Vol. 53: 13–24, 2012
doi: 10.3354/cr01071

CLIMATE RESEARCH
Clim Res

Published online May 3

OPEN ACCESS

Reducing overdispersion in stochastic weather generators using a generalized linear modeling approach

Y. Kim[1,*], R. W. Katz[2], B. Rajagopalan[3], G. P. Podestá[4], E. M. Furrer[5]

[1]Department of Statistics, Yeungnam University, Daegu 712-749, South Korea
[2]Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, Colorado 80307, USA
[3]Department of Civil, Environmental and Architectural Engineering, and Cooperative Institute for Research in Environmental Science (CIRES), University of Colorado, Boulder, Colorado 80309, USA
[4]Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, Florida 33149, USA
[5]Institute for Social and Preventive Medicine, Biostatistics Unit, University of Zürich, 8001 Zürich, Switzerland

ABSTRACT: Stochastic weather generators are commonly used to simulate time series of daily weather, especially minimum (Tmin) and maximum (Tmax) temperature and amount of precipitation. Recently, generalized linear models (GLM) have been proposed as a convenient approach to fitting weather generators. One limitation of weather generators is a marked tendency to underestimate the observed interannual variance in monthly, seasonal, or annual total precipitation and mean temperature, termed the 'overdispersion' phenomenon. In this study, aggregated statistics, consisting of seasonal total precipitation and mean Tmin and Tmax, are introduced as additional covariates into the GLM weather generator. With an appropriate degree of smoothing of these aggregated statistics, this approach is shown to virtually eliminate overdispersion when applied to 2 sites, Pergamino and Pilar, in the Argentine Pampas. The addition of these covariates does not distort the performance of the weather generator in other respects, such as annual cycles in the probability of precipitation and in the mean Tmin and Tmax. For seasonal total precipitation, the reduction in overdispersion is partially attributable to a corresponding reduction in the overdispersion of the frequency of precipitation occurrence, as well as to apparent temporal trends or 'regime' shifts. For seasonal mean Tmin and Tmax, the reduction in overdispersion is largely due to temporal trends on an interannual time scale.

KEY WORDS: Stochastic weather generator · Generalized linear model · Overdispersion · Locally weighted scatterplot smoothing

1. INTRODUCTION

Stochastic weather generators are commonly used to simulate time series of weather, especially for the variables minimum and maximum temperature (Tmin and Tmax, respectively) and amount of precipitation, on a daily time scale (Wilks & Wilby 1999). Among other uses, these models constitute one technique to produce sequences of daily weather consistent with seasonal climate forecasts or longer-term climate change projections (Maraun et al. 2010, Wilks 2010). For example, in a project on agricultural decision-making in the Argentine Pampas, scenarios of daily weather are needed consistent with plausible variations in climate (Podestá et al. 2009).

Recently, generalized linear models (GLMs; McCullagh & Nelder 1989) have been proposed as a technique to fit stochastic weather generators to daily data (Furrer & Katz 2007). Through the use of covariates, the GLM approach makes it straightforward to incorporate annual cycles and long-term trends, as well as to condition the model on indices of large-

*Email: ykkim@yu.ac.kr

scale atmospheric or oceanic circulation such as the El Niño–Southern Oscillation phenomenon (ENSO). For more background on the use of GLMs in climate applications, see Chandler (2005), Chandler & Wheater (2002).

One important limitation of stochastic weather generators is their marked tendency to underestimate the observed interannual variance of monthly, seasonally, or annually aggregated variables (e.g. Buishand 1978, Katz & Parlange 1998), especially pronounced for precipitation. This behavior of the data relative to a given statistical model is conventionally termed the 'overdispersion' phenomenon (i.e. the model is 'underdispersed' relative to the data). The extent to which overdispersion is attributable to an inadequate model for high frequency (i.e. daily) variations in weather (Katz & Parlange 1998), as opposed to a failure to take into account low frequency (i.e. interannual) variations in climate such as 'regime' shifts (Katz & Zheng 1999), is not clear. Overdispersion implies that impact assessments involving the risk of climate variations on interannual time scales will be unrealistic if they rely on scenarios of daily weather from stochastic generators.

In the present study we propose a modified GLM-based weather generator that takes into account low frequency variations. To reduce the overdispersion phenomenon, we incorporate time series consisting of seasonal total precipitation and seasonal mean Tmin and Tmax into the GLM weather generator, as additional covariates. These seasonal time series need to be smoothed to avoid introducing underdispersion (i.e. too much variance instead of not enough variance). We use locally weighted scatterplot smoothing (LOESS; Cleveland 1979, Hastie & Tibshirani 1990) because of its simplicity and flexibility, although other common smoothers such as moving averages could have been used instead. Wilks (1989) conditioned a stochastic model for daily precipitation on monthly total precipitation, and Hansen & Mavromatis (2001) adjusted the parameters of a stochastic weather generator in an ad hoc fashion to correct for overdispersion. The ad hoc adjustments of Hansen & Mavromatis (2001) entail the risk that the performance of the weather generator may deteriorate in other respects. Less ad hoc approaches include conditioning or nesting the daily generator within another generator for a longer time scale such as monthly or annual (Dubrovsky et al. 2004, Srikanthan & Pegram 2009).

We briefly review the basic GLM approach to stochastic weather generators, and introduce the extension involving the use of aggregated climate statistics as covariates (Section 2). These extended models are then fitted to time series of daily weather at Pergamino and Pilar, 2 important agricultural locations in the Argentine Pampas, and the model fit in terms of overdispersion is evaluated (Section 3). The extent to which the addition of these covariates affects the performance of the GLM weather generator in other respects, such as annual cycles in the probability of precipitation and in mean Tmin and Tmax, is examined. Possible sources of the reduction of overdispersion are identified, whether corresponding to reductions in subcomponent processes, such as the frequency of wet days for precipitation, or to long-term temporal trends or apparent 'regime' shifts (Section 4). Finally, some implications of the results are discussed in Section 5.

2. GLM WEATHER GENERATOR

2.1. Original model

The GLM approach to stochastic weather generators introduced by Furrer & Katz (2007) focuses on the simplest form of generator first proposed by Richardson (1981). In the present study we only briefly describe this basic GLM weather generator, referring to Furrer & Katz (2007) for details (see also www.image.ucar.edu/~eva/GLMwgen/). For the ease of interpretation of the results concerning overdispersion, the ENSO phenomenon is not used as a covariate, unlike in Furrer & Katz (2007).

The precipitation occurrence and intensity components of the GLM stochastic weather generator of Furrer & Katz (2007) are essentially the same as those in Stern & Coe (1984), who used GLM to model daily precipitation amount as a chain-dependent process with annual cycles in the parameters.

2.1.1. Precipitation occurrence

Let J_t denote the precipitation occurrence state on day t of a given year (i.e. $J_t = 1$ if precipitation occurs, $J_t = 0$ otherwise), and let $p_t = \Pr\{J_t = 1\}, t = 1, 2, \ldots$, denote the probability of a wet day. Equivalent to a first-order, 2-state Markov chain, the logistic transformation of the probability of precipitation is modeled conditional on the occurrence state on the previous day J_{t-1}:

$$\ln(p_t/1 - p_t) = \mu + \alpha J_{t-1} + \beta_1 C_t + \beta_2 S_t + \gamma_1 C_t J_{t-1} + \gamma_2 S_t J_{t-1} \quad (1)$$

where $C_t = \cos(2\pi t/365)$ and $S_t = \sin(2\pi t/365)$. Besides the intercept term (or mean) μ, the coefficient α

permits the conditional probability of precipitation to shift depending on whether or not precipitation occurred on the previous day (strictly speaking, α is not a correlation coefficient because of the logistic transformation on the left-hand side of Eq. 1), β_1 and β_2 determine the phase and amplitude of the sine wave for the annual cycle in these conditional probabilities, and γ_1 and γ_2 allow this annual cycle to be separate for the 2 conditional probabilities.

2.1.2. Precipitation intensity

The daily precipitation intensity (i.e. precipitation amount conditional on $J_t = 1$) is modeled as a gamma distribution (e.g. Stern & Coe 1984), with an annual cycle in the form of a sine wave for mean intensity, denoted by μ_t:

$$\ln(\mu_t) = \mu + \beta_{\mu,1} C_t + \beta_{\mu,2} S_t \quad (2)$$

Besides the intercept term μ, the coefficients $\beta_{\mu,1}$ and $\beta_{\mu,2}$ determine the phase and amplitude of the sine wave for the annual cycle in the mean intensity. Eq. (2) is equivalent to allowing the scale parameter, but not the shape parameter, of the gamma distribution to have an annual cycle. This constraint on the shape parameter appears reasonable at both Pergamino (as already verified in Furrer & Katz 2007) and Pilar, but could be relaxed if necessary.

2.1.3. Tmin and Tmax

Let (X_t, Y_t) denote the Tmin and Tmax (respectively) on day t of a given year, jointly modeled as a bivariate first-order autoregressive AR(1) process (as in Richardson 1981). In the GLM approach of Furrer & Katz (2007), this bivariate process is modeled indirectly through 2 univariate linear models:

$$X_t = \mu_{X,0} + \mu_{X,1} J_t + \varphi_X X_{t-1} + \psi_X Y_{t-1} + \beta_{X,1} C_t + \beta_{X,2} S_t + \varepsilon_{X,t} \quad (3)$$

$$Y_t = \mu_{Y,0} + \mu_{Y,1} J_t + \varphi_Y Y_{t-1} + \psi_Y X_t + \beta_{Y,1} C_t + \beta_{Y,2} S_t + \varepsilon_{Y,t} \quad (4)$$

Here the 2 error terms, $\varepsilon_{X,t}$ and $\varepsilon_{Y,t}$, besides being normally distributed with mean = 0, have no autocorrelation or cross correlation, unlike the conventional representation of a bivariate AR(1) process in which the error terms need to be cross correlated (i.e. a general bivariate white noise process). The term involving J_t (i.e. coefficients $\mu_{X,1}$ in Eq. 3 and $\mu_{Y,1}$ in Eq. 4) allows for a shift in the conditional mean Tmin and Tmax depending on whether or not precipitation occurs (as in Richardson 1981), and the terms involving C_t and S_t (i.e. coefficients $\beta_{X,1}$ and $\beta_{X,2}$ in Eq. 3; $\beta_{Y,1}$ and $\beta_{Y,2}$ in Eq. 4) model the annual cycle in mean Tmin and Tmax as sine waves. Autocorrelation is included through a lag term consisting of the same temperature variable on the previous day (i.e. coefficients φ_X in Eq. 3 and φ_Y in Eq. 4). Cross correlation is introduced into the Tmin X_t through a term involving the Tmax on the previous day, Y_{t-1} (i.e. coefficient ψ_X in Eq. 3), and into the Tmax Y_t through a term involving the Tmin on the same day, X_t (i.e. coefficient ψ_Y in Eq. 4). Note that it would be straightforward to include additional Fourier series terms in Eqs. (3 & 4), as well as in Eqs. (1 & 2), if needed.

2.2. Model with aggregated covariates

The basis of our statistical approach is to relate long-term (i.e. interannual) temporal scale predictor variables to short-term (i.e. daily) temporal scale predictands. For example, indices of large-scale atmospheric or oceanic circulation, such as the ENSO, can be used as covariates in the daily precipitation model. Instead, we incorporate time series of seasonal climate statistics, namely total precipitation and mean Tmin and Tmax in the GLM weather generator as covariates in the manner of disaggregation. Retaining ENSO as a covariate would make the interpretation of the model more difficult. However, our approach indirectly takes into account the effects of ENSO on daily weather statistics, because of the well-established ENSO signal in these aggregated climate statistics in the Argentine Pampas (Grondona et al. 2000, Letson et al. 2005).

As will be seen in Section 3, using the observed (i.e. unsmoothed) seasonal climate statistics as covariates may introduce excessive noise into the daily weather statistics and result in 'underdispersion' for the aggregated climate statistics. Thus, we consider smoothed seasonal climate statistics as covariates in the GLM weather generator, and adopt LOESS as a smoothing tool (Cleveland 1979). LOESS combines much of the simplicity of linear least squares regression with the flexibility of nonlinear regression and resistance to outliers, and is descriptively known as locally weighted polynomial regression. It is a computationally intensive method, requires fairly large, densely sampled data sets in order to produce good models, and does not produce a regression function explicitly represented by a mathematical formula. Nevertheless, it is a very simple and flexible proce-

dure (e.g. LOESS does not require the specification of a function to fit a model to all of the data, except for a smoothing parameter called the 'span' and another parameter specifying the degree of the local polynomial; e.g. the function 'loess' in the open source statistical programming language R). Here we use the value 2 for the degree of the local polynomial, and the value of the span parameter is selected to minimize overdispersion (i.e. through increasing the variance produced by the statistical model). A moving average (or running mean), more commonly used as a smoother in climate research, would be less flexible than LOESS.

Our approach involves introducing LOESS smoothed seasonally aggregated climate statistics into the basic GLM weather generator specified by Eqs. (1) to (4) as follows:

$$\ln(p_t/1-p_t) = \mu + \alpha J_{t-1} + \beta_1 C_t + \beta_2 S_t + \gamma_1 C_t J_{t-1} + \gamma_2 S_t J_{t-1} + \beta_s I_t P_t^s + \beta_w (1-I_t) P_t^w \quad (5)$$

$$\ln(\mu_t) = \mu + \beta_{\mu,1} C_t + \beta_{\mu,2} S_t + \beta_{\mu,s} I_t P_t^s + \beta_{\mu,w} (1-I_t) P_t^w \quad (6)$$

$$X_t = \mu_{X,0} + \mu_{X,1} J_t + \varphi_X X_{t-1} + \psi_X Y_{t-1} + \beta_{X,1} C_t + \beta_{X,2} S_t + \beta_{X,s} I_t N_t^s + \beta_{X,w} (1-I_t) N_t^w + \varepsilon_{X,t} \quad (7)$$

$$Y_t = \mu_{Y,0} + \mu_{Y,1} J_t + \varphi_Y Y_{t-1} + \psi_Y X_{t-1} + \beta_{Y,1} C_t + \beta_{Y,2} S_t + \beta_{Y,s} I_t M_t^s + \beta_{Y,w} (1-I_t) M_t^w + \varepsilon_{Y,t} \quad (8)$$

where I_t is a seasonal indicator (i.e. $I_t = 1$ in austral summer [October–March] and $I_t = 0$ in austral winter [April–September]), P_t^s and P_t^w are LOESS smoothed summer and winter seasonal total precipitation, and N_t^s and N_t^w (M_t^s and M_t^w) are LOESS smoothed summer and winter seasonal mean Tmin (Tmax). Note that the summer and winter time-series are smoothed separately, and that the smoothed climate statistics do not vary depending on the day t, but remain constant over a given season (the use of the subscript 't' is solely for convenience). The seasonal indicators in Eqs. (5) to (8) allow for different relationships with the aggregated covariates depending on the season. The value of the LOESS smoothing parameter minimizing overdispersion is determined through trial and error, ranging from the case of no smoothing (i.e. span = 0) to as smooth as possible (i.e. span = 1).

3. FIT OF GLM WEATHER GENERATOR TO DATA

3.1. Study area and data

As an application of the basic GLM approach, time series of daily precipitation (mm) and daily Tmin and Tmax (°C) at 2 locations in the Argentine Pampas, Pergamino and Pilar (Fig. 1a), are considered. Both locations have a marked wet season in the Southern Hemisphere summer, with Pilar being somewhat drier (Fig. 1b,c). The Pergamino data were already modeled by Furrer & Katz (2007), the only difference is that the present study omits an index of the ENSO phenomenon as a covariate. The annual precipitation cycle in this region has a clear maximum in late spring and summer and a marked winter minimum. Data are available for the time period 1932–2003, but several years were excluded from the analysis because they contain too many missing values (Pergamino: 1954–1956 and 1964–1966; Pilar: 1956–1960 and 1968), such that a total of 66 yr of data were analyzed at each location. There are further missing values in the rest of the record, more so for temperature than for precipitation (especially at Pilar), but they are relatively scarce and do not prevent the GLM framework from being used. Data corresponding to February 29 in leap years were removed for simplicity. Note that Furrer & Katz (2007) applied a more stringent criterion for excluding years with missing data, analyzing only 63 yr for Pergamino, in part to facilitate the use of ENSO as a covariate.

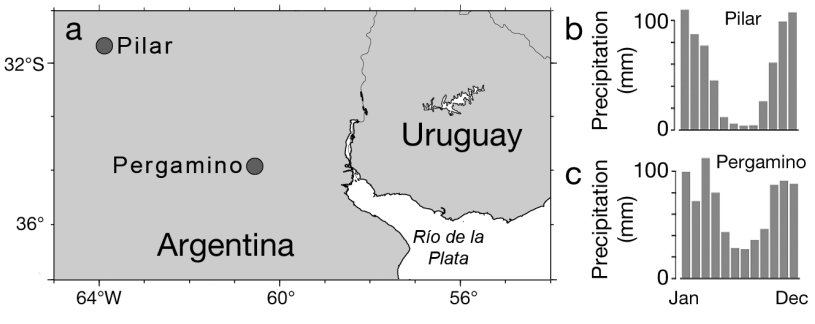

Fig. 1. (a) Pampas region of Argentina showing study locations Pergamino and Pilar. (b,c) Monthly mean total precipitation for (b) Pilar and (c) Pergamino

3.2. Fit of original model

Table 1 lists the parameter estimates and SEs for the original GLM weather generator (i.e. without the aggregated climate statistics as covariates). Using the open source software R (www.r-project.org), these results were obtained through the application of the 'glm' function to fit Eqs. (1) & (2) and 'lm,' a special case of

Table 1. Estimated coefficients (Coef.; estimate ± SE) for all components of the original GLM weather generator (i.e. without aggregated climate statistics as covariates) at Pergamino and Pilar, Argentine Pampas. Precip.: precipitation, Tmin: daily minimum temperature, Tmax: daily maximum temperature, AIC: Akaike's information criterion, BIC: Bayesian information criterion

Covariate category	Precip. occurrence Term	Coef.	Precip. intensity (mm) Term	Coef.	Tmin (°C) Term	Coef.	Tmax (°C) Term	Coef.
Pergamino								
Mean	μ	−1.56 ± 0.019	μ	2.43 ± 0.019	μ	−2.76 ± 0.109	μ	9.16 ± 0.112
Autocorrelation	J_{t-1}	1.10 ± 0.034	–	–	X_{t-1}	0.42 ± 0.005	Y_{t-1}	0.52 ± 0.006
Dependence	–	–	–	–	Y_{t-1}	0.36 ± 0.005	X_t	0.22 ± 0.006
	–	–	–	–	J_t	1.89 ± 0.048	J_t	−1.83 ± 0.050
Seasonality	C_t	0.45 ± 0.027	C_t	0.28 ± 0.028	C_t	0.72 ± 0.044	C_t	2.21 ± 0.044
	S_t	0.03 ± 0.027	S_t	0.14 ± 0.026	S_t	0.42 ± 0.030	S_t	0.44 ± 0.030
Interaction	$C_t J_{t-1}$	−0.57 ± 0.048	–	–	–	–	–	–
	$S_t J_{t-1}$	−0.01 ± 0.046	–	–	–	–	–	–
AIC		25842		38451		123202		123794
BIC		25891		38473		123249		123840
Pilar								
Mean	μ	−1.76 ± 0.021	μ	2.07 ± 0.022	μ	−1.23 ± 0.094	μ	9.11 ± 0.117
Autocorrelation	J_{t-1}	1.44 ± 0.037	–	–	X_{t-1}	0.50 ± 0.005	Y_{t-1}	0.57 ± 0.006
Dependence	–	–	–	–	Y_{t-1}	0.27 ± 0.004	X_t	0.17 ± 0.007
	–	–	–	–	J_t	1.27 ± 0.044	J_t	−2.41 ± 0.057
Seasonality	C_t	0.92 ± 0.030	C_t	0.54 ± 0.031	C_t	1.26 ± 0.040	C_t	2.06 ± 0.049
	S_t	0.13 ± 0.028	S_t	0.04 ± 0.028	S_t	0.62 ± 0.026	S_t	0.15 ± 0.033
Interaction	$C_t J_{t-1}$	−0.87 ± 0.052	–	–	–	–	–	–
	$S_t J_{t-1}$	−0.05 ± 0.048	–	–	–	–	–	–
AIC		24398		35125		119654		131020
BIC		24447		35149		119703		131069

glm, to fit Eqs. (3) & (4). For Pergamino, the estimated coefficients for the remaining variables are virtually the same as those obtained by Furrer & Katz (2007), despite the omission of ENSO and adjusting for the difference in the cosine and sine term definitions. To select the best fitting model, we use Akaike's information criterion (AIC) and Bayesian information criterion (BIC), with both criteria penalizing the maximized log likelihood function for the number of parameters estimated (e.g. Venables & Ripley 2002). The model with minimum AIC (or BIC) is selected as best fitting. Consistent with the results obtained by Furrer & Katz (2007), both the AIC and BIC indicate that each covariate category is statistically significant for Pergamino. Similarly, the AIC and BIC both support the same terms in the model for Pilar as for Pergamino (detailed results concerning model selection not included).

For Pergamino, Furrer & Katz (2007) already determined that this form of GLM weather generator underestimates the observed SD of annual and summer total precipitation (by roughly 15%), annual, summer, and winter mean Tmin (by roughly 20 to 30%), and to a lesser extent annual, summer, and winter mean Tmax, notwithstanding the inclusion by Furrer & Katz (2007) of an ENSO index as a covariate. In Section 3.3, we attempt to reduce this overdispersion.

3.3. Fit of model with aggregated covariates

Table 2 lists the estimated coefficients and associated SEs for all components (i.e. including the smoothed aggregated statistics as covariates) of the GLM weather generator fitted to Pergamino and Pilar. Comparing AIC and BIC values in these tables with the corresponding values in Table 1, the AIC always selects, and the BIC nearly always selects, the model with the aggregated covariates as being a better fit. The estimated coefficients of the remaining covariates do not change very much (especially those for the categories labeled 'autocorrelation' and 'dependence' in the tables) when the aggregated covariates are included.

Fig. 2 provides 2 examples of how the span parameter in LOESS can be chosen to minimize overdispersion. Time series of daily weather were simulated using the same 66 yr for which observations were

Table 2. Estimated coefficients (Coef; estimate ± SE) for all components of the GLM weather generator with aggregated climate statistics as covariates at Pergamino and Pilar, Argentine Pampas. Note: in precipitation models for convenience to make the results easy to present in a compact format, daily mean rate is used as a covariate instead of precipitation total. *Model preferred by AIC or BIC over corresponding model in Table 1. For definitions, see Table 1

Covariate category	Precip. occurrence Term	Coef.	Precip. intensity (mm) Term	Coef.	Tmin (°C) Term	Coef.	Tmax (°C) Term	Coef.
Pergamino								
Mean	μ	−1.72 ± 0.139	μ	1.61 ± 0.166	μ	−7.34 ± 0.307	μ	0.27 ± 1.664
Summer	$I_tP_t^s$	0.08 ± 0.038	$I_tP_t^s$	0.24 ± 0.045	$I_tN_t^s$	0.35 ± 0.021	$I_tM_t^s$	0.32 ± 0.061
Winter	$(1-I_t)P_t^w$	0.02 ± 0.083	$(1-I_t)P_t^w$	0.42 ± 0.099	$(1-I_t)N_t^w$	0.67 ± 0.045	$(1-I_t)N_t^w$	0.50 ± 0.090
Autocorrelation	J_{t-1}	1.10 ± 0.034	–	–	X_{t-1}	0.41 ± 0.005	Y_{t-1}	0.51 ± 0.006
Dependence	–	–	–	–	Y_{t-1}	0.37 ± 0.005	X_t	0.23 ± 0.006
	–	–	–	–	J_t	1.88 ± 0.047	J_t	−1.83 ± 0.050
Seasonality	C_t	0.29 ± 0.051	C_t	0.20 ± 0.058	C_t	0.36 ± 0.071	C_t	2.50 ± 0.073
	S_t	0.03 ± 0.027	S_t	0.14 ± 0.026	S_t	0.40 ± 0.029	S_t	0.44 ± 0.030
Interaction	$C_t J_{t-1}$	−0.58 ± 0.048	–	–	–	–	–	–
	$S_t J_{t-1}$	−0.01 ± 0.046	–	–	–	–	–	–
AIC	25830*		38415*		122922*		123743*	
BIC	25895		38456*		122987*		123809*	
Pilar								
Mean	μ	−2.39 ± 0.140	μ	1.65 ± 0.176	μ	−5.60 ± 0.230	μ	−2.11 ± 1.062
Summer	$I_tP_t^s$	0.21 ± 0.042	$I_tP_t^s$	0.12 ± 0.053	$I_tN_t^s$	0.30 ± 0.015	$I_tM_t^s$	0.40 ± 0.038
Winter	$(1-I_t)P_t^w$	0.65 ± 0.180	$(1-I_t)P_t^w$	0.53 ± 0.224	$(1-I_t)N_t^w$	0.61 ± 0.030	$(1-I_t)M_t^w$	0.57 ± 0.053
Autocorrelation	J_{t-1}	1.43 ± 0.037	–	–	X_{t-1}	0.48 ± 0.005	Y_{t-1}	0.56 ± 0.006
Dependence	–	–	–	–	Y_{t-1}	0.27 ± 0.004	X_t	0.18 ± 0.007
	–	–	–	–	J_t	1.28 ± 0.044	J_t	−2.40 ± 0.056
Seasonality	C_t	0.79 ± 0.053	C_t	0.55 ± 0.063	C_t	1.12 ± 0.063	C_t	2.31 ± 0.080
	S_t	0.12 ± 0.028	S_t	0.04 ± 0.028	S_t	0.63 ± 0.026	S_t	0.13 ± 0.033
Interaction	$C_t J_{t-1}$	−0.87 ± 0.052	–	–	–	–	–	–
	$S_t J_{t-1}$	−0.05 ± 0.048	–	–	–	–	–	–
AIC	24373*		35121*		119229*		130896*	
BIC	24438*		35162		119294*		130962*	

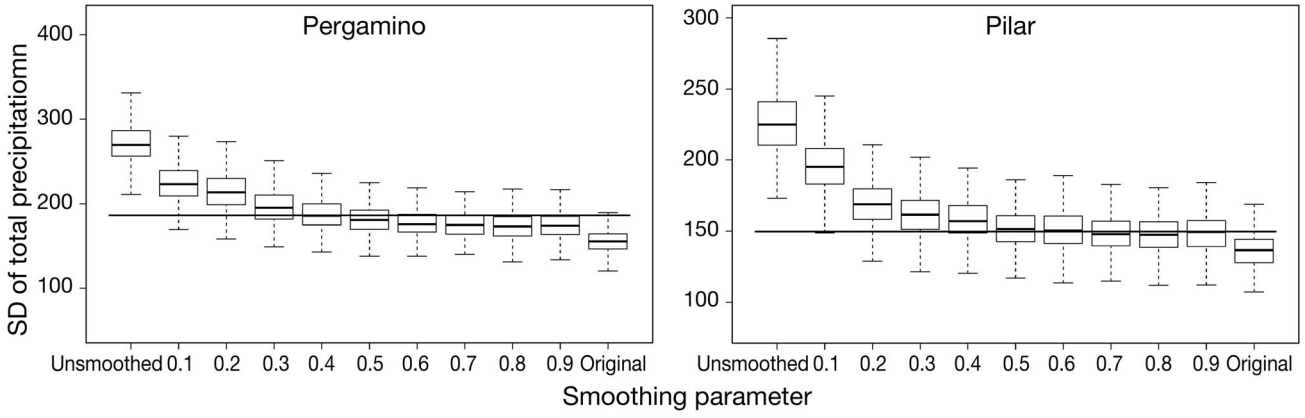

Fig. 2. Boxplots (giving the minimum, lower quartile, median, upper quartile, and maximum; box indicates middle half of data, dashed line the range) of simulated SD of summer total precipitation (mm) as a function of the LOESS span smoothing parameter for the GLM weather generator with aggregated climate statistics as covariates at Pergamino and Pilar. Horizontal solid line: corresponding observed value for the data series (below the line indicates overdispersion, above the line underdispersion)

available and aggregated statistics calculated, with the simulation exercise repeated 500 times. Shown are boxplots of the SD of the aggregated statistics, along with the corresponding values for the 66 yr data series and including cases of unsmoothed aggregated covariates and temperature models with linear temporal trend covariates but no aggregated covariates. For summer total precipitation at both Pergamino and Pilar, the overdispersion present in the original model (i.e. without any aggregated statistics as covariates) gradually disappears, eventually becoming underdispersed as the case of no smoothing (i.e. span = 0) is approached. For Pergamino, a span parameter of 0.4 virtually eliminates any overdispersion; for Pilar, a span of 0.6 is necessary. If a finer grid of values of the span parameter were used, then the overdispersion could be completely eliminated. Nevertheless, it is clear from Fig. 2 that the degree of overdispersion is not very sensitive to the choice of value of the span parameter.

Using the same simulation approach as in Fig. 2, Fig. 3 illustrates how our proposed model (i.e. with aggregated statistics as covariates), with a suitable choice of smoothing parameter, performs in reproducing variances of summer and winter total precipitation and mean Tmin and Tmax at both Pergamino

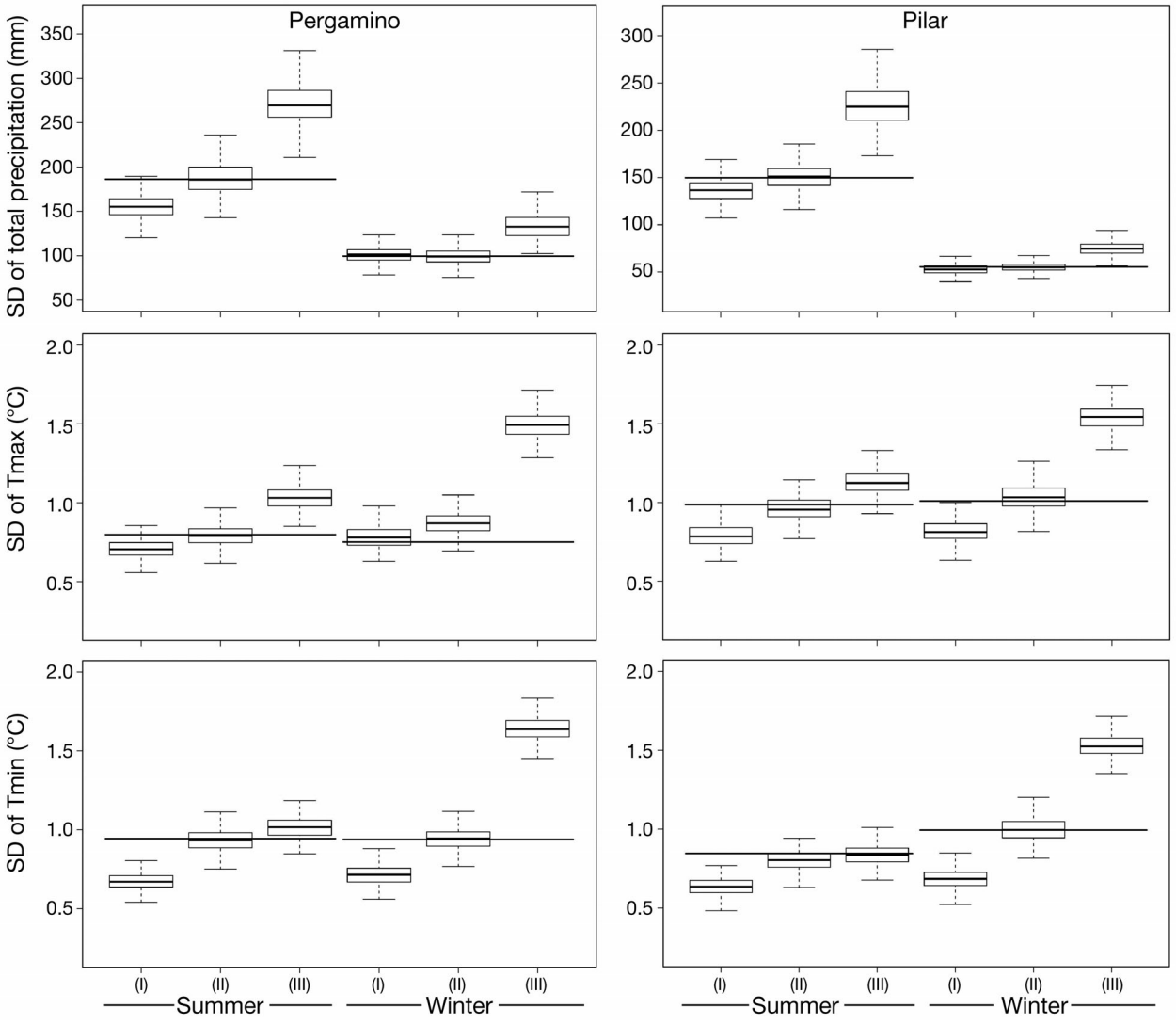

Fig. 3. Boxplots (see Fig. 2 for definitions) of simulated SD of summer (S) and winter (W) total precipitation (mm, top), mean maximum temperature (Tmax) (°C, middle) and mean minimum temperature (Tmin) (°C, bottom), for Pergamino (left) and Pilar (right), based on the GLM weather generator without aggregated covariates (I), with smoothed aggregated covariates (II; Pergamino span — mean precipitation [S] 0.4, [W] 0.4; mean Tmax [S] 0.3, [W] 1.0; mean Tmin [S] 0.2, [W] 0.7. Pilar span — mean precipitation [S] 0.6, [W] 0.4; mean Tmax [S] 0.5, [W] 0.2; mean Tmin [S] 0.0, [W] 0.9), and unsmoothed aggregated covariates (III). Horizontal solid line: corresponding observed value for the data series. For boxplot definitions, see Fig. 2

and Pilar. The proposed model virtually eliminates the overdispersion phenomenon in nearly all cases, with the value of the span parameter ranging from 0.4 to 0.6 for seasonal total precipitation, from 0 to 0.9 for Tmin, and from 0.2 to 1 for Tmax. The wider range in span parameter for temperature may be related to the presence of long-term trends, considered in Section 4. Although the winter mean Tmax at Pergamino does not appear to be overdispersed, this might reflect sampling error in estimating the seasonal SD. The GLM weather generator with unsmoothed aggregated covariates tends to overestimate inter-annual variances (i.e. underdispersion), and the introduction of a linear temporal trend (i.e. without the smoothed aggregated temperatures as covariates) in the temperature models is not enough to correct overdispersion. Note that precipitation in the winter season is simply not as variable as in the summer season, at least in absolute terms.

4. EVALUATION OF GLM WEATHER GENERATOR

4.1. Daily statistics

We examine how well the GLM weather generator, with and without the aggregated climate statistics as covariates, reproduces some daily statistics (a subset of those examined in Furrer & Katz 2007), focusing on the results for Pilar. The Markov chain model for daily precipitation occurrence can be fully characterized by the 2 transition probabilities $p_{11}(t) = \Pr\{J_t = 1 \mid J_{t-1} = 1\}$, the conditional probability of a wet day given the previous day was wet, and $p_{01}(t) = \Pr\{J_t = 1 \mid J_{t-1} = 0\}$, the conditional probability of a wet day given the previous day was dry. From these transition probabilities, it is straightforward to derive the unconditional probability of a wet day, $\pi(t) = \Pr\{J_t = 1\}$, and the first-order autocorrelation coefficient (or 'persistence' parameter), $\rho(t) = \mathrm{Corr}(J_{t-1}, J_t)$, of the occurrence process (see Furrer & Katz 2007).

As a function of the time of year, Figs. 4 to 7 show $p_{11}(t)$ and $p_{01}(t)$, $\pi(t)$ and $\rho(t)$, the mean and SD of daily precipitation intensity, and the mean daily Tmin and Tmax, respectively, at Pilar. In each case, the curves for the GLM weather generator, both with and without the aggregated climate statistics as covariates, are included along with the observed daily statistics. Like the mean daily Tmin and Tmax, the transition probability $p_{01}(t)$, the unconditional probability $\pi(t)$, and the mean and SD of intensity all have quite noticeable maxima in mid-summer. Only the persistence parameter exhibits a maximum in mid-winter. The GLM weather generator captures all of these seasonal patterns quite well, with virtually no difference depending on whether or not the aggregated climate statistics are included as covariates. The results obtained for Pergamino are quite similar (not shown, but included in Furrer & Katz 2007).

4.2. Sources of reduction in overdispersion

Long-term trends or more abrupt shifts in 'regimes' are one possible source of overdispersion. For precipitation, it can also be informative to decompose the variance of seasonal total precipitation into 2 components, one involving the variance of the number of wet days, the other the variance of daily precipitation intensity (Katz & Parlange 1998).

Using the same approach as in Section 3, Fig. 8 shows boxplots of the simulated SDs of the number of wet days in summer and winter at Pergamino and

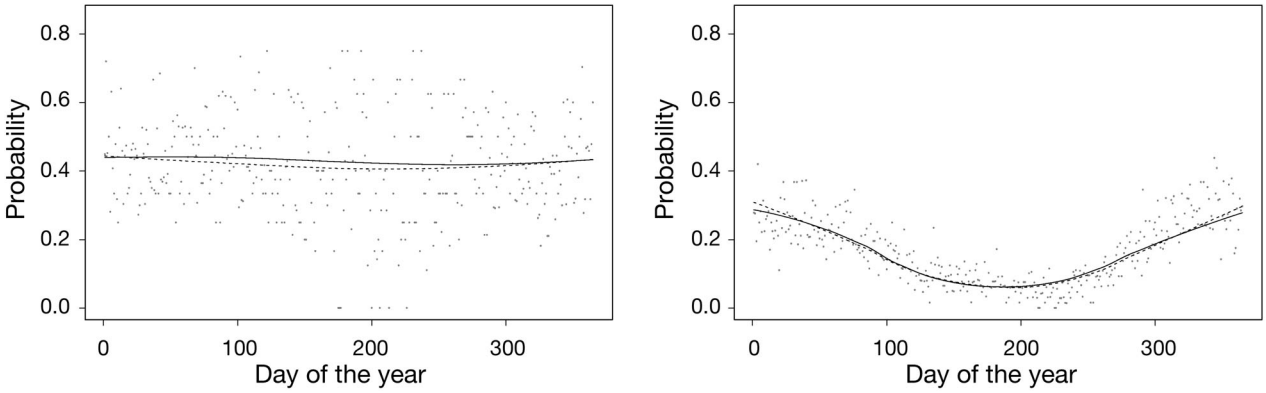

Fig. 4. Modeled transition probabilities $p_{11}(t)$ (left) and $p_{01}(t)$ (right) with (solid line) and without (dashed line) smoothed aggregated covariates, at Pilar. Dots: empirical transition probabilities, i.e. frequencies of observed transitions calculated separately on each day of the year

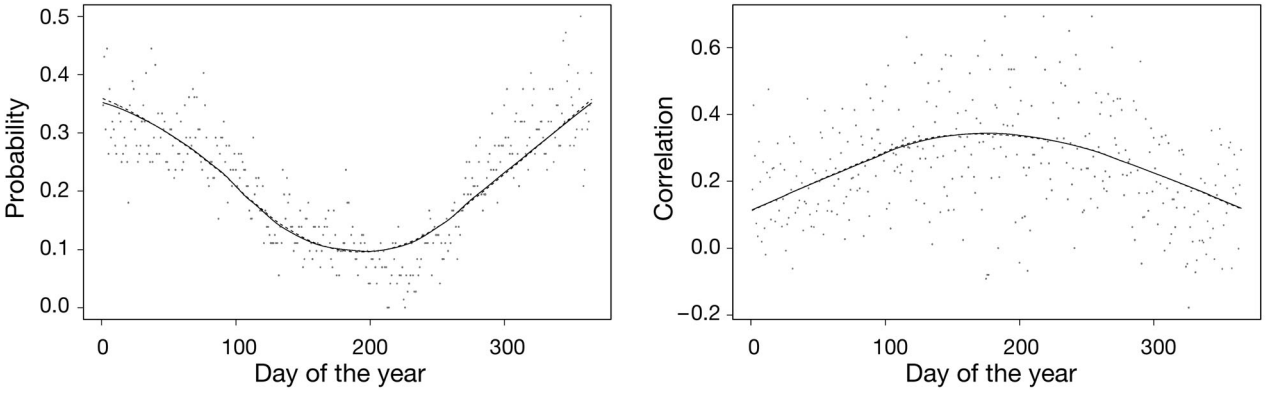

Fig. 5. Modeled unconditional probability of rain $\pi(t)$ (left) and first-order autocorrelation coefficient $\rho(t)$ (right) with (solid line) and without (dashed line) smoothed aggregated covariates, at Pilar. Dots: empirical probabilities, i.e. frequencies of rain on each day of the year (left) and empirical autocorrelation coefficients (Pearson's correlation coefficient between occurrence on consecutive days on each day of the year) (right)

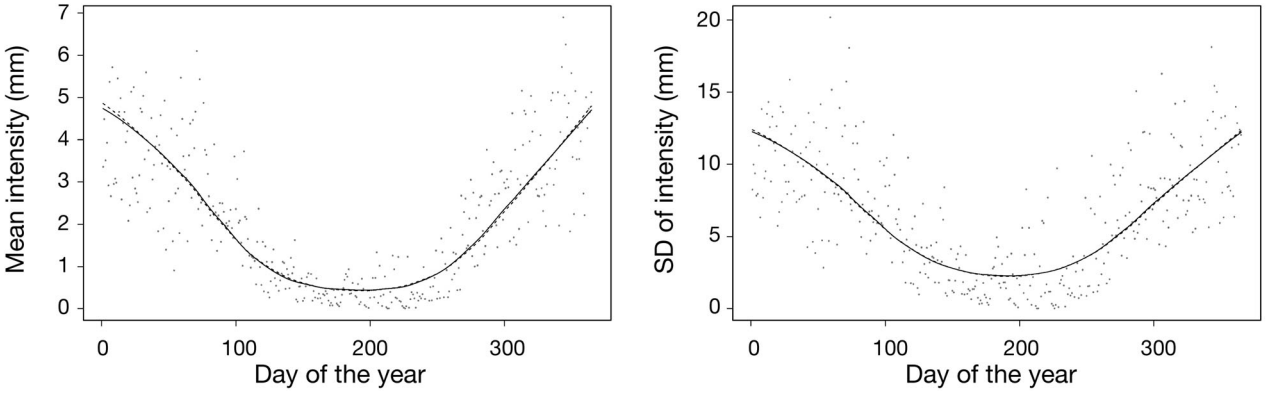

Fig. 6. Modeled (left) mean and SD (right) of precipitation intensity with (solid line) and without (dashed line) smoothed aggregated covariates, at Pilar. Dots: empirical means and SD calculated separately for each day of the year

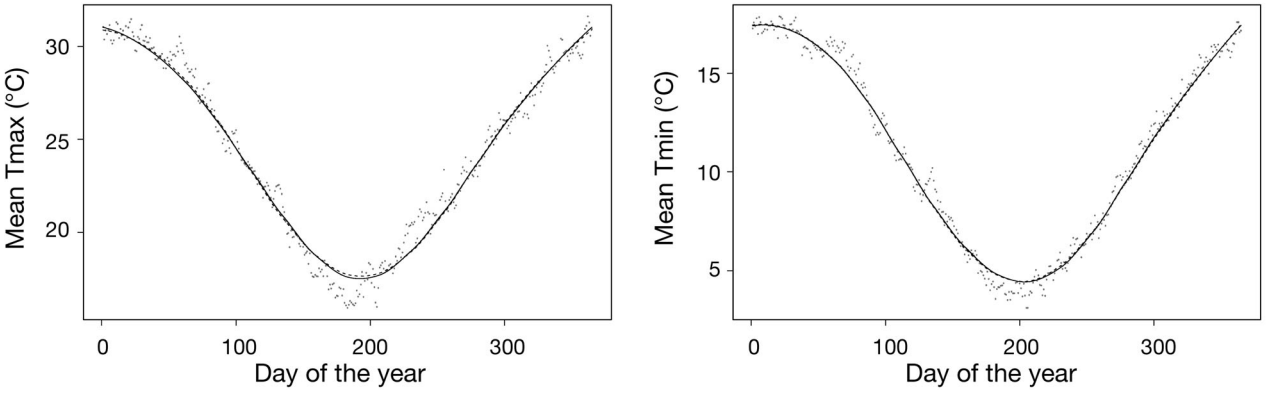

Fig. 7. Modeled mean daily maximum (Tmax; left) and minimum (Tmin; right) temperatures with (solid line) and without (dashed line) smoothed aggregated covariates, at Pilar. Dots: empirical mean temperatures calculated separately for each day of the year

Pilar. The overdispersion in this statistic in summer at Pilar is essentially removed with the modified GLM weather generator. On the other hand, because the seasonal total precipitation covariates in the precipitation occurrence component of the GLM weather generator are only barely statistically significant in summer at Pergamino (see SEs in Table 2), the overdispersion cannot be reduced much at all in this case. One way to ensure the elimination of overdispersion in the number of wet days would be to modify how the transition probabilities are modeled, replacing the seasonal total precipita-

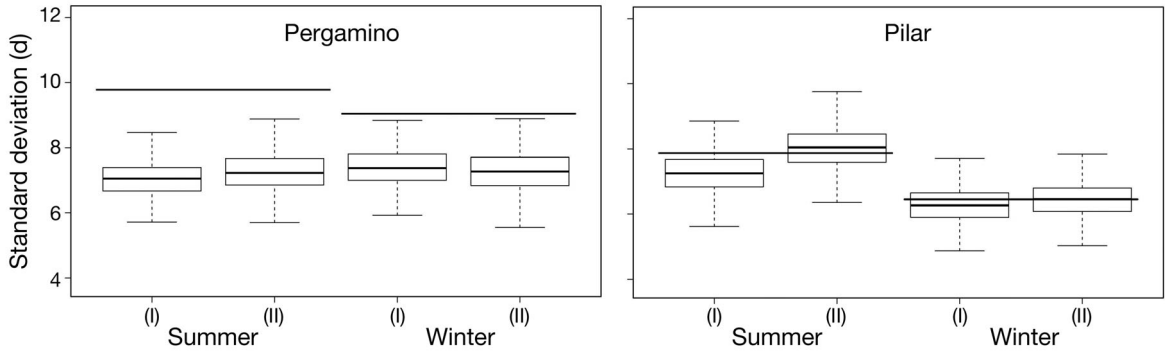

Fig. 8. Boxplots of simulated SD of summer and winter number of wet days for Pergamino and Pilar based on the GLM weather generator without aggregated covariates (I) and with smoothed aggregated covariates (II). Horizontal solid line: corresponding observed value for the data series. For boxplot definitions, see Fig. 2

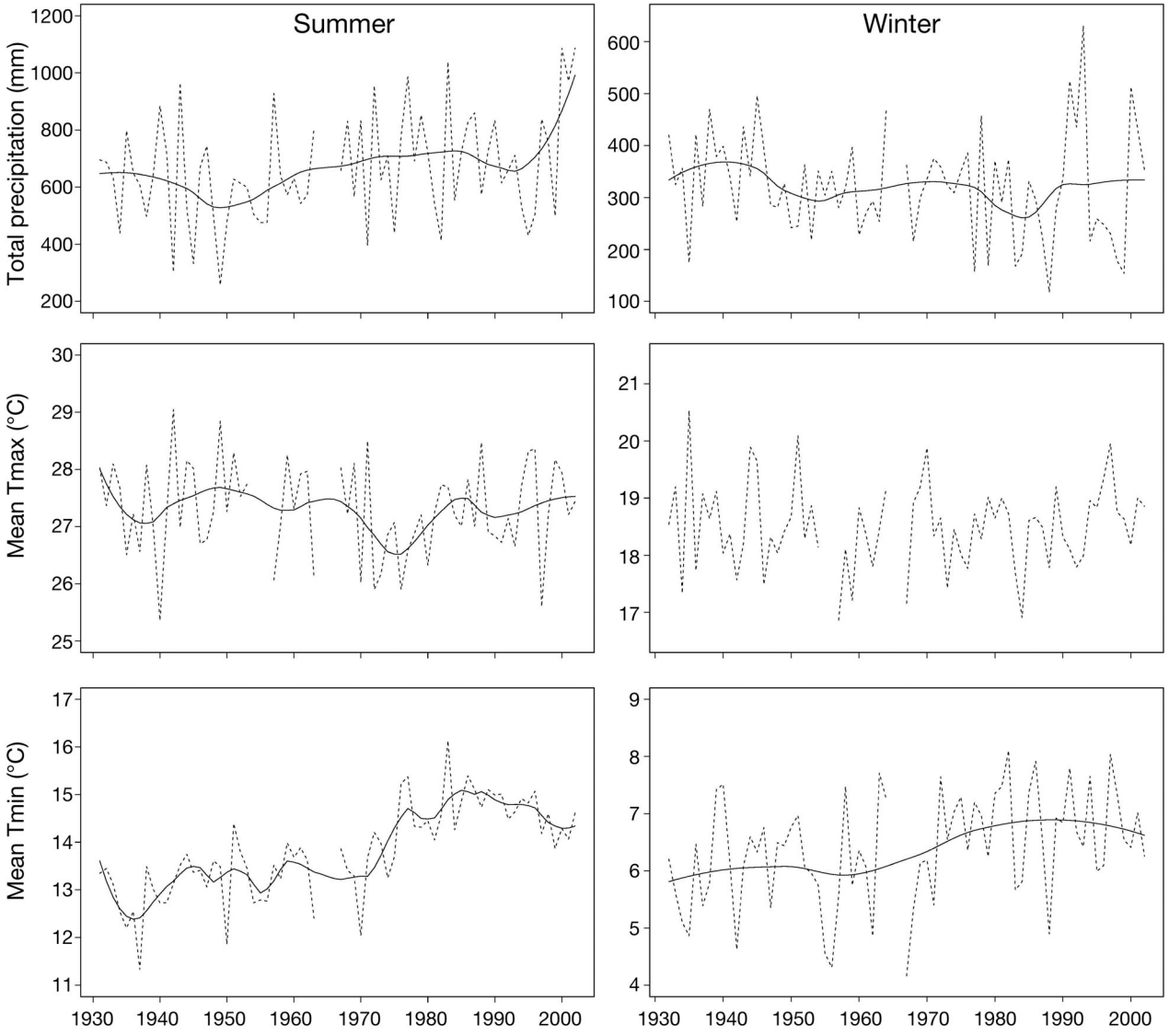

Fig. 9. Optimal smoothed aggregated covariates of total precipitation (top), mean maximum (Tmax; middle) and minimum (Tmin; bottom) temperatures during summer (left) and winter (right) for Pergamino. Dashed lines: corresponding observed values of the data series. Note that for winter Tmax, no smoothed covariate is used

tion covariates in Eq. (5) with the corresponding number of wet days.

Figs. 9 & 10 show the time series of the seasonal aggregated climate statistics, both raw and optimally smoothed, at Pergamino and Pilar. Marked trends of increasing Tmin are evident at Pergamino in summer and at Pilar in both summer and winter; such patterns were also identified by Messina et al. (1999), Magrin et al. (2005). Weaker trends of decreasing Tmax may be present, especially at Pilar in summer. The fact that the use of linear trends, instead of aggregated temperature statistics, as covariates did not eliminate overdispersion (as mentioned in Section 3.3) suggests that these trends may be somewhat nonlinear. For seasonal precipitation, while no marked trends are evident, there are at least hints of a shift in recent decades to a wetter regime in summer at both Pergamino and Pilar (Podestá et al. 2009). Any such shifts would be automatically incorporated into the model through the seasonally aggregated covariates.

5. DISCUSSION

It is shown how the GLM approach to stochastic weather generators can been extended to effectively eliminate the overdispersion phenomenon in seasonally aggregated climate statistics. Consequently, scenarios of daily weather can be produced with more

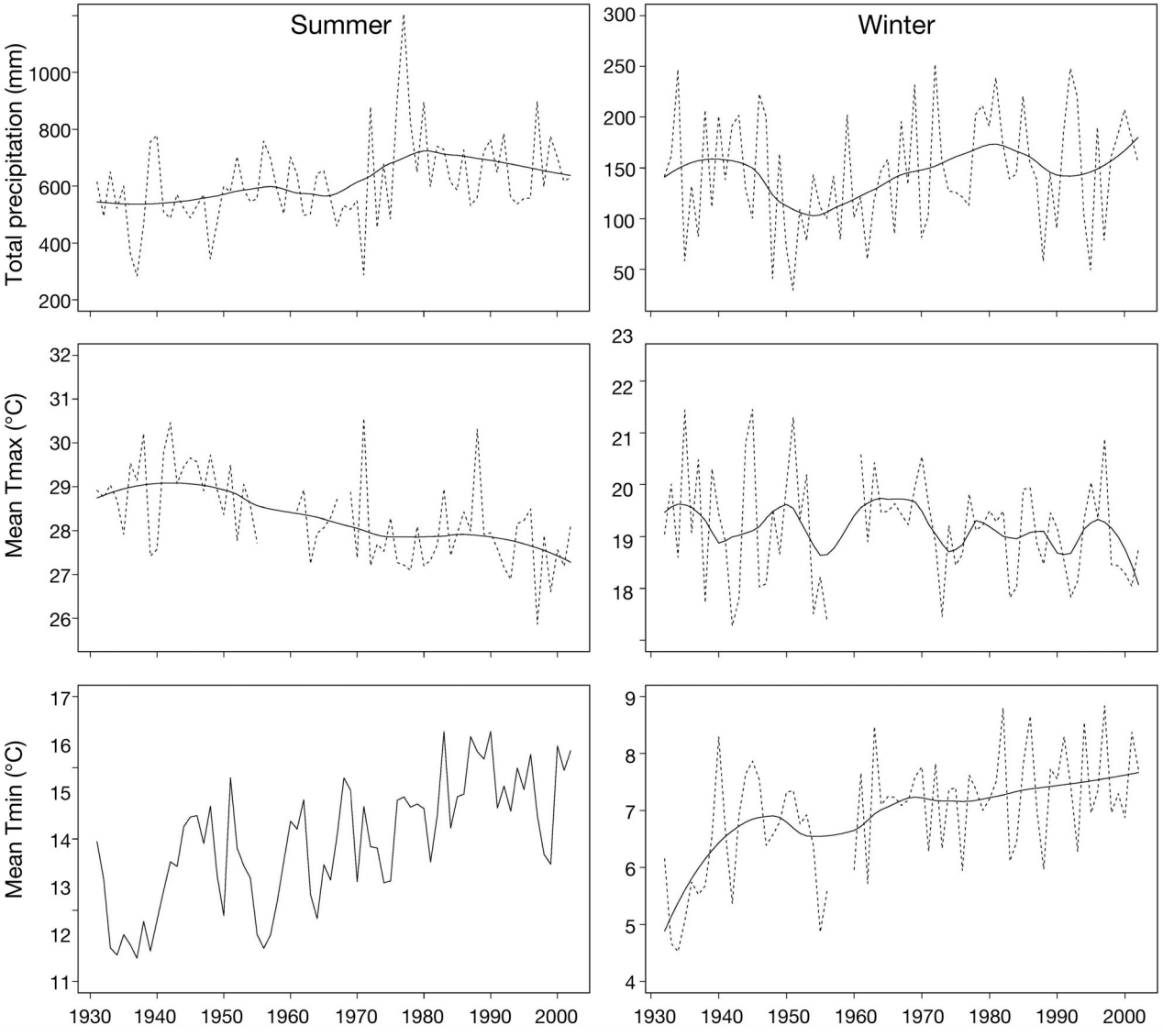

Fig. 10. Optimal smoothed aggregated covariates of the weather time series at Pilar. Other details as in Fig. 9. Note that for summer Tmin the observed and smoothed values coincide

realistic low frequency statistical properties, without any apparent deterioration in high frequency characteristics. This extension involves the incorporation of smoothed (using LOESS) seasonally aggregated climate statistics into the GLM weather generator as covariates. The only non-automatic feature of this extension is the need to determine the degree of smoothing that minimizes overdispersion, but the results are not very sensitive to the exact choice of span parameter in LOESS. With this improvement, climate impact assessments using scenarios of daily weather produced by such generators should be more realistic. Concerning climate change simulations, the proposed method would not necessarily be straightforward to apply unless the seasonal climate statistics were available (e.g. as obtained from simulations by a numerical model of the climate system).

An alternative approach to removing overdispersion would involve replacing an observed covariate with a hidden variable to reflect unobserved shifts in climate regimes on inter-annual or longer (e.g. decadal) time scales. Using a hidden Markov model (HMM, Zucchini & MacDonald 2009) to represent this regime state would allow for long-term persistence, as well as having the advantage of being a fully probabilistic approach (i.e. explicitly modeling the uncertainty about which climate regime is presently occurring). Although HMMs with a hidden daily state variable have been incorporated into time series modeling of daily precipitation (e.g. Hughes et al. 1999), stochastic weather generators with a hidden seasonal state variable have not yet been developed.

Acknowledgements: We gratefully acknowledge the comments of 3 anonymous reviewers. Y.K. acknowledges the support of the National Center for Atmospheric Research (NCAR) as a Visiting Scientist. This research was partially supported by the National Science Foundation (NSF) Coupled Natural Human Systems Program grant CNH-0709681. NCAR is sponsored by the NSF.

LITERATURE CITED

Buishand TA (1978) Some remarks on the use of daily rainfall models. J Hydrol 47:235–249

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: 10.1029/2001WR000906

Cleveland WS (1979) Robust locally-weighted regression and smoothing scatterplots. J Am Stat Assoc 74:829–836

Dubrovsky M, Buchtele J, Zalud Z (2004) High-frequency and low-frequency variability in stochastic daily weather generator and its effect on agricultural and hydrologic modelling. Clim Change 63:145–179

Furrer EM, Katz RW (2007) Generalized linear modeling approach to stochastic weather generators. Clim Res 34: 129–144

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

Hansen JW, Mavromatis T (2001) Correcting low-frequency variability bias in stochastic weather generators. Agric For Meteorol 109:297–310

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman & Hall, London

Hughes JP, Guttorp P, Charles SP (1999) A non-homogeneous hidden Markov model for precipitation occurrence. Appl Stat 48:15–30

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

Magrin G, Travasso M, Rodriguez G (2005) Changes in climate and crop production during the 20th century in Argentina. Clim Change 72:229–249

Maraun D, Wetterhall F, Ireson AM, Chandler RE and others (2010) Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. Rev Geophys 48:RG3003. doi:10.1029/2009RG000314

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

Messina CD, Hansen JW, Hall AJ (1999) Land allocation conditioned on ENSO phases in the Pampas of Argentina. Agric Syst 60:197–212

Podestá G, Bert F, Rajagopalan B, Apipattanavis S and others (2009) Decadal climate variability in the Argentine Pampas: regional impacts of plausible climate scenarios on agricultural systems. Clim Res 40:199–210

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

Srikanthan R, Pegram GGS (2009) A nested multisite daily rainfall stochastic generation model. J Hydrol 371:142–153

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, NY

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

Wilks DS (2010) Use of stochastic weather generators for precipitation downscaling. Wiley Interdisciplinary Reviews: Climate Change 1:898–907

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

Zucchini W, MacDonald IL (2009) Hidden Markov models for time series: an introduction using R. Chapman & Hall, London

Editorial responsibility: Mikhail Semenov, Harpenden, UK

Submitted: May 16, 2011; Accepted: November 23, 2011
Proofs received from author(s): April 19, 2012