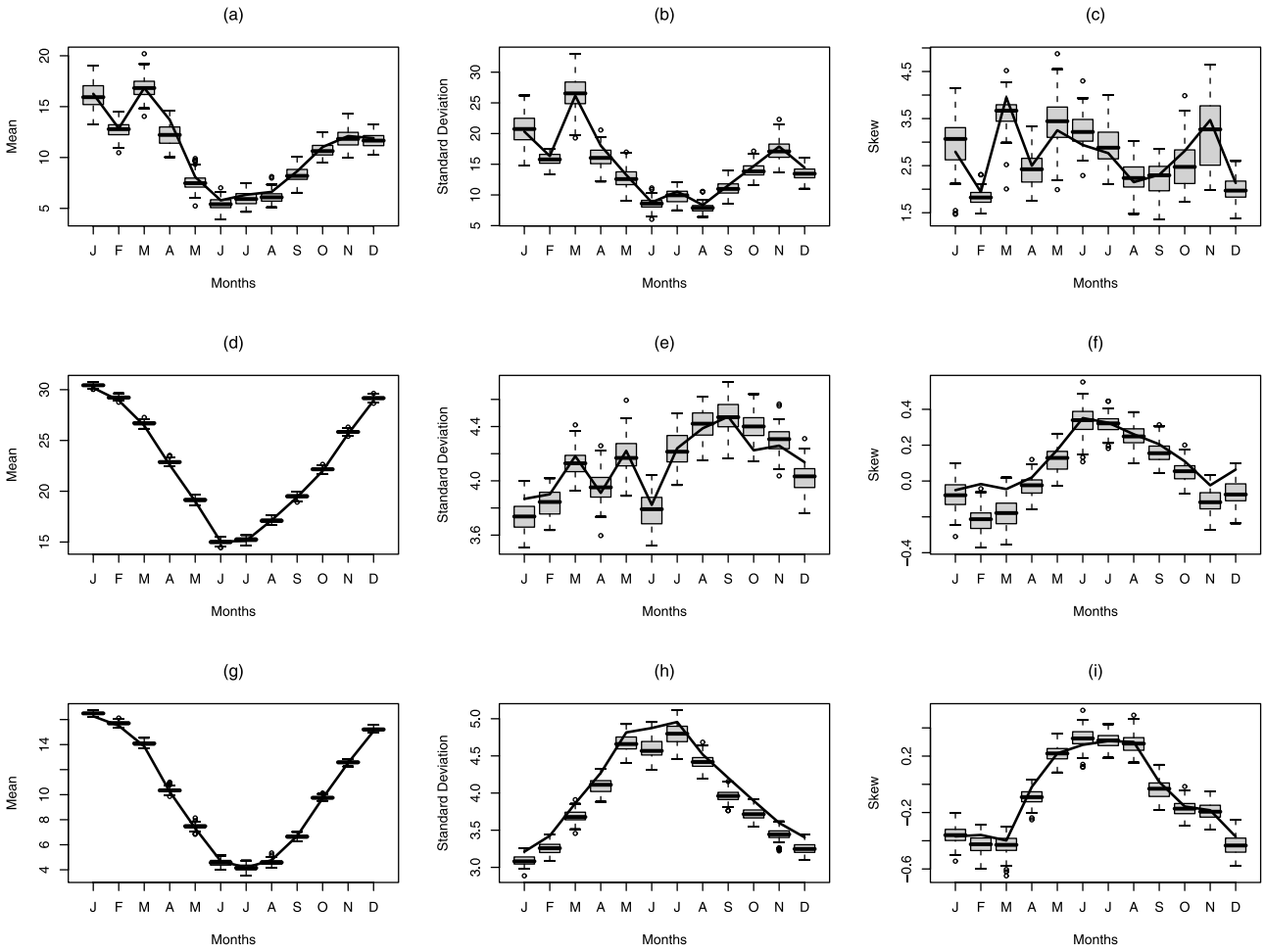

Figure 12. Basic distribution properties for station II (Junin) simulated from multisite generator using the semiparametric model. Statistics of daily precipitation (mm); mean, standard deviation, and coefficient of skew are shown in Figures 12a, 12b, and 12c, respectively. Similarly, Figures 12d, 12e, and 12f are the same as Figures 12a, 12b, and 12c but of maximum temperature (°C), and Figures 12g, 12h, and 12i are the same as Figures 12a, 12b, 12c but of minimum temperature (°C). Same format as in Figure 7.

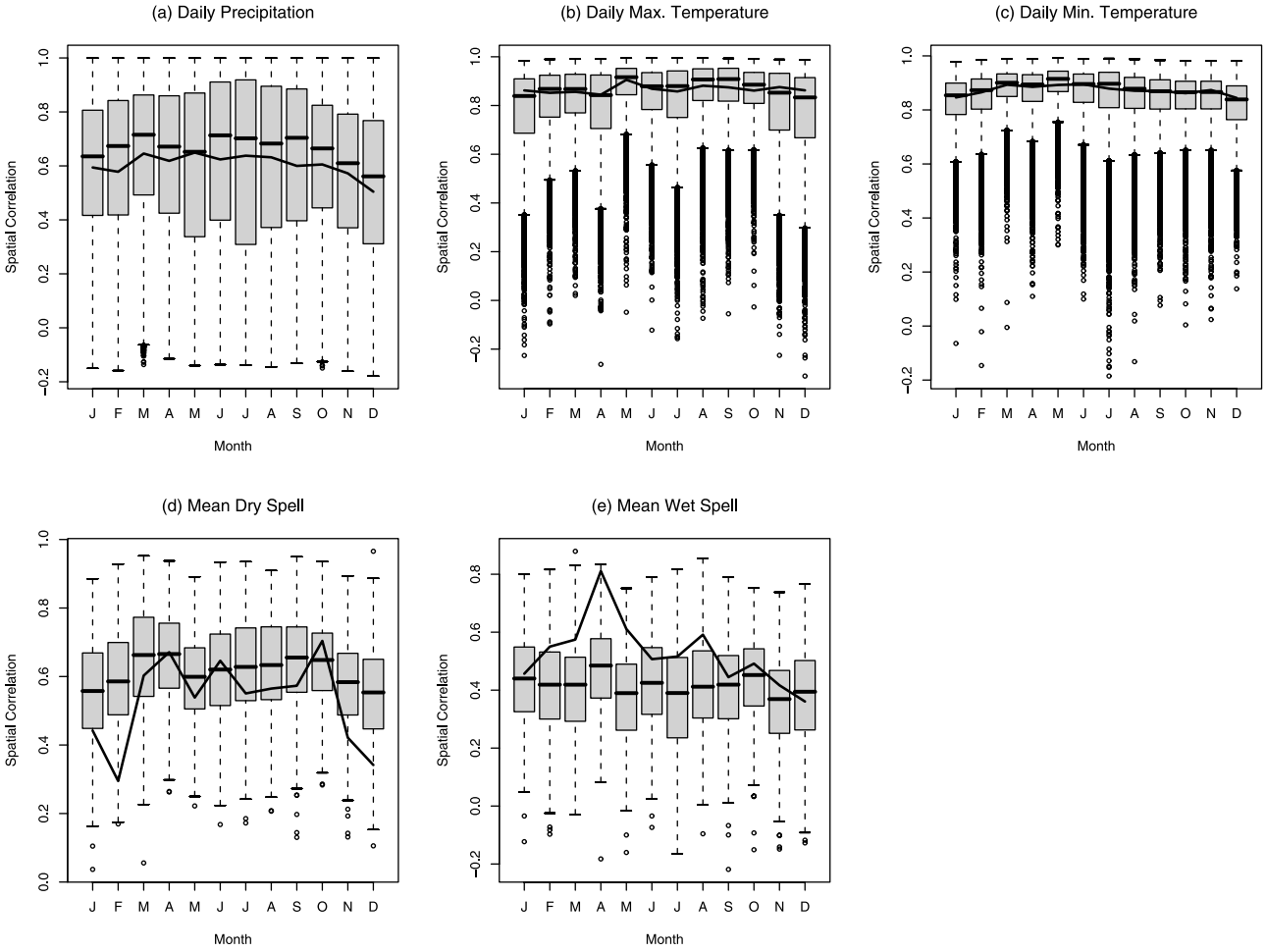

Figure 13. Regional spatial correlations as simulated from multisite generator shown in the form of monthly boxplots for daily precipitation (Figure 13a), daily maximum temperature (Figure 13b), and daily minimum temperature (Figure 13c), while Figures 13d and 13e are for monthly mean dry- and wet-spell lengths, respectively. Same format as in Figure 3.

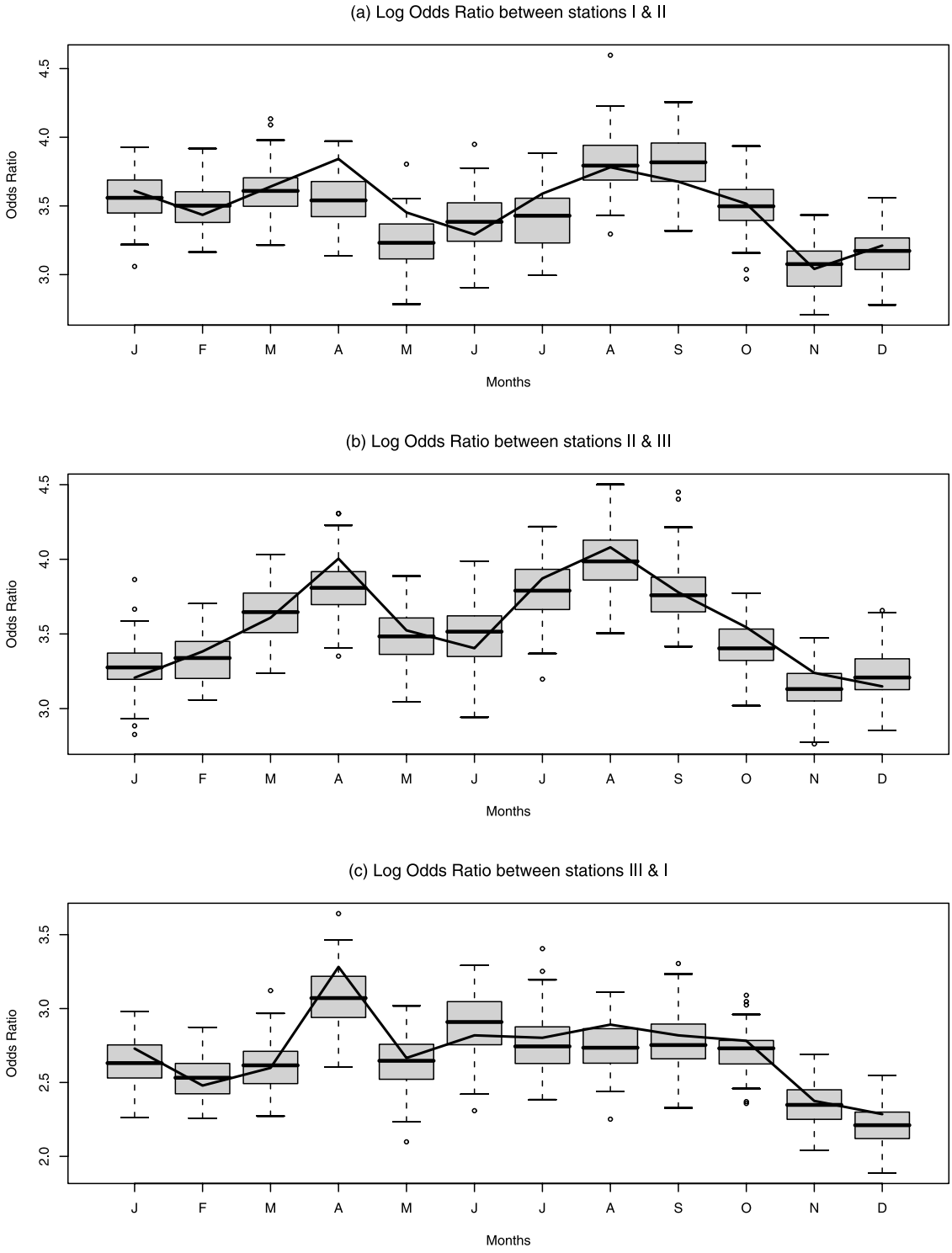

Figure 14. Log-odds ratios for precipitation occurrence at pairs of sites as simulated from multisite generator. Same format as in Figure 3.

first-order Markov chain to model the precipitation state. The precipitation state is first generated from the Markov chain, and the suite of daily weather variables are then generated from the *k*-NN resampling method conditioned on the precipitation state of the current and the next day. The Markov chain reproduces fairly well all the spell statistics, which is important in process models such as crop yield and hydrologic models. Furthermore, the three-state Markov chain seems to enable a better simulation of the statistics of the total annual precipitation. We also developed an approach to use this weather generator for simulating daily weather sequences conditioned on a seasonal climate forecast. Furthermore, the proposed method was extended to multisite simulation.

[43] We demonstrated the utility of the proposed weather generator to daily weather data from Pergamino, Argentina,

and compared the simulations from a traditional *k*-NN weather generator, showing that the spell statistics are better captured in the modified method. The conditional simulation also provided an appropriate shift in the PDF of the seasonal precipitation consistent with the observed. Of course, the skill of the conditional simulations will depend on the skill of the seasonal climate forecast. This approach has also been applied to simulating number of freezing and/or heavy precipitation days for modeling delays in construction due to extreme weather [*Xi et al.*, 2005]. The next step is to drive the process models (i.e., watershed model, crop model, etc.) by the weather sequences and to investigate the skill in the decision variables.

[44] Further investigations are needed to improve the proposed model, especially in the performance of multisite weather generation over diverse regions. Also, instead of spatially averaged daily time series to obtain the representative time series, other approaches such as weighted average should be considered to account for spatial heterogeneity. The neighbor selection and the definition of the distance [*Bárdossy et al.*, 2005] are also interesting ideas to adapt to weather generator context. Ability in generating values not seen in history is an attractive alternative for improving the *k*-NN part of our model as well [*Prairie et al.*, 2006]. There is great potential for combining parametric and resampling approaches especially for improving treatment of extremes for multisite weather generation. This needs to be explored.

[45] **Acknowledgments.** We gratefully acknowledge the funding of this research by the National Science Foundation's Biocomplexity in the Environment program grant BE-0410348. We also thank the three anonymous reviewers and the associate editor for their insightful comments and suggestions for improving the manuscript.

References

Bárdossy, A. (1998), Generating precipitation time series using simulated annealing, *Water Resour. Res.*, *34*(7), 1737–1744.

Bárdossy, A., G. G. S. Pegram, and L. Samaniego (2005), Modeling data relationships with a local variance reducing technique: Applications in hydrology, *Water Resour. Res.*, *41*, W08404, doi:10.1029/2004WR003851.

Beersma, J. J., and T. A. Buishand (2003), Multi-site simulation of daily precipitation and temperature conditional on the atmospheric circulation, *Clim. Res.*, *23*, 121–133.

Briggs, W. M., and D. S. Wilks (1996), Extension of the climate prediction center long-lead temperature and precipitation outlooks to general weather statistics, *J. Clim.*, *9*, 3496–3504.

Buishand, T. A., and T. Brandsma (2001), Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling, *Water Resour. Res.*, *37*(11), 2761–2776.

Caskey, J. E., Jr. (1963), A Markov chain model for the probability of precipitation occurrence in intervals of various lengths, *Mon. Weather Rev.*, *91*, 298–301.

Clark, M. P., S. Gangopadhyay, D. Brandon, K. Werner, L. Hay, B. Rajagopalan, and D. Yates (2004), A resampling procedure for generating conditioned daily weather sequences, *Water Resour. Res.*, *40*, W04304, doi:10.1029/2003WR002747.

Dubrovsky, M., Z. Zalud, and M. Stastna (2000), Sensitivity of CERES-Maize yields to statistical structure of daily weather series, *Clim. Change*, *46*, 447–472.

Efron, B. (1979), Bootstrap methods: Another look at the jackknife, *Ann. Stat.*, *7*, 1–26.

Foufoula-Georgiou, E., and K. P. Georgakakos (1991), Hydrologic advances in space-time precipitation modeling and forecasting, in *Recent Advances in the Modeling of Hydrologic Systems*, edited by D. S. Bowles and P. E. O'Connell, pp. 47–65, Springer, New York.

Furrer, E. M., and R. W. Katz (2007), Generalized linear modeling approach to stochastic weather generators, *Clim. Res*, in press.

Gabriel, R., and J. Neuman (1962), A Markov chain model for daily rainfall occurrence at Tel Aviv, Israel, *Q. J. R. Meteorol. Soc.*, *88*, 90–95.

Gangopadhyay, S., M. Clark, and B. Rajagopalan (2005), Statistical downscaling using K-nearest neighbors, *Water Resour. Res.*, *41*, W02024, doi:10.1029/2004WR003444.

Goddard, L., A. G. Barnston, and S. J. Mason (2003), Evaluation of the IRI's "net assessment" seasonal climate forecasts 1997–2001, *Bull. Am. Meteorol. Soc.*, *84*, 1761–1781.

Gregory, J. M., T. M. L. Wigley, and P. D. Jones (1993), Application of Markov models to area-average daily precipitation series and interannual variability in seasonal totals, *Clim. Dyn.*, *8*, 299–310.

Haan, C. T., D. M. Allen, and J. O. Street (1976), A Markov chain model of daily rainfall, *Water Resour. Res.*, *12*(3), 443–449.

Hardle, W., and A. W. Bowman (1988), Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands, *J. Am. Stat. Assoc.*, *83*, 102–110.

Hutchinson, M. F. (1987),Methods of generation of weather sequences, in *Agricultural Environments: Characterization, Classification and Mapping*, edited by A. H. Bunting, pp. 149–157, CAB Int., Cambridge, Mass.

Jones, W., R. C. Rex, and D. E. Threadgill (1972), A simulated environmental model of temperature, evaporation, rainfall, and soil moisture, *Trans. ASAE*, *15*, 366–372.

Katz, R. W. (1977), Precipitation as a chain-dependent process, *J. Appl. Meteorol.*, *16*(7), 671–676.

Katz, R. W., and M. B. Parlange (1995), Generalizations of chain-dependent processes: Application to hourly precipitation, *Water Resour. Res.*, *31*(5), 1331–1341.

Kottegoda, N. T., L. Natale, and E. Raiteri (2003), A parsimonious approach to stochastic multisite modelling and disaggregation of daily rainfall, *J. Hydrol.*, *274*, 47–61.

Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for time series resampling, *Water Resour. Res.*, *32*(3), 679–693.

Lettenmaier, D. (1995), Stochastic modeling of precipitation with applications to climate model downscaling, in *Analysis of Climate Variability: Applications of Statistical Techniques*, edited by H. von Storch and A. Navarra, pp. 197–212, Springer, New York.

Mehrotra, R., and A. Sharma (2006), Conditional resampling of hydrologic time series using multiple predictor variables: A k-nearest neighbour approach, *Adv. Water Resour.*, *29*, 978–999.

Nicks, A. D., and J. F. Harp (1980), Stochastic generation of temperature and solar radiation data, *J. Hydrol.*, *48*, 1–7.

Parlange, M. B., and R. W. Katz (2000), An extended version of the Richardson model for simulating daily weather variables, *J. Appl. Meteorol.*, *39*(5), 610–622.

Prairie, R. J., B. Rajagopalan, T. J. Fulp, and E. A. Zagona (2006), Modified K-NN model for stochastic streamflow simulation, *J. Hydrol. Eng.*, *11*, 371–378.

Qian, B., J. Corte-Real, and H. Xu (2002), Multisite stochastic weather models for impact studies, *Int. J. Climatol.*, *22*, 1377–1397.

Rajagopalan, B., and U. Lall (1999), A k-nearest-neighbor simulator for daily precipitation and other variables, *Water Resour. Res.*, *35*(10), 3089–3101.

Rajagopalan, B., U. Lall, D. G. Tarboton, and D. S. Bowles (1997), Multivariate nonparametric resampling scheme for generation of daily weather variables, *Stochast. Hydrol. Hydraul.*, *11*(1), 523–547.

Richardson, C. W. (1981), Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resour. Res.*, *17*(1), 182–190.

Semenov, M. A., and J. R. Porter (1995), Climatic variability and the modeling of crop yields, *Agric. For. Meteorol.*, *73*, 265–283.

Sharma, A., D. G. Tarboton, and U. Lall (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, *33*(2), 291–308.

Skidmore, E. L., and J. Tatarko (1990), Stochastic wind simulation for erosion modeling, *Trans. ASAE*, *33*(6), 1893–1899.

Stephenson, D. B. (2000), Use of the "odds ratio" for diagnosing forecast skill, *Weather Forecasting*, *15*, 221–232.

Trigo, R. M., and J. P. Palutikof (1999), Simulation of daily temperatures for climate change scenarios over Portugal: A neural network model approach, *Clim. Res.*, *13*, 45–59.

Wilby, R. W., O. J. Tomlinson, and C. W. Dawson (2003), Multi-site simulation of precipitation by conditional resampling, *Clim. Res.*, *23*, 183–194.

Wilks, D. S. (1997), Forecast value: Prescriptive decision studies, in *Economic Value of Weather and Climate Forecasts*, edited by R. W. Katz and A. H. Murphy, pp. 109– 145, Cambridge Univ. Press, New York.

Wilks, D. S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *J. Hydrol.*, *210*, 178–191.

Wilks, D. S. (1999a), Multisite downscaling of daily precipitation with a weather generator, *Clim. Res.*, *11*, 125–136.

Wilks, D. S. (1999b), Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain, *Agric. For. Meteorol.*, *96*, 85–101.

Wilks, D. S., and R. L. Wilby (1999), The weather generation game: A review of stochastic weather models, *Prog. Phys. Geogr.*, *23*, 329–357.

Woolhiser, D. A. (1992), Modeling daily precipitation: Progress and problems, in *Statistics in the Environmental and Earth Sciences*, edited by A. T. Walden and P. Guttorp, pp. 71–89, Edward Arnold, London.

Xi, Y., B. Rajagopalan, and K. Molenaar (2005), Quantify construction delays due to weather, final report of technology study, Dep. of Civ., Environ., and Architect. Eng., Univ. of Colo., Boulder.

Yakowitz, S. (1993), Nearest neighbor regression estimation for null-recurrent Markov time series, *Stochast. Processes Appl.*, *48*, 311–318.

Yakowitz, S., and M. Karlsson (1987), Nearest neighbor methods with application to rainfall/runoff prediction, in *Stochastic Hydrology*, edited by J. B. Macneil and G. J. Humphries, pp. 149–160, Springer, New York.

Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek (2003), A technique for generating regional climate scenarios using a nearest–neighbor algorithm, *Water Resour. Res.*, *39*(7), 1199, doi:10.1029/2002WR001769.

Young, K. C. (1994), A multivariate chain model for simulating climatic parameters from daily data, *J. Appl. Meteorol.*, *33*, 661–671.

Zucchini, W., and P. T. Adamson (1989), Bootsfstrap confidence intervals for design storms from exceedance series, *Hydrol. Sci. J.*, *34*(1), 41–48.

S. Apipattanavis, Office of Research and Development, Royal Irrigation Department, Nonthaburi 11120, Thailand. (apipatta@colorado.edu)

R. W. Katz, National Center for Atmospheric Research, Boulder, CO 80307, USA.

G. Podestá, Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, FL 33149, USA.

B. Rajagopalan, Department of Civil, Environmental and Architectural Engineering, University of Colorado, Boulder, CO 80309, USA.