Vol. 34: 129–144, 2007

CLIMATE RESEARCH
Clim Res

Published July 19

Generalized linear modeling approach to stochastic weather generators

Eva M. Furrer*, Richard W. Katz

National Center for Atmospheric Research, PO Box 3000, Boulder, Colorado 80307-3000, USA

ABSTRACT: Stochastic weather generators are a popular method for producing synthetic sequences of daily weather. We demonstrate that generalized linear models (GLMs) can provide a general modeling framework, allowing the straightforward incorporation of annual cycles and other covariates (e.g. an index of the El Niño-Southern Oscillation, ENSO) into stochastic weather generators. We apply the GLM technique to daily time series of weather variables (i.e. precipitation and minimum and maximum temperature) from Pergamino, Argentina. Besides annual cycles, the fit is significantly improved by permitting both the transition probabilities of the first-order Markov chain for daily precipitation occurrence, as well as the means of both daily minimum and maximum temperature, to depend on the ENSO state. Although it is more parsimonious than typical weather generators, the GLM-based weather generator performs comparably, particularly in terms of extremes and overdispersion.

KEY WORDS: Weather generator · Generalized linear models · GLMs · El Niño-Southern Oscillation · ENSO

Resale or republication not permitted without written consent of the publisher

1. INTRODUCTION

Stochastic weather generators are a popular method for producing synthetic sequences of daily weather, particularly minimum and maximum temperature and precipitation amount (Wilks & Wilby 1999). Such sequences are used, for instance, as inputs to crop–climate models to study the effect of climate variability on crop yields (Grondona et al. 2000). The goal is to capture the basic statistical features of daily weather variables, especially temporal dependence for individual variables and contemporaneous dependence between variables. Generally based either on parametric models (Richardson 1981) or on resampling (Rajagopalan & Lall 1999), or on some combination of these 2 methods (Apipattanavis et al. 2007), complications include the need to incorporate annual cycles and the desire to condition the model on large-scale atmospheric or oceanic circulation patterns such as the El Niño-Southern Oscillation (ENSO) phenomenon (Podestá et al. 1999).

Generalized linear models (GLMs) can greatly simplify the effort involved in the stochastic modeling of daily weather variables. An extension of the more familiar multiple regression analysis, GLMs can handle variables with non-normal distributions (e.g. amount of precipitation on a wet day or 'intensity') as well as discrete variables (e.g. precipitation occurrence). As in conventional time series analysis, temporal dependence can be incorporated through the introduction of lagged variables as predictors, termed 'covariates' in the field of statistics. For a detailed treatment of GLMs, see McCullagh & Nelder (1989); for a user-oriented summary, see Venables & Ripley (2002).

Stern & Coe (1984) demonstrated that it is relatively straightforward to model a time series of daily precipitation using GLMs, even in a region with a very marked wet season. Their GLM is equivalent to a 2-state Markov chain model for daily precipitation occurrence and a gamma distribution for the distribution of daily precipitation intensity. Annual cycles in both of these components are represented through the use of sine waves as covariates.

Recently, the GLM approach to the stochastic modeling of daily weather variables has been revisited, showing that it can be applied to essentially any variable, including temperature and wind speed (Yan et al.

*Email: eva@ucar.edu

© Inter-Research 2007 · www.int-res.com

2002, Chandler 2005). Examples are provided in which the covariates need not be restricted to simple deterministic functions, such as sine waves to account for annual cycles or trends to account for climate change, but can be geophysical variables such as an index of the North Atlantic Oscillation (Chandler & Wheater 2002). The GLM approach has the advantage of being able to treat circulation indices as continuous variables, avoiding the non-parsimonious approach of fitting different stochastic models for the daily weather variable conditional on a categorization of the index into a few, somewhat arbitrary, discrete states (e.g. Katz & Parlange 1996).

In the present study, we show that the GLM approach can be applied to parametric weather generators, modeling more than one daily weather variable at a given site simultaneously. The basic form of the weather generator originally proposed by Richardson (1981) is adopted. Daily precipitation occurrence is modeled as a first-order Markov chain and daily precipitation intensity is modeled using a gamma distribution, with the GLM approach to fitting the precipitation component being essentially the same as in Stern & Coe (1984). As in the Richardson model, the conditional means of minimum (Tmin) and maximum (Tmax) daily temperature are permitted to depend on whether or not precipitation occurs. So that covariates can be more readily introduced, coupled univariate models are used for daily Tmin and Tmax, essentially equivalent to the bivariate first-order autoregressive process in the Richardson model (Richardson 1981). Note that the GLM framework also provides a straightforward possibility to model precipitation conditional on temperature.

We apply the GLM technique to fit a stochastic weather generator to time series of daily precipitation, Tmin, and Tmax from Pergamino, Argentina; a region with a marked wet season and known teleconnections with ENSO (Grondona et al. 2000). A stochastic weather generator will be used at this location in a multidisciplinary project concerned with assessing the economic impact of interannual variations in climate, particularly those associated with ENSO, on agriculture in the Pampas region of Argentina ('Climate, Agriculture, and Complexity in the Argentine Pampas' www.rsmas.miami.edu/groups/agriculture/, Letson et al. 2005). In all components of this GLM-based weather generator, annual cycles are modeled and ENSO is considered as a possible covariate. Because the parameters of a GLM are estimated by maximum likelihood, another advantage of the GLM approach is that testing whether covariates improve the fit is straightforward (Chandler 2005).

Besides providing a simplified framework for the development of stochastic weather generators, the approach has the additional benefit of being readily amenable to uncertainty analysis (e.g. taking into account the source of uncertainty attributable to estimating the parameters of the model from a limited sample). The appropriate treatment of uncertainty is an important issue in assessments of the impacts of climate variability and change (Katz 2002). For these reasons, it is anticipated that the framework presented here will be especially appealing to the climate impacts research community.

2. GLM APPROACH

This section starts by giving a very brief background to GLMs and then describes in more detail the modeling strategy for each of the 3 parts of the proposed weather generator: precipitation occurrence, precipitation intensity, and daily Tmin and Tmax.

2.1. Background

GLMs are used to link response variables to covariates, i.e. explanatory variables, for Gaussian and non-Gaussian settings. Similar to simple linear models, the probability distribution of the response variable is modeled through the dependence of its conditional mean on covariates. In contrast to simple linear models, the response variable does not necessarily need to be Gaussian, and the link between mean and covariates can be given by e.g. the logarithm or another suitable function ('link function'). Obviously, simple linear models are a special case of this larger class of statistical models.

Considering the response variable Y, e.g. precipitation or temperature, and covariates $X_1,...,X_p$, e.g. seasonal cycles, climate indices, etc., a simple linear model assumes Y to be Gaussian, its conditional mean (E) to be given by

$$E(Y|X_1 = x_1,...,X_p = x_p) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \quad (1)$$

where β_0, $\beta_1 ... \beta_p$ are parameters and the conditional variance to be independent of the covariates. In contrast, a generalized linear model assumes Y to have a distribution from a certain family, containing among others such diverse distributions as the gamma and the binomial distribution as well as the normal distribution. The conditional mean of Y is then given by

$$g\left[E(Y|X_1 = x_1,...,X_p = x_p)\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \quad (2)$$

where the so-called link function g is, for example, chosen to constrain the range to permissible values. Generally, the conditional variance is a function of the conditional mean and hence depends on the covariates and the regression coefficients. Therefore, it is necessary to use a weighting scheme for the estimation of the parameters by least squares.

Parameter estimates for GLMs are obtained using iterative weighted least squares, since the weights at each step can only be estimated using current estimates of the regression coefficients. A variant of the Newton-Raphson algorithm is used for this, replacing the Hessian by its expected value. Convergence of the algorithm is judged by a criterion based on the likelihood. This approach is equivalent to maximum likelihood estimation for a certain family of distributions, as are ordinary least squares for simple linear models. All model fitting in this study is done with R (free software environment for statistical computing and graphics), using the functions 'lm' and 'glm', see R Development Core Team (2005).

Our objective is to consider the effects of different covariates in the models for precipitation and temperature. Which set of covariates best fits the data is decided on the basis of the Bayesian information criterion (BIC). This technique balances the better fit of a more sophisticated model against the increase in the number of parameters to be estimated, the model with the smallest BIC value being preferred (Schwarz 1978). More precisely, the BIC is given by

$$\mathrm{BIC(model)} \;=\; -2\text{log-likelihood(model)} + p\ln(\mathrm{n}) \qquad (3)$$

where p is the number of estimated parameters, and n the sample size. Note that the results of the application to weather data from Argentina would essentially be the same if using other model selection criteria such as the Akaike information criterion (AIC) or a likelihood ratio test. BIC selects (slightly) less complex models only in a few cases, some of them mentioned in Section 5.

2.2. Precipitation occurrence

Precipitation occurrence is modeled as a first-order Markov chain. This essentially means that the conditional probability of occurrence of precipitation on a specific day only depends on the occurrence of precipitation the day before. Denoting by $J_t = 1$ that it rained on Day t and by $J_t = 0$ that it did not, the first-order Markov chain model is characterized by the transition probabilities

$$p_{ij}(t) \;=\; \Pr(J_t \;=\; j|J_{t-1} \;=\; i), \quad i,j \;=\; 0,1 \qquad (4)$$

where Pr denotes probability. The appropriate model for occurrence data, i.e. a 0 or 1 variable, is a binomial GLM and we choose to use a logistic link function, $g(x) = \ln[x(1-x)^{-1}]$. The expected value of a binomial distribution with a single trial (also called a Bernoulli distribution) is the underlying success probability, i.e. in our case, the probability of rain.

We use a binomial GLM with logistic link function to link covariates to precipitation occurrence via p_t the conditional probability of rain on Day t, see also Chandler & Wheater (2002). Our primary covariate is the occurrence of rain on the previous day J_{t-1}, such that we actually have a first-order Markov chain. Further covariates, such as a seasonal cycle or a climatic index, e.g. the ENSO index, can be introduced sequentially and the BIC criterion is used to decide whether they should be part of the model or not. Additionally, interaction terms between these covariates and the occurrence of rain on the previous day, J_{t-1}, allow different effects of covariates depending on the value of J_{t-1}, i.e. depending on whether it rained the previous day or not. Interactions of a seasonal cycle with J_{t-1}, for example, allow for $p_{01}(t)$ and $p_{11}(t)$ to have different seasonal cycles. More precisely, p_t is linked to J_{t-1} and the vector of covariates $\mathbf{Z}_t$, through the equation

$$\ln\left(\frac{p_t}{1-p_t}\right) \;=\; \mu + \alpha J_{t-1} + \mathbf{Z}_t'\boldsymbol{\beta} + J_{t-1}\mathbf{Z}_t'\boldsymbol{\gamma} \qquad (5)$$

where μ, α and the vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are parameters and the prime symbol (′) denotes transpose. Note that we do not necessarily need to consider interaction terms of all components of $\mathbf{Z}_t$ with J_{t-1} as suggested by these equations. By setting $J_{t-1} = i$, $i = 0,1$ in this equation and rearranging, we retrieve the transition probabilities of our Markov chain

$$p_{i1}(t) \;=\; \frac{\exp(\mu + \alpha i + \mathbf{Z}_t'\boldsymbol{\beta} + i\mathbf{Z}_t'\boldsymbol{\gamma})}{1+\exp(\mu + \alpha i + \mathbf{Z}_t'\boldsymbol{\beta} + i\mathbf{Z}_t'\boldsymbol{\gamma})} \quad i = 0,1 \qquad (6)$$

The connection between the transition probabilities $p_{ij}(t)$ and the unconditional probability of rain on Day t, as well as the first-order autocorrelation coefficient, is discussed in Appendix 1; see also Katz & Parlange (1995). Obviously, it is straightforward to introduce a second or even higher-order Markov structure within the GLM framework.

2.3. Precipitation intensity

Precipitation intensity I_t on wet days is modeled using a gamma GLM with a logarithmic link function, assuming its conditional distribution has a constant shape parameter, independent of the vector of covariates $\mathbf{Z}_t$ and the mean. In other words, I_t is considered to be gamma distributed with conditional mean μ_t on Day t given by

$$\ln(\mu_t) \;=\; \mu + \mathbf{Z}_t'\boldsymbol{\beta} \qquad (7)$$

where t is such that $J_t = 1$ and μ and the vector $\boldsymbol{\beta}$ are parameters. It is commonly assumed that, conditional on precipitation occurrence, the intensity on a given day does not depend on the intensity or occurrence the day before, i.e. we do not consider an autoregressive-type term for precipitation intensity (but it would be straightforward to incorporate this feature into the GLM framework). The probability density function of the gamma distribution is given by

$$f(x) \;=\; \frac{1}{s^a\Gamma(a)}x^{a-1}\exp\left(-\frac{x}{s}\right),\ \text{for } x > 0,\ a > 0,\ s > 0 \qquad (8)$$

Here, s is a scale parameter governing the spread of the distribution, a is a shape parameter mainly being used to set the degree of skewness and Γ is the gamma function. The mean of the gamma distribution is as and the variance as^2. In this study the shape parameter is estimated by maximum likelihood, see e.g. McCullagh & Nelder (1989; their Section 8.3.6). Again, covariates, such as a seasonal cycle or a climatic index, but not necessarily the same as for precipitation occurrence, can be introduced sequentially and the BIC criterion is used for model selection.

2.4. Minimum and maximum daily temperature

Tmin and Tmax are modeled using separate first-order autoregressive AR(1) processes with covariates, which are coupled through the introduction of lagged values of the respective other temperature variable as a covariate. Consequently, we assume that the conditional distribution of the temperature variables at Time t depends only on their value at Time t – 1 and that, given some appropriate covariates such as a seasonal cycle, they are normally distributed. The coupling approach is a simplified way to introduce dependence between Tmin and Tmax, which allows us to consider other types of covariates more easily than in a classical bivariate setting, as, e.g. in Parlange & Katz (2000). Let X_t denote Tmin and Y_t denote Tmax on Day t, then we write the coupled model as

$$X_t \;=\; \mu_{X,0} + \mu_{X,1}J_t + \phi_X X_{t-1} + \psi_X Y_{t-1} + \mathbf{Z}_t'\boldsymbol{\beta}_X + \varepsilon_{X,t} \qquad (9)$$

$$Y_t \;=\; \mu_{Y,0} + \mu_{Y,1}J_t + \phi_Y Y_{t-1} + \psi_{Y},X_{t-1} + \mathbf{Z}_t'\boldsymbol{\beta}_Y + \varepsilon_{Y,t} \qquad (10)$$

where the error terms $\varepsilon_{X,t}$ and $\varepsilon_{Y,t}$ are uncorrelated (i.e. no cross correlation), zero-mean normal errors with constant conditional variances $\sigma_X{}^2$ and $\sigma_Y{}^2$ (i.e. independent of t), additional covariates, such as a seasonal cycle or a climate index, are given by the vector $\mathbf{Z}_t$. The parameters $\mu_{X,1}$ and $\mu_{Y,1}$ allow for the mean Tmin and Tmax, respectively, to be dependent on whether or not precipitation occurs. The parameters ϕ_X and ϕ_Y allow for dependence on the same temperature variable on the previous day, they correspond to first-order autocorrelation coefficients conditionally on all the other variables in the models (i.e. they are likewise assumed independent of t). The parameters ψ_X and ψ_Y allow for dependence between Tmin and Tmax, they correspond to the lag 1 cross correlation for Tmin and the lag 0 cross correlation for Tmax conditionally on all the other variables in the models. The influence of the covariates $\mathbf{Z}_t$ on the temperature variables and the interpretation of the coefficients in the vectors $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Y$ is rather complex. For example, the influence of $\mathbf{Z}_t$ on X_t is described by $\boldsymbol{\beta}_X$ but there is indirect influence via ϕ_X, ψ_X, $\mu_{X,1}$ and $\boldsymbol{\beta}_Y$ as well.

This modeling strategy differs from the formulation of the Richardson model in several ways. We include covariates and therefore systematically model seasonal cycle and ENSO effect as opposed to fitting separate models for each month or for each ENSO category. The proposed model is not formally presented as a bivariate AR(1) process, even though—leaving out covariates—the 2 formulations are essentially equivalent. As a consequence, we do not use Yule-Walker equations to estimate the parameters, as is usually done within the classical Richardson model, but fit the 2 models separately via least squares.

The coupled modeling idea has an order aspect to it, since Tmax is modeled subsequently to Tmin. The reasoning is that Tmin will be observed in the very early morning hours of each Day t and Tmax usually in the early afternoon. This means that the Tmin of Day t is closer in a sense of temporal distance to the Tmax of Day t – 1 than the Tmax of Day t is to the Tmin of Day t – 1.

One of the potential disadvantages of the coupled modeling approach, compared to a classical bivariate approach, is that covariates need to be selected separately for Tmin and Tmax, possibly leading to 2 different sets of covariates. Again, for the 2 separate models, the BIC criterion is used for model selection and the selected covariates do not need to be the same as for precipitation occurrence or intensity.

Note that this approach does not guarantee that Tmax is higher than Tmin. One way to guarantee this property from a modeling point of view would be to consider a temperature range (maximum minus minimum) and a temperature mean (average of maximum and minimum) with a gamma and a normal model respectively (Jolliffe & Hope 1996). The disadvantage of this idea is the loss of the simple AR(1) model, which is based on Gaussianity, although it would still be possible to implement within the GLM framework. As can be seen in Section 4.1, the occurrence of days for which the simulated Tmin is higher than the simulated Tmax is very rare and we chose to retain the simpler model and deal with this issue in an ad-hoc way.

3. APPLICATION

This section discusses the application of our proposed weather generator to a set of weather data from

Table 1. Notation and description of the variables used

Symbol	Description
J_t	Precipitation occurrence on Day t
I_t	Precipitation intensity on Day t
X_t	Tmin on Day t
Y_t	Tmax on Day t
C_t	$\cos(2\pi \times t/365)$
S_t	$\sin(2\pi \times t/365)$
E_t	ENSO index on Day t

Argentina. We start by briefly describing the data we used and then address each part of the weather generator: precipitation occurrence, precipitation intensity, and temperature, in a separate subsection. For each part, the fundamental model assumptions: first-order Markov for precipitation occurrence, auto-regression for the temperature variables, and the distributional assumptions of gamma and normal distributions for precipitation intensity and temperature variables, respectively, have been checked and are sufficiently well satisfied for these data. Table 1 provides a list of the notation of variables used in the following analyses.

3.1. Data

The GLM approach to weather generation has been tested using a series of daily precipitation (mm) and daily Tmin and Tmax (°C) at Pergamino (33° 56′ S, 60° 33′ W) in the Pampas region of central-eastern Argentina. The annual precipitation cycle at Pergamino has a clear maximum in late spring and summer and a marked winter minimum. Full years of data are available for 1932 to 2003, several years have been excluded from the analysis since they contain too many missing values (1954 to 1957 and 1963 to 1966), such that a total of 63 yr of data have been analyzed. There are a few more missing values in the rest of the record period, more so for the temperature variables than for precipitation, but they are scarce enough that the results concerning the relationship with ENSO are not affected and they certainly do not prevent the GLM framework from being used. Data corresponding to February 29 of leap years have been removed for simplicity. The ENSO index is based on monthly mean sea surface temperature anomalies (°C) for Niño Region 3 (bounded by 90° W–150° W and 5° S–5° N), see Kaplan et al. (1998). In our context we use daily weather data and therefore assume the ENSO index to be constant over each month of the record period.

3.2. Precipitation occurrence

Besides a mean signal and the occurrence of rain on the previous day, we introduce a seasonal cycle using $C_t = \cos(2\pi \times t/365)$ and $S_t = \sin(2\pi \times t/365)$, as well as the ENSO index E_t to the binomial GLM for precipitation occurrence. Additionally, we consider interaction terms between the seasonal cycle and the previous day's occurrence, i.e. $J_{t-1}C_t$ and $J_{t-1}S_t$:

$$\ln\left(\frac{p_t}{1-p_t}\right) = \mu + \alpha J_{t-1} + \beta_1 C_t + \beta_2 S_t + \beta_3 E_t + \gamma_1 C_t J_{t-1} + \gamma_2 S_t J_{t-1} \tag{11}$$

All of the mentioned covariates are selected using the BIC criterion, values of BIC along with estimated coefficients of the selected model are given in Table 2. Due to the greatest decrease in BIC when introducing

Table 2. Estimated coefficients (Coef.) and associated Bayesian information criterion (BIC) values for all components of the stochastic weather generator; each row refers to the model containing the terms in all preceding rows. BIC values are given in **bold** for conceptually necessary parts of the model. Parentheses: non-selected parts of a model. Chosen models are independent of the order of entry of predictors

Covariate	Occurrence			Intensity[a]			Minimum			Maximum		
category	Term	Coef.	BIC	Term	Coef.	BIC	Term	Coef.	BIC	Term	Coef.	BIC
Mean	μ	–1.57	**24336.6**	μ	2.44	**34610.1**	μ	–2.77	**147511.8**	μ	9.20	**151162.4**
Autocorrelation	J_{t-1}	1.12	23399.2	–	–	–	X_{t-1}	0.42	**122839.8**	Y_{t-1}	0.52	**120729.7**
Dependence	–	–	–	–	–	–	Y_{t-1}	0.36	**116605.1**	X_t	0.23	**119257.5**
	–	–	–	–	–	–	J_t	1.89	115266.8	J_t	–1.84	117905.6[b]
Season	C_t	0.45	23277.3	C_t	0.29	34497.2	C_t	0.74	114945.3	C_t	2.21	115711.2
	S_t	0.01		S_t	0.12		S_t	0.37		S_t	0.38	
Interaction	$C_t J_{t-1}$	–0.56	23174.7	–	–	–	–	–	–	–	–	–
	$S_t J_{t-1}$	0.01		–	–		–	–		–	–	
ENSO	E_t	0.07	23275.4	E_t	(0.16)	(34505.2)	E_t	0.20	114882.0	E_t	–0.15	115685.2

[a] Estimated shape parameter of the gamma distribution: 0.59

[b] For maximum temperature BIC would decrease more if introducing the seasonal cycle before precipitation (116950.6)

the seasonal cycle, we can conclude that it is the strongest of the considered signals. Because of the interaction terms, $p_{11}(t)$ and $p_{01}(t)$ are allowed different cyclic behavior in this model: seasonal signal = $(\beta_1+\gamma_1)C_t$ + $(\beta_2+\gamma_2)S_t$ for $p_{11}(t)$; and seasonal signal = $\beta_1 C_t$ + $\beta_2 S_t$ for $p_{01}(t)$.

Note that an interaction term between the ENSO index and the previous day's occurrence turned out not to be selected by BIC, i.e. the influence of the ENSO index on the probability of rain is modeled as the same for both previous wet and dry days.

The modeled effects on the probability of a transition from a wet day to a wet day, $p_{11}(t)$, can be 'predicted' as a function of the time of the year and of the value of the ENSO index by setting the previous day's precipitation occurrence to 1 in the fitted model. Similarly, the effects on $p_{01}(t)$ can be 'predicted' by setting the previous day's precipitation occurrence to 0. Since we use the fitted model in an unobserved situation, we talk of 'predicting' these probabilities, because it is not guaranteed that the ENSO index assumes all of the observed values on each day of the year.

Fig. 1 shows the different seasonal cycles as well as the different magnitudes of $p_{11}(t)$ and $p_{01}(t)$. The effect of the ENSO index on both probabilities is such that higher (positive) ENSO values induce a higher probability of rain regardless if the day before was wet or dry. This behavior could be interpreted in the sense that it is raining more frequently during El Niño events, which correspond to positive values of the ENSO index, and it is raining less frequently during La Niña events, which correspond to negative values of the ENSO index.

Fixing the value of the ENSO index to 0 we can again predict both transition probabilities for the entire year. The resulting curves correspond to the 'neutral' situation of no effect of the ENSO index. Empirical transition probabilities, i.e. frequencies of observed transitions calculated separately on each day of the year from the 63 yr of data, and a smoothed version thereof, are shown together with the modeled transition probabilities.

Fig. 2 illustrates more easily the different seasonal cycles as well as the different scales of $p_{11}(t)$ and $p_{01}(t)$. $p_{01}(t)$ has a more pronounced dependency on the season, the chances of rain after a dry day being smaller in (austral) winter than in (austral) summer. On the other hand $p_{11}(t)$ is a comparatively flat function of the day of the year, the chances of rain after an already wet day being higher in winter. The empirical transition probabilities and more so their smoothed versions substantiate our modeling strategy since magnitude and shape throughout the year of these very crude, empirical estimators of the transition probabilities agree very well with their modeled counterparts.

Similarly, running the recursive Eq. (A1) of Appendix 1 for a few years and using the values of $p_{11}(t)$ and $p_{01}(t)$ together with a starting value of 0.5, we predict the unconditional probability of rain and the first-order autocorrelation coefficient for a fixed ENSO index of 0 and for the entire year. Again, empirical probabilities (frequencies of rain on each day of the year) and empirical autocorrelation coefficients (Pearson's correlation coefficient between occurrence on consecutive days on each day of the year) and smoothed versions thereof are shown together with the model curves.

Fig. 3 shows that rain is again, as expected, less probable in winter than in summer. The first-order autocorrelation is stronger in winter than in summer, meaning that in winter the occurrence of rain more

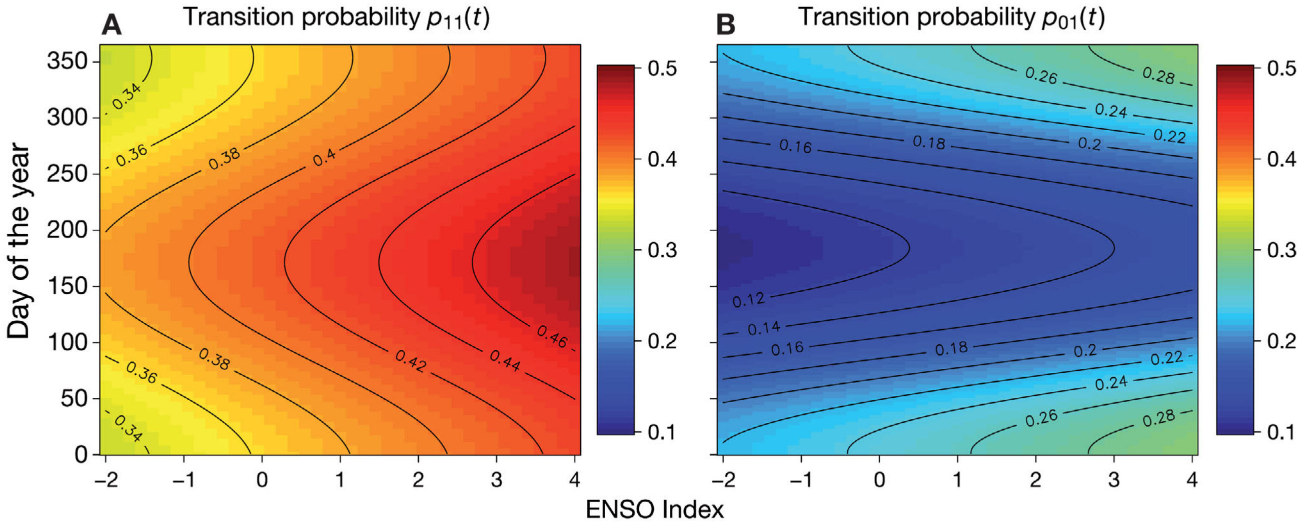

Fig. 1. Modeled transition probabilities (A) $p_{11}(t)$ and (B) $p_{01}(t)$ as functions of the day of the year and of the range of the values of the ENSO index

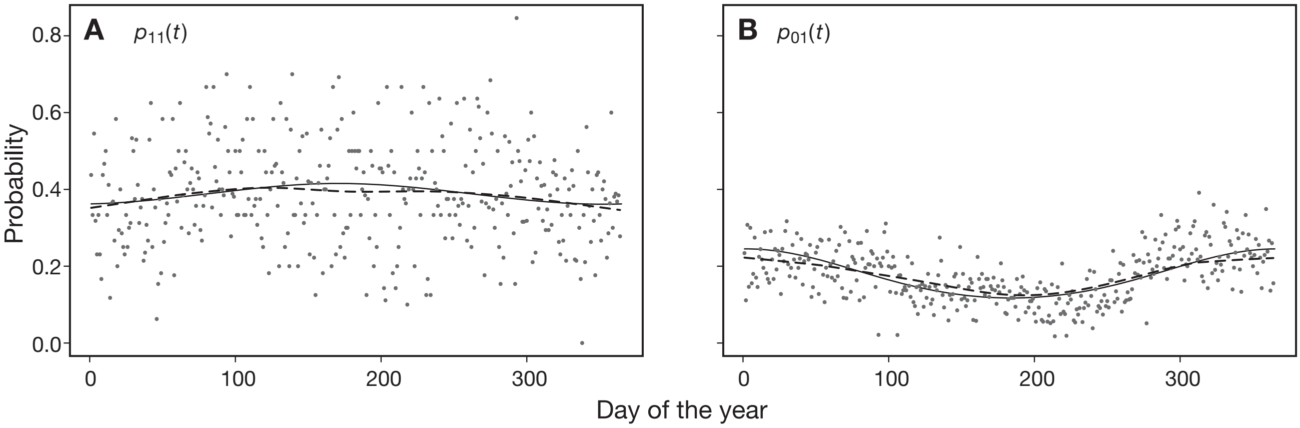

Fig. 2. Modeled transition probabilities (A) $p_{11}(t)$ and (B) $p_{01}(t)$ for ENSO = 0 (solid lines). Dots: empirical transition probabilities, i.e. frequencies of observed transitions calculated separately on each day of the year from the 63 yr of data; dashed lines: smoothed empirical values

strongly depends on the previous day's occurrence, or from the other viewpoint, in summer it can rain almost independently from the previous day being rainy or not. In both cases, the empirical values and their smoothed versions again indicate the validity of our modeling strategy.

3.3. Precipitation intensity

As above, we consider the seasonal cycle given by C_t and S_t and the ENSO index E_t as covariates in the gamma GLM

$$\ln(\mu_t) = \mu + \beta_1 C_t + \beta_2 S_t + \beta_3 E_t \qquad (12)$$

The BIC selects the model containing a mean signal as well as the seasonal cycle, the introduction of the ENSO index leads to an increase in BIC. Values of BIC along with estimated coefficients of the selected model are given in Table 2; for the ENSO index the estimated coefficient corresponds to the (non-selected) model (Eq. 12). The mean and standard deviation (SD) of precipitation intensity as a function of the day of the year are given in Fig. 4. As expected, mean intensity is lower, and intensity is less variable during the winter months. The empirical means as well as the empirical SD calculated separately for each day of the year from the 63 yr of data and their smoothed version indicate a good agreement of our model with the data, although it seems that the modeled mean as well as the modeled SD might be a little too high during approximately November to February. Note that we do not explicitly model the SD, the functional form is induced by the model for the mean (see Section 2.3).

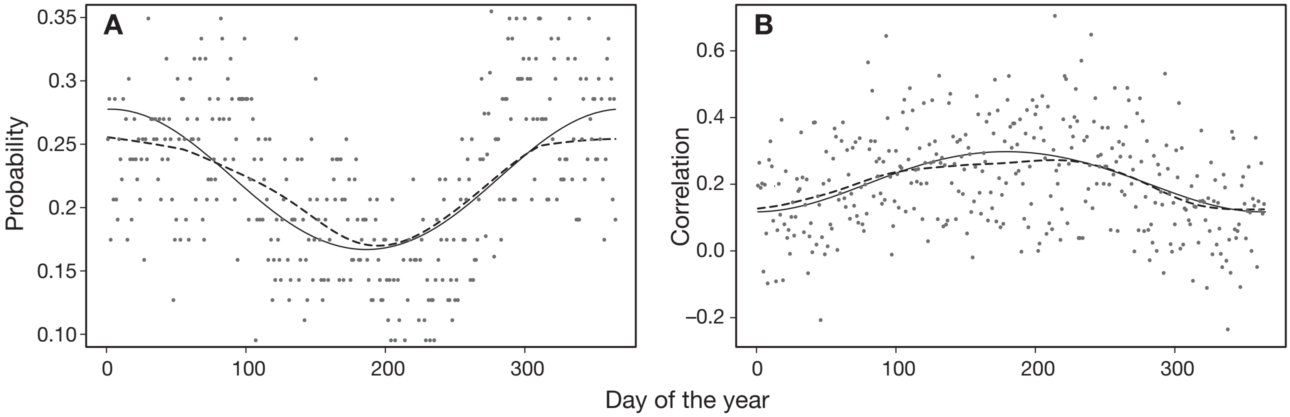

Fig. 3. Modeled (A) unconditional probability of rain and (B) first-order autocorrelation coefficient for ENSO = 0 (solid lines). Dots: empirical probabilities (frequencies of rain on each day of the year) in (A) and empirical autocorrelation coefficients (Pearson's correlation coefficient between occurrence on consecutive days on each day of the year) in (B); dashed lines: smoothed empirical values

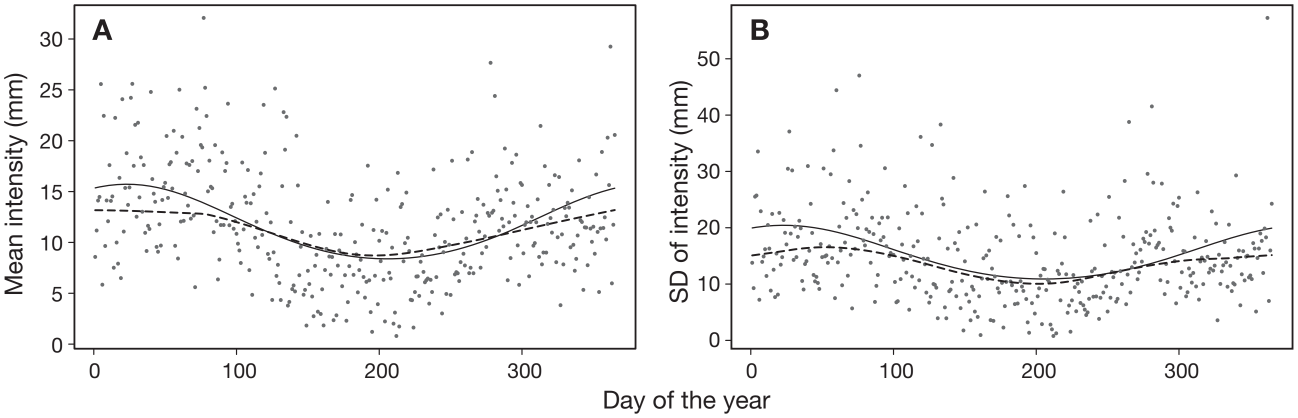

Fig. 4. Modeled (A) mean and (B) standard deviation (SD) of precipitation intensity (solid lines). Dots: empirical means and empirical SDs calculated separately for each day of the year from the 63 yr of data, conditioned on occurrence; dashed lines: smoothed empirical values

The assumption that the estimated shape parameter of the gamma distribution is constant is restrictive, since fitting separate models for precipitation intensity in summer and winter leads to estimates of the shape parameter of about 0.63 in summer and 0.55 in winter. On the other hand, it is far from trivial to relax this assumption within the proposed framework, which provides a flexible and more parsimonious approach than fitting separate models for summer and winter.

3.4. Minimum and maximum temperature

The idea of the modeling approach for Tmin and Tmax are 2 separate AR(1) models which are coupled through a dependence term corresponding to the other temperature variable. Therefore, our basic models will contain these 2 terms as well as a mean signal. On top of that we add the occurrence of precipitation on the same day J_t, a seasonal cycle given by C_t and S_t as well as the ENSO index E_t to the models

$$X_t = \mu_{X,0} + \mu_{X,1}J_t + \phi_X X_{t-1} + \psi_X Y_{t-1} + \beta_{X,1}C_t + \beta_{X,2}S_t + \beta_{X,3}E_t + \varepsilon_{X,t} \tag{13}$$

$$Y_t = \mu_{Y,0} + \mu_{Y,1}J_t + \phi_Y Y_{t-1} + \psi_Y X_t + \beta_{Y,1}C_t + \beta_{Y,2}S_t + \beta_{Y,3}E_t + \varepsilon_{Y,t} \tag{14}$$

The BIC criterion selects all of the considered covariates for both Tmin and Tmax, values of BIC along with estimated coefficients of the selected model are given in Table 2. Comparing the decrease in BIC when introducing precipitation occurrence of the same day and seasonal cycle indicates that for Tmin precipitation is the stronger signal than the seasonal cycle, whereas for Tmax it is the contrary. For both temperature variables the ENSO index is, although leading to a decrease in BIC, of lesser importance.

The coefficients concerning precipitation indicate that Tmin is higher by about 2°C when precipitation occurs whereas Tmax is lower by about the same amount.

The positive value of the coefficient corresponding to the ENSO index for Tmin means that for positive values of the ENSO index (corresponding to El Niño events) Tmin is increased and for negative values of the index (La Niña events) Tmin is decreased. The value of this increase/decrease is rather small, 0.74/–0.35 for the maximum and minimum in the observed ENSO index series.

The negative value of the coefficient corresponding to the ENSO index for Tmax means that for positive values of the ENSO index (corresponding to El Niño events) Tmax is decreased and for negative values of the index (La Niña events) Tmax is increased. The value of this decrease/increase is rather small, –0.55/0.26 for the maximum and minimum in the observed ENSO index series.

One of the indirect effects of the ENSO index on temperature is that precipitation is more frequent for positive values of the ENSO index, such that the effects of the (direct) ENSO predictor and the precipitation predictor combine.

Similar to the transition probabilities for precipitation occurrence, we predict the conditional means of Tmin and Tmax for the entire year and all observed values of the ENSO index. First, we calculate the unconditional probability of rain $\pi(t) = \Pr\{J_t = 1\}$ for a fixed value of the ENSO index within the observed range using the same technique as in Section 3.2 for ENSO = 0. Then we predict the mean of both temperature variables for the entire year for this fixed value of ENSO by taking the mean in the model equations and running the resulting recursive equations for a few years using the corresponding $\pi(t)$, i.e. the mean of J_t,

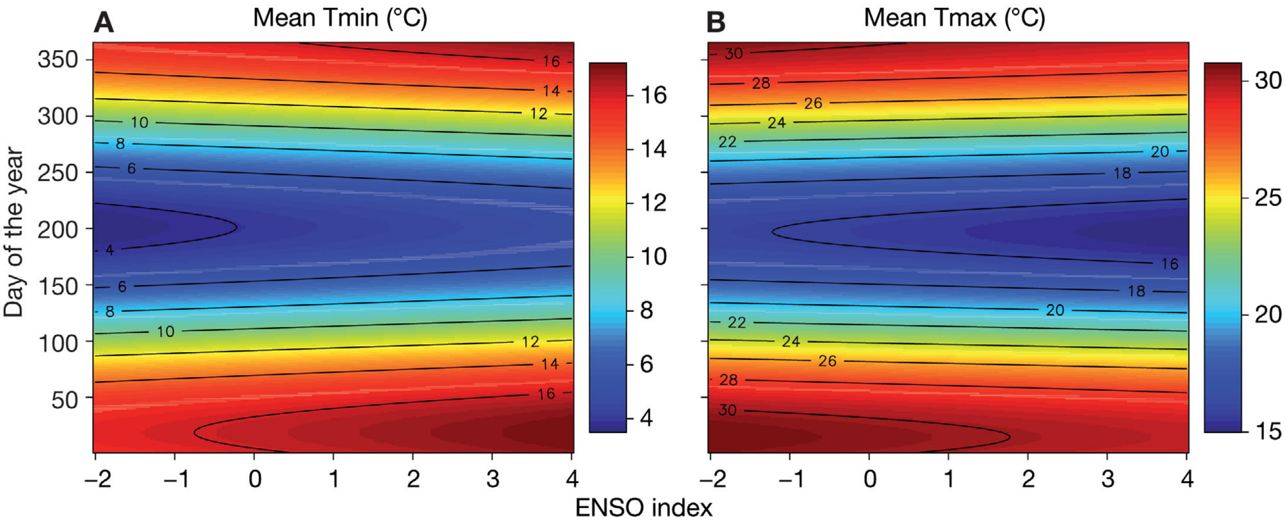

Fig. 5. Modeled mean of (A) Tmin and (B) Tmax as functions of the day of year and of the range of ENSO index values

together with the observed temperatures on Day 1 as starting values. Repeating these calculations for a fine enough grid over the observed ENSO range results in Fig. 5.

We observe that mean Tmin is higher for positive values of the ENSO index and lower for negative values. In contrast, mean Tmax is higher for negative values of the ENSO index and lower for positive values. In other words, the temperature range is smaller for positive values of the ENSO index, i.e. for El Niño events, and larger for negative values of the ENSO index, i.e. La Niña events. This can also be deduced be merely looking at the parameter estimates corresponding to the ENSO index covariate. Letson et al. (2005) report similar influence of the ENSO signal on temperature variables in the Pampas region.

In order to compare the modeled conditional means of the temperature variables with the observed data, we look at the 'neutral' situation of no effect of the ENSO index. That is, we show the curve for ENSO = 0 from Fig. 5 together with empirical mean temperatures calculated separately for each day of the year from the 63 yr of data and a smoothed version thereof in Fig. 6.

The empirical means and their smoothed versions in Fig. 6 show that also for the temperature variables the principal shape of our model agrees very well with the data, although it seems that the minimum over the year for the modeled Tmax is shifted with respect to the data.

4. VALIDATION

This section provides validation of our stochastic weather model through the generation of a large number of weather series of the same length as the data series, for which we calculate various types of statistics

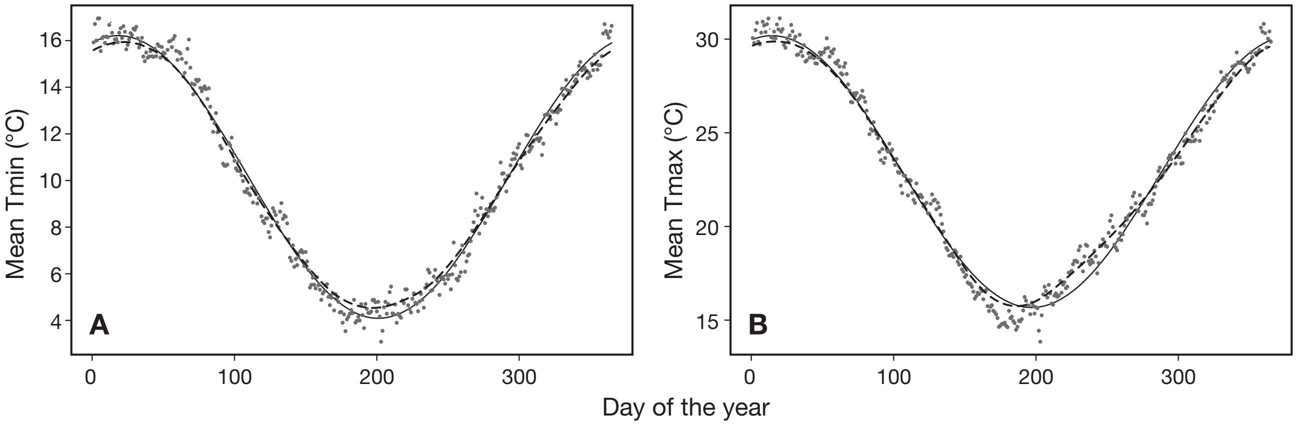

Fig. 6. Modeled mean of (A) Tmin and (B) Tmax daily temperature for ENSO = 0 (solid lines). Dots: empirical mean temperatures calculated separately for each day of the year from the 63 yr of data; dashed lines: smoothed empirical values

and compare those to the respective statistics of the data series. Details on how to generate weather series from the proposed model are given in Appendix 2. It should be kept in mind that the estimation technique we use (i.e. maximum likelihood) is not designed to reproduce sample statistics, nor should it be the goal to precisely model apparent sample behavior that is not necessarily real.

4.1. Daily statistics

We first check statistics derived from the series of daily data. Starting with precipitation, simple statistics are the average number of rainy days, the first-order autocorrelation coefficient of precipitation occurrence and the mean and SD of precipitation intensity on days rain occurred, each of these calculated separately for each month of the year from the 63 yr of data. The correspondence of statistics of 500 simulated series of length 63 yr to the data values has been summarized in Fig. 7.

For the average number of rainy days per month, we first note that the seasonal cycle in the data values is not very regular. Under these difficult circumstances it seems that the simulated values reproduce the data values reasonably well. The first-order autocorrelation coefficient of precipitation occurrence per month is reproduced very well, the simulated values are for most of the months not very far off from the data values and the apparent seasonal cycle in the data is very well reproduced.

The mean precipitation intensity per month is reproduced very well, except for the months of March and October, already indicating that high precipitation intensities are problematic for our model. Note also the extreme average numbers of wet days for these months. For the SD of precipitation intensity per month the reproduction of the data values by the generated statistics is a little less good than for the mean, especially for the months of March, May, September and October.

A more complex issue for precipitation are spells of consecutive dry or wet days, dry spells being more difficult to model using a Markov chain; see, for example, Wilks & Wilby (1999). For the 500 series of generated weather we calculate the number of spells of consecutive dry/wet days of any occurring length. For each spell length 5%, 50% and 95% quantiles of the sample of 500 spell counts are then obtained (for longer spells the sample might not be 500 but smaller depending on how many of the simulated series contained spells of the specific length). For wet spells, the corresponding count from the data series lies within the 5% and 95% quantiles of the simulated values for all spell lengths. For dry spells the picture is more complicated, see Fig. 8. First, in the data as well as in the simulated series longer spell lengths than for wet spells occur.

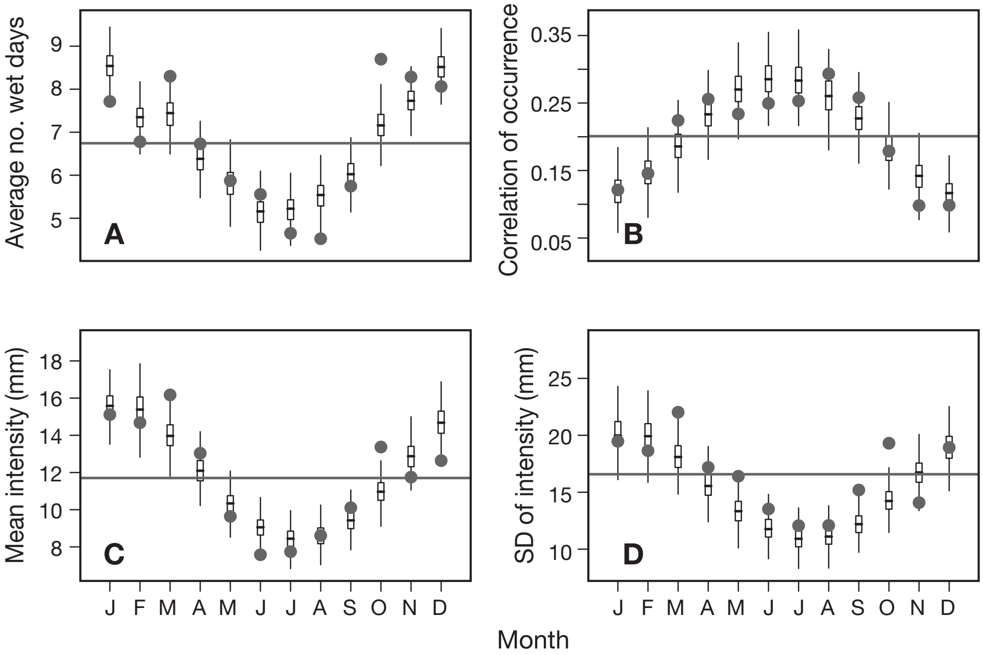

Fig. 7. (A) Average number of wet days, (B) first-order autocorrelation of occurrence, (C) mean precipitation intensity and (D) standard deviation (SD) of precipitation intensity, calculated separately for each month of the year from the 63 yr of data. Boxplots: results from 500 generated series; box height = interquartile range (lower to upper quartile); horizontal line within box = median; vertical lines extend to the most extreme data points (no more than 1.5× the interquartile range from the box). Gray dots: data values of the statistics, horizontal gray lines: mean value of each statistic over the year

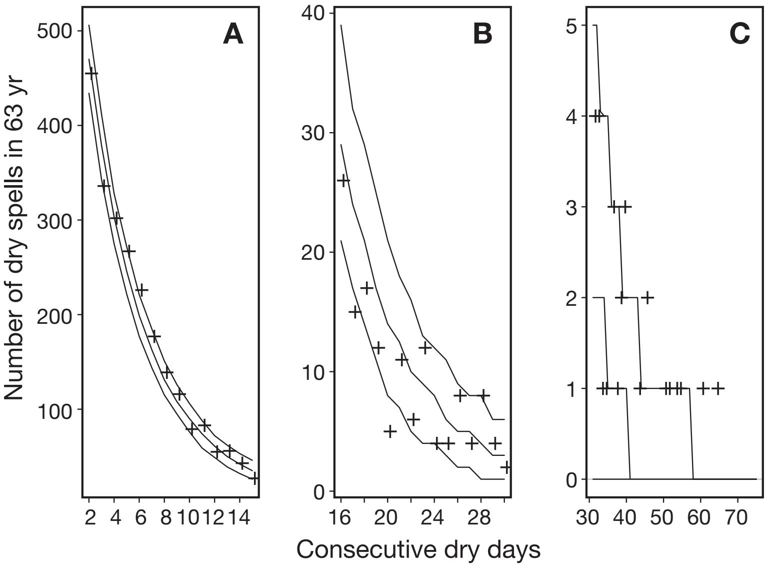

Fig. 8. Dry spells of different lengths: (A) 0 to 15 d, (B) 16 to 30 d and (C) 31 to 75 d. Lines in graphs: 5, 50 and 95% quantiles of the dry spell counts of 500 generated series; (+) corresponding counts of the data series. Note the change of scales in the 3 panels and that the 5% quantiles are always zero in (C)

The longer the spell lengths the less the model is able to reproduce them, this is the most obvious in the right-most panel of Fig. 8. Overall, the performance of our model is quite good, considering the fact that we did not use higher-order Markov chains, see, e.g. Wilks & Wilby (1999).

For Tmin and Tmax, we consider mean and SD as well as auto- and cross correlations calculated separately for each month of the year from the 63 yr of data. The correspondence of statistics of 500 simulated series of length 63 yr to the data values has been summarized in Fig. 9.

The means per month of both temperature variables are reproduced very well by the generated series. The SDs of Tmin and Tmax per month of the data series both show a rather pronounced and fairly regular seasonal cycle (only June and July for Tmax depart seriously). As expected, the generated SDs do not match this seasonal cycle

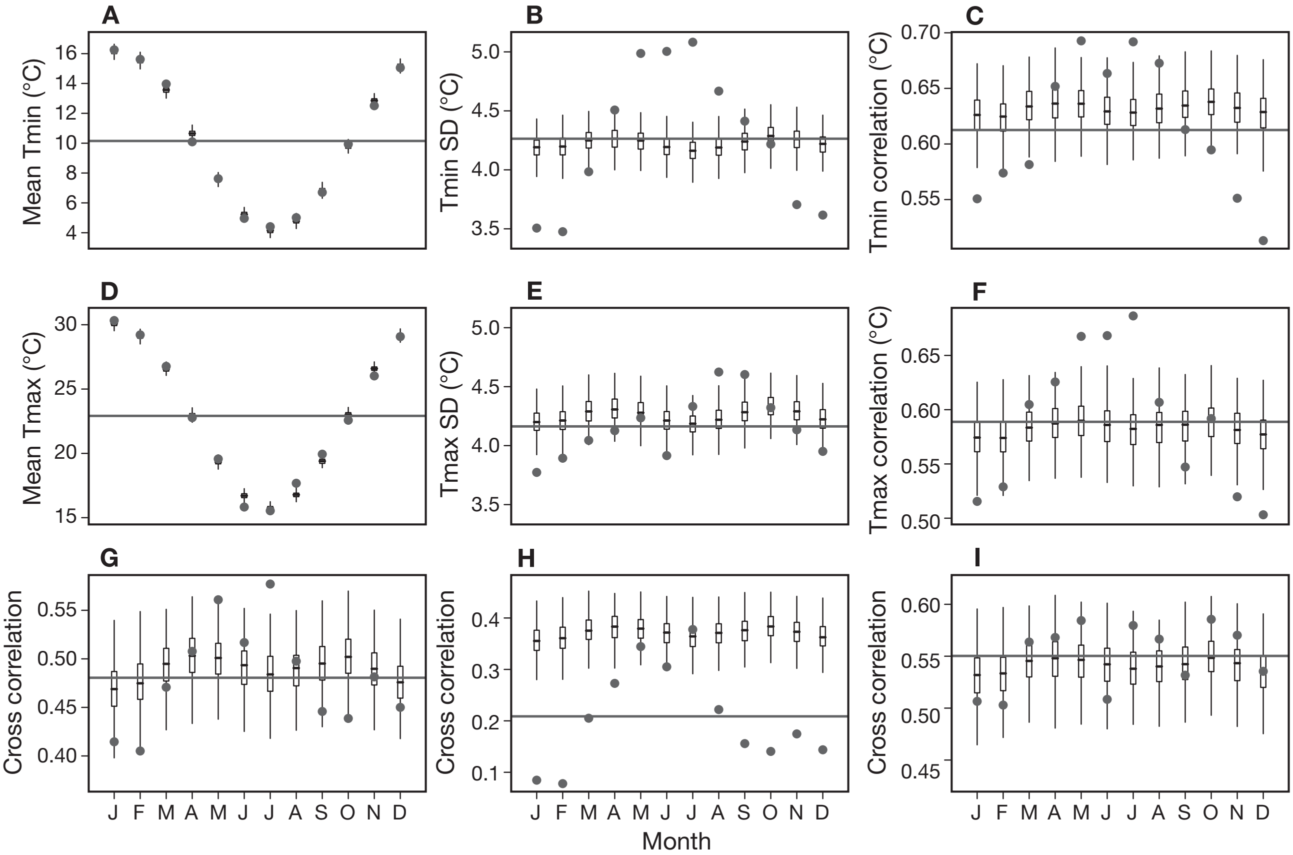

Fig. 9. (A,D) Mean, (B,E) standard deviation and (C,F) first-order autocorrelation of Tmin and Tmax as well as (G) zero-order and (H,I) first-order cross correlations, calculated separately for each month of the year from the 63 yr of data. For (H) first-order correlations, Tmin is at *t*−1 and Tmax is at *t*; for (I) first-order cross correlations, Tmin is at *t*, and Tmax is at *t*–1. Boxplots: results from 500 generated series (see Fig. 7 for box and whisker definitions); gray dots: data values of the statistics; horizontal gray lines: mean value of each statistic over the year

since this feature has only been explicitly modeled for the conditional means. In terms of nominal value, the generated SD are close to the mean of the monthly SDs over the year. The first-order autocorrelation of Tmin and of Tmax per month of the data series show again a rather pronounced and more or less regular seasonal cycle. And again, as expected, the generated autocorrelations do not match this seasonal cycle but are not too far from the mean of the monthly autocorrelations over the year.

The cross correlation between Tmin and Tmax per month of the data series shows a less regular seasonal cycle and for several months of the year the generated cross correlations match this statistic better than for the previous statistics. Again, they are not too far from the mean of the monthly cross correlations over the year. The first-order cross correlation between Tmin on Day t – 1 and Tmax on Day t per month is the statistic for which the generated statistics are the most off the data values. Note that the temporal distance between Tmax on Day t and Tmin on Day t – 1 is usually much more than 24 h, which was our motivation not to introduce this type of dependency in our model. For this reason it is not surprising that our model performs worse for this statistic than for others. On the other hand, the first-order cross correlation between Tmin on Day t and Tmax on Day t – 1 is explicitly modeled and is reproduced rather well.

The above deficiencies of the temperature models, which are in our opinion only minor problems in view of the simplicity and parsimony of the model, certainly point to the need of a more involved modeling of the variability of the temperature variables. See Section 5 for a discussion on ideas how to incorporate such a feature into our model.

Finally, we need to assess the number of instances when the generated Tmax is lower than the generated Tmin. The overall percentage of days for which the generated Tmax is lower than the generated Tmin is 0.2% for the total 31 500 yr of generated data. Therefore, we conclude that this situation occurs too rarely to be a reason to change our modeling strategy as dramatically as it would be necessary to overcome this problem. As a quick and simple solution we switch Tmin and Tmax values whenever the deficiency occurs.

4.2. Extremes

Obviously, our approach to weather generation is not specifically constructed to reproduce observed extremes in precipitation or temperature or even more so to generate extreme events of greater magnitude than observed. On the other hand, we believe that it is amenable to modifications aimed at a better treatment of extremes (see Section 6). In the following, we briefly examine how our proposed model performs with respect to extremes and point out the most important points that need improvement.

The simplest approach to study extremes from a statistical point of view is to model the maxima of a block of observations using the so-called generalized extreme value (GEV) distribution, which is, under weak conditions, the asymptotically correct model for maxima. As an example, we take maxima over the summer (October to March) of precipitation intensity and Tmax and minima of Tmin. One of the advantages of the GEV distribution function, denoted by G, is that it can be inverted analytically. For a given probability p, the quantile $z_p = G^{-1}(1 - p)$ is commonly called the return level associated with the return period $1/p$, since it is reasonable to expect z_p to be exceeded on average once every $1/p$ yr. Obviously, return levels can be calculated (or approximated) for other types of distributions. Return levels for, say, 20, 50 or more years are a quite intuitive way to describe the tail behavior of a distribution. See Coles (2001) for formulas concerning the GEV distribution and its return levels as well as an introduction to extreme value statistics.

To the samples of length 62 yr of summer maxima/minima (the first and the last year contain only an incomplete summer) we fit the GEV distribution and calculate the 50 yr return levels. Besides the data, the GEV has been fitted to summer maxima/minima of 500 generated weather series, the resulting 50 yr return levels are shown in Fig. 10 along with the corresponding data values. For precipitation, the discrepancy between data and simulated values is largest, the 50 yr return level estimated from the data being 238 mm, whereas even the 95% quantile of the 500 simulated return levels is only 189 mm. Clearly, the extremal behavior of the generated precipitation series is insufficient, not having a heavy enough upper tail, which is typical for a gamma model, see e.g. Wilks (1999). While the extremal behavior of the generated Tmin series is not perfect, the deficiency for this variable is much less serious than for precipitation. Since estimated temperature return levels from the simulated series are higher than the corresponding data value for Tmax, the generated temperature return levels are conservative estimates in contrast to those for precipitation.

4.3. Aggregative statistics

Stochastic models fitted to time series of daily precipitation are known to underestimate the inter-annual variance of aggregative statistics like monthly or seasonal total precipitation (Buishand 1978, Wilks 1989). Our model has not been constructed in view of over-

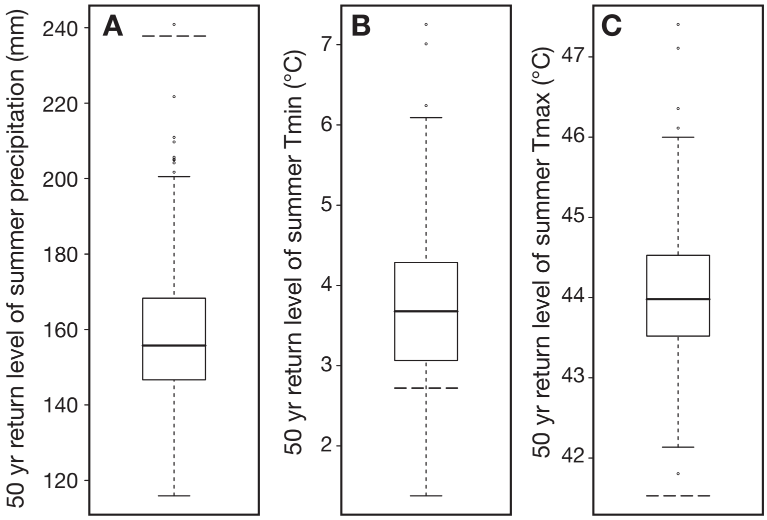

Fig. 10. 50 yr return levels of summer (A) precipitation maxima, (B) Tmin minima, and (C) Tmax maxima of 500 generated series based on the generalized extreme value (GEV) distribution. Horizontal dashed lines: corresponding values of the data series. Boxplots: see Fig. 7 for definitions, dots show outliers

coming this issue, termed 'overdispersion'; for possible approaches to do this at least partially, see Katz & Parlange (1998), Katz & Zheng (1999). This subsection summarizes how our model performs in reproducing inter-annual variances of annual as well as summer and winter total precipitation and also of average Tmax and Tmin; see e.g. Mavromatis & Hansen (2001), Qian et al. (2004) concerning overdispersion in temperature variables.

To this end, 500 weather series of length 63 yr have been generated. For each of them we calculate annual, summer and winter precipitation totals as well as annual, summer and winter averages of the temperature variables. Boxplots of the inter-annual SD of these aggregated statistics along with the corresponding values of the data series over the 63 yr are given in Fig. 11.

Comparing the median of the 500 simulated SD of precipitation totals to the corresponding value of the data series, we have 14% overdispersion for annual total precipitation, 13% for summer and 1% for winter, where overdispersion is defined as the relative error of the generated inter-annual SD compared to that of the data series. Note that the winter season is simply not as variable as the summer season, therefore our model is able to reproduce the variability in winter more easily. Also, the percentages we obtain with our model are of the same order of magnitude as that obtained by other models using first-order Markov chains, see e.g. Katz & Parlange (1998).

For Tmin, comparing the median of the 500 simulated SD to the corresponding value of the data series, we have 32% overdispersion for the annual average, 27% for the summer average and 19% for the winter average. These percentages are a little higher than those mentioned for monthly averages in Mavromatis & Hansen (2001) for a comparable model (without covariates). Our own simulations from a model without covariates show, that for annual and seasonal averages the percentages from this simpler model are somewhat higher.

For Tmax, comparing the median of the 500 simulated SD to the corresponding value of the data series, we have 4% overdispersion for the annual average, 7% for the summer average and 1% for the winter average. These results are much better than those reported in Mavromatis & Hansen (2001) for monthly averages, but obviously this could be due to special circumstances in Pergamino. Again, these percentages are somewhat lower than those obtained from our own simulations from a model without covariates.

In summary, while our model does not overcome the problem of overdispersion (which it has not been designed to do), in terms of percentages the results are comparable to other models using first-order Markov chain models for precipitation and AR(1) type models for the temperature variables. Note that the introduction of the ENSO index as a covariate does not correct

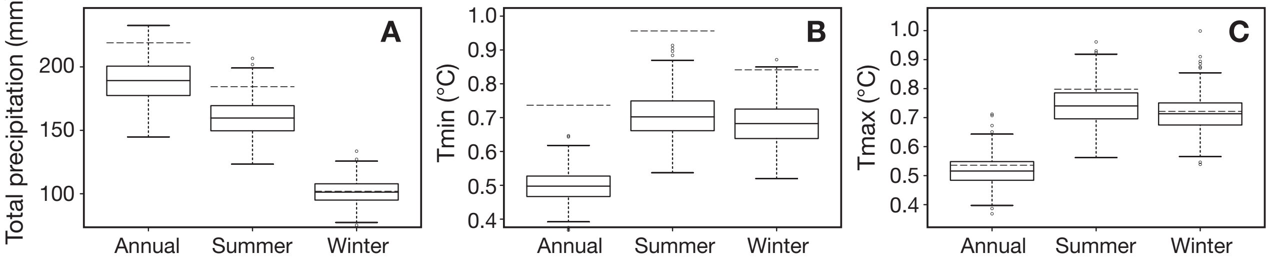

Fig. 11. Standard deviations of annual, summer and winter (A) total precipitation, average (B) Tmin and (C) Tmax of 500 generated series. Horizontal dashed lines: corresponding values of the data series; boxplots: see Figs. 7 & 10 for definitions

overdispersion, as one would expect in view of previous literature. It seems that, the ENSO signal on, e.g. precipitation at Pergamino is not strong enough to do so.

5. REFINEMENTS

The results of Section 4.1 clearly indicate the need to model error variability or other second moment statistics for the temperature variables depending on the season. Considering residuals of the fitted models for Tmin and Tmax, i.e. estimates of the error terms, is a direct way to identify potential weaknesses of the models we propose. We calculate SD of the residuals of Tmin and Tmax separately for each month of the year. The obtained residual SD shown in Fig. 12 support the already mentioned deficiency of our models.

In the case of temperature, explicit modeling of the variance is possible, see e.g. Chandler (2005), but beyond the scope of the present study. Simple estimates of different error variances for winter and summer, for example, can be obtained from the fit of any model, by dividing the residuals into winter and summer and calculating the sample variances of the 2 samples. Using the model chosen in Section 3.4, we obtain for Tmin a winter SD of 3.25°C and a summer SD of 2.72. For Tmax these are 2.96 and 3.12 respectively.

For an indicator variable I_t—which equals 1 if t is a day in (austral) winter, i.e. between April and September, and zero otherwise—a more involved approach is to introduce covariates of the type $I_{t-1}X_{t-1}$ for Tmin and $I_{t-1}Y_{t-1}$ for Tmax. An interaction between the seasonal cycle and the autoregressive terms X_{t-1} and Y_{t-1}, respectively, would refine this idea even further. Both alternatives allow the autocorrelation coefficients to be different depending on the season. This induces different variances of the temperature variables (not the error variances) as well, since in an AR(1) model the variance of the process is determined by error variance and autocorrelation coefficient. Another simple, but unfortunately not systematic nor parsimonious, way to allow for some heteroscedasticity of temperature is to fit separate seasonal models, e.g. for summer and winter. This simple approach allowed us to confirm that the model selection results of Section 3.4 are similar if summer/winter differences are considered.

Similarly, fitting separate models for precipitation intensity in summer and winter is a possibility to somewhat relax the assumption of a constant shape parameter of the gamma distribution. As mentioned before, fitting separate models leads to estimates of the shape parameter of about 0.63 in summer and 0.55 in winter. Again, it is far from trivial to relax this assumption within the proposed framework and more fundamental statistical research is required for this purpose.

Previous literature in atmospheric sciences indicates that, in general, the relationship of the ENSO index with weather variables appears to differ depending on the season. A possible strategy to include such behavior in the proposed weather generator could be to model different ENSO signals depending on winter and summer. Note that an interaction term between the ENSO index and the seasonal cycle has been considered but not selected by BIC for precipitation occurrence, Tmin and Tmax. Such an interaction term would allow a systematically changing effect of ENSO throughout the year.

We briefly explore the possibility of a seasonally changing ENSO signal for the different parts of the proposed weather generator; a thorough treatment is beyond the scope of the present study. For precipitation occurrence we add I_tE_t as a covariate into the model, where I_t is the summer/winter indicator variable from above. Let the coefficient of this covariate be δ, then the effect of the ENSO index on the log odds of precipitation is $(\beta_3+\delta)E_t$ in winter and β_3E_t in summer. For the gamma GLM for precipitation intensity, we have already found that the ENSO index does not improve the model fit. We can, nevertheless, consider

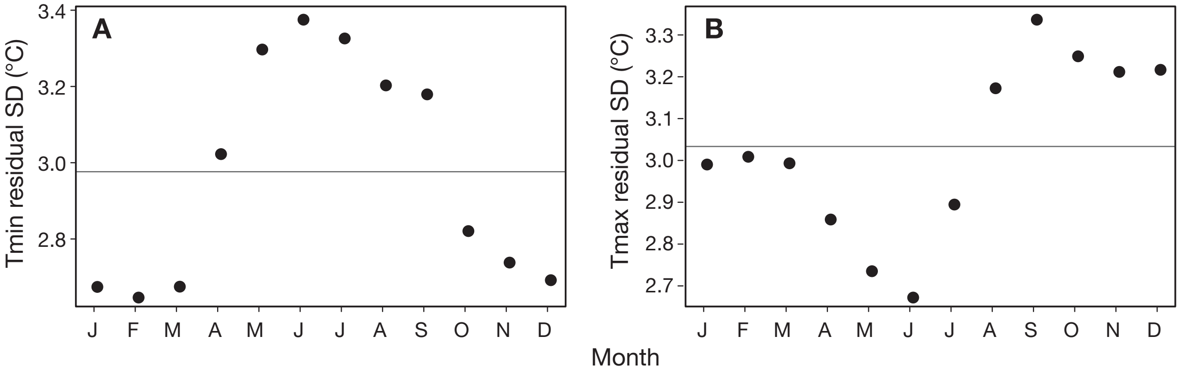

Fig. 12. Residual standard deviation (SD) for (A) Tmin and (B) Tmax calculated separately for each month of the year from the 63 yr of data. Horizontal gray lines: mean value of the SDs over the year

the additional covariate $I_t E_t$ in addition. For the temperature variables we use the same strategy and introduce $I_t E_t$ as a covariate, denoting the respective coefficients for Tmin and Tmax by δ_X and δ_Y, the effect of the ENSO index on Tmin is $(\beta_{X,3}+\delta_X)E_t$ in winter and $\beta_{X,3}E_t$ in summer, similarly for Tmax. For all parts of the proposed weather generator, the introduction of this type of term leads to an increase in BIC, we conclude therefore that we do not obtain a better model fit by considering seasonally different ENSO effects.

6. DISCUSSION

We have demonstrated that an approach based on generalized linear models can provide a general modeling framework for stochastic weather generators. In particular, this approach permits the inclusion of annual cycles as well as other covariates such as an index of the ENSO phenomenon. One advantage of this technique is that the determination of whether or not a given covariate ought to be included in the model is essentially as straightforward as in multiple regression analysis. A related advantage is the relative ease of performing uncertainty analysis (e.g. taking into account the uncertainty in the relationship between ENSO and daily weather), as needed in climate impact assessments. In the Argentina Pampas project (see 'Introduction'), the proposed approach provides a method by which such uncertainties will be taken into account in the input of time series of daily weather to dynamic, process-level crop simulation models (Letson et al. 2005).

One limitation of the GLM approach to stochastic weather generators is the difficulty in permitting the variability of daily weather variables to depend on covariates (e.g. allowing for annual cycles in temperature variability, not just the mean). Nevertheless, we have suggested ways in which such annual cycles can be included in the model. Another limitation concerns the difficulty in incorporating more than 2 other weather variables into the model in addition to precipitation. In principle, this limitation can be dealt with through an extension of the presented GLM technique by successively conditioning on the appropriate variables.

As indicated in our evaluation of the performance of the GLM-based weather generator, improved treatment of extremes is needed, consistent with the results of other evaluations of weather generators. The simulation of high precipitation amounts could be improved by, instead of using the gamma, substituting a heavier-tailed distribution for precipitation intensity. Ideally, such improvements can be achieved without completely abandoning the GLM framework.

Also indicated in the evaluation of our weather generator is the inability to produce enough variability in seasonal or annual aggregated weather statistics for both precipitation and temperature. This overdispersion phenomenon is a well-known weakness of stochastic weather generators more generally. In theory, this tendency to underestimate variances can be reduced through the inclusion of additional covariates in the model, not only large-scale indices like ENSO, but also regional indices of atmospheric or oceanic circulation.

For many applications in climate impact assessment, daily weather sequences at multiple sites are required. Although our work has focused on the case of only a single site, the extension of the GLM approach to multi-site stochastic weather generators appears feasible. In this regard, GLMs have already been applied to model a single weather variable at multiple sites. For example, Yang et al. (2005) have modeled multi-site daily precipitation using GLMs.

Acknowledgements. Research was partially supported by NSF Grant DMS-0355474 to the NCAR Geophysical Statistics Project and by NSF's Biocomplexity in the Environment program Grant BE-0410348. NCAR is sponsored by the National Science Foundation. We thank Guillermo Podestá and Balaji Rajagopalan for helpful advice on this research. We would also like to thank 3 anonymous referees for numerous suggestions that refined and improved the manuscript.

LITERATURE CITED

Apipattanavis S, Podestá GP, Rajagoplan B, Katz RW (2007) A semiparametric multivariate and multisite weather generator. Water Resour Res (in press)

Buishand TA (1978) Some remarks on the use of daily rainfall models. J Hydrol 36:295–308

Chandler RE (2005) On the use of generalized linear models for interpreting climate variability. Environmetrics 16:699–715

Chandler RE, Wheater HS (2002) Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland. Water Resour Res 38:1192, doi:10.1029/2001 WR000906

Coles S (2001) An introduction to statistical modeling of extreme values. Springer, London

Grondona MO, Podestá GP, Bidegain M, Marino M, Hordij H (2000) A stochastic precipitation generator conditioned on ENSO phase: a case study in southeastern South America. J Clim 13:2973–2986

Jolliffe IT, Hope PB (1996) Bounded bivariate distributions with nearly normal marginals. Am Stat 50:17–20

Kaplan A, Cane M, Kushnir Y, Clement A, Blumenthal M, Rajagopalan B (1998) Analyses of global sea surface temperature 1856–1991. J Geophys Res C 103:18567–18589

Katz RW (2002) Techniques for estimating uncertainty in climate change scenarios and impact studies. Clim Res 20:167–185

Katz RW, Parlange MB (1995) Generalizations of chain-dependent processes: application to hourly precipitation. Water Resour Res 31:1331–1341

Katz RW, Parlange MB (1996) Mixtures of stochastic processes: application to statistical downscaling. Clim Res 7:185–193

Katz RW, Parlange MB (1998) Overdispersion phenomenon in stochastic modeling of precipitation. J Clim 11:591–601

Katz RW, Zheng X (1999) Mixture model for overdispersion of precipitation. J Clim 12:2528–2537

Letson D, Podestá GP, Messina CD, Ferreyra R (2005) The uncertain value of perfect ENSO phase forecasts: stochastic agricultural prices and intra-phase climatic variations. Clim Change 69:163–196

Mavromatis T, Hansen JW (2001) Interannual variability characteristics and simulated crop response of four stochastic weather generators. Agric For Meteorol 109:283–296

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London

Parlange MB, Katz RW (2000) An extended version of the Richardson model for simulating daily weather variables. J Appl Meteorol 39:610–622

Podestá GP, Messina CD, Grondona MO, Magrin GO (1999) Associations between grain crop yields in central-eastern Argentina and El Niño-Southern Oscillation. J Appl Meteorol 38:1488–1498

Qian B, Gameda S, Hayhoe H, De Jong R, Bootsma A (2004) Comparison of LARS-WG and AAFC-WG stochastic weather generators for diverse Canadian climates. Clim Res 26: 175–191

R Development Core Team (2005) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. www.R-project.org

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

Richardson CW (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour Res 17: 182–190

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Stern RD, Coe R (1984) A model fitting analysis of daily rainfall data. J R Stat Soc Ser A 147:1–34

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York

Wilks DS (1989) Conditioning stochastic daily precipitation models on total monthly precipitation. Water Resour Res 25: 1429–1439

Wilks DS (1999) Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. Agric For Meteorol 93:153–169

Wilks DS, Wilby RL (1999) The weather generation game: a review of stochastic weather models. Prog Phys Geogr 23:329–357

Yan Z, Bate S, Chandler RE, Isham V, Wheater HS (2002) An analysis of daily maximum windspeed in northwestern Europe using generalized linear models. J Clim 15:2073–2088

Yang C, Chandler RE, Isham V, Wheater HS (2005) Spatial-temporal rainfall simulation using generalized linear models. Water Resour Res 41:W11415, doi:10.1029/2004WR003739

Appendix 1. Unconditional probability of rain and autocorrelation

For a fixed value of the ENSO index, the transition probabilities $p_{ij}(t)$ only depend on the day of the year corresponding to t, i.e. the same values are repeated after 365 d. Then the connection between $p_{ij}(t)$ and the unconditional probability $\pi(t) = \Pr(J_t = 1)$ of rain, as well as the first-order autocorrelation coefficient $\rho_1(t) = \mathrm{Corr}(J_{t-1}, J_t)$, is given by

$$\pi(t) = p_{01}(t) + \pi(t-1)[p_{11}(t) - p_{01}(t)], \qquad \text{(A1)}$$

$$\rho_1(t) = [p_{11}(t) - p_{01}(t)]\{\pi[t-1][1-\pi(t-1)]\}^{1/2}\{\pi[t][1-\pi(t)]\}^{-1/2}$$

for $t = 1,\ldots,365$ and with the convention that $\pi(0) = \pi(365)$ (see Katz & Parlange 1995)

Appendix 2. Implementation in R

All model fitting in this study was done with the free software environment for statistical computing and graphics R (2.1.1); see R Development Core Team (2005). We used the functions:

- 'glm' with argument 'family=binomial()' for occurrence and 'family=Gamma(link="log")' for intensity.
- 'gamma.shape' of the library 'MASS' for the maximum likelihood estimate of the shape parameter.
- 'lm' for Tmin and Tmax, which is mathematically, but not computationally equivalent to the use of 'glm'.

For the generation of weather series, we cycle through the following steps for each Day t, for which we desire to generate weather data. We use observed values of precipitation occurrence and minimum and maximum temperature as starting values for this procedure.

(1) Calculate p_t (using the inverse of the logistic link function, the estimated coefficients, the generated J_{t-1} and the other covariates at t) and generate the occurrence J_t according to p_t.

(2) Generate Tmin on Day t using the estimated coefficients, generated Tmin and Tmax on $t-1$, the generated J_t and the other covariates at t, and adding a normal noise with the estimated standard deviation for Tmin.

(3) Generate Tmax on Day t similarly to Tmin.

After having generated the entire series of precipitation occurrence and Tmax and Tmin, we:

(1) Switch the values of Tmin and Tmax where Tmin > Tmax.

(2) Generate precipitation intensity on the days for which precipitation occurrence has been generated. Using the estimated coefficients and the appropriate covariates, we calculate the mean μ_t, the scale parameter is given by $s_t = \mu_t/a$, where a is the estimated shape parameter. The function 'rgamma' is used to generate gamma variables with the prescribed mean and scale.

Note that the observed values of the ENSO index enter into these calculations, therefore we are generating weather series corresponding to the years for which these values were observed.

Editorial responsibility: Eric DeWeaver,
Madison, Wisconsin, USA

Submitted: January 23, 2007; Accepted: May 10, 2007
Proofs received from author(s): July 12, 2007