Journal of the Korean Data & Information Science Society
2012, **23**(6), 1309–1317

http://dx.doi.org/10.7465/jkdi.2012.23.6.1309
한국데이터정보과학회지

Stochastic precipitation modeling based on Korean historical data

Yongku Kim[1] · Hyeonjeong Kim[2]

[12]Department of Statistics, Yeungnam University

Received 30 October 2012, revised 15 November 2012, accepted 20 November 2012

Abstract

Stochastic weather generators are commonly used to simulate time series of daily weather, especially precipitation amount. Recently, a generalized linear model (GLM) has been proposed as a convenient approach to fitting these weather generators. In this paper, a stochastic weather generator is considered to model the time series of daily precipitation at Seoul in South Korea. As a covariate, global temperature is introduced to relate long-term temporal scale predictor to short-term temporal predictands. One of the limitations of stochastic weather generators is a marked tendency to underestimate the observed interannual variance of monthly, seasonal, or annual total precipitation. To reduce this phenomenon, we incorporate time series of seasonal total precipitation in the GLM weather generator as covariates. It is verified that the addition of these covariates does not distort the performance of the weather generator in other respects.

Keywords: Generalized linear model, overdispersion, precipitation, stochastic weather generator.

1. Introduction

Stochastic weather generators constitute one technique to temporally downscale such climate information. They are commonly used to simulate time series of daily weather, especially minimum and maximum temperature and precipitation amount (Wilks and Wilby, 1999). Among other things, 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, scenarios of daily weather are needed consistent with projected variations in climate.

The recently introduced approach for stochastic weather generators, based generalized linear modeling (GLM), is convenient for this purpose, especially with covariates to account for seasonality and teleconnections with the El Niño-Southern Oscillation phenomenon (McCullagh and Nelder, 1989; Furrer and Katz, 2007). Yet one important limitation of stochastic weather generators is a marked tendency to underestimate the observed interannual variance of seasonally aggregated variables (e.g., Buishand, 1978; Katz and Parlange, 1998). Such a thing is especially pronounced for precipitation. This behavior of a statistical model relative

[1] Corresponding author: Assistant professor, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea. E-mail: ykkim@yu.ac.kr
[2] Graduate student, Department of Statistics, Yeungnam University, Kyungsan 712-749, Korea.

to the data is termed the "overdispersion" phenomenon. It is not clear the extent to which this phenomenon is attributable to an inadequate model for weather variation, as opposed to a failure to take into account climate variation (Katz and Zheng, 1999).

To reduce the overdispersion phenomenon, Kim *et al.* (2012) incorporated time series consisting of seasonal total precipitation and seasonal mean minimum and maximum temperature into the GLM weather generator as additional covariates. These seasonal time series are smoothed using locally weighted scatterplot smoothing (LOESS; Cleveland, 1979; Hastie and Tibshirani, 1990) to avoid introducing underdispersion. It should be noted that Wilks (1989) conditioned a stochastic model for daily precipitation on monthly total precipitation, and that Hansen and Mavromatis (2001) adjusted the parameters of a stochastic weather generator in an ad hoc fashion to correct for overdispersion. In the present paper, a stochastic weather generator is considered to model the time series of daily precipitation at Seoul in South Korea. A similar approach is adopted to reduce overdispersion phenomenon.

In Section 2, the basic GLM approach to stochastic weather generators is briefly reviewed, and then the extension involving the introduction of aggregated climate statistics as covariates is treated. In Section 3, these models are fitted to time series of daily precipitation at Seoul, evaluating the model fit in terms of overdispersion (Section 3). Section 4 consists of a discussion.

2. GLM weather generator

2.1. Basic model

The GLM approach to stochastic weather generators introduced by Furrer and Katz (2007) focuses on the simplest form of generator first proposed by Richardson (1981). Here we only briefly describe this basic GLM weather generator, referring to for more details (also see http://www.image.ucar.edu/˜eva/GLMwgen/). The precipitation occurrence and intensity components of the GLM stochastic weather generator of Furrer and Katz (2007) are essentially the same as in Stern and Coe (1984), who modeled daily precipitation amount as a chain-dependent process, with annual cycles in the parameters, using GLM. Here golobal temperature is used as a covariate instead of the El Niño phenomenon, unlike in Furrer and Katz (2007).

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 $p_t = \mathrm{P}\{J_t = 1\}$, $t = 1, 2, \ldots$, denote the probability of a wet day. Equivalent to a first-order, two-state Markov chain, the probability of precipitation is modeled conditional on the occurrence state on the previous day J_{t-1}:

$$\ln\left(\frac{p_t}{1-p_t}\right) = \mu + \alpha J_{t-1} + \beta_1 C_t + \beta_2 S_t + \beta_3 G_t + \gamma_1 C_t J_{t-1} + \gamma_2 S_t J_{t-1}, \qquad (2.1)$$

where $C_t = \cos(2\pi(t-181)/365)$, $S_t = \sin(2\pi(t-181)/365)$ and G_t is monthly global temperature (land and ocean combined into an anomaly) index. Here the coefficient α permits the conditional probability of precipitation to shift depending on whether or not precipitation occurred on the previous day, β_1 and β_2 determine the phase and amplitude of the sine wave for the annual cycle in these conditional probabilities, β_3 explains the effect of global warming, and γ_1 and γ_2 allow this annual cycle to be separate for the two conditional probabilities. Note that the $(t - 181)$ term is used to make the estimates consistent with those in Furrer and Katz (2007).

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

$$\ln(\mu_t) = \mu + \beta_1 C_t + \beta_2 S_t. \tag{2.2}$$

Here the coefficients β_1 and β_2 determine the phase and amplitude of the sine wave for the annual cycle in the mean intensity.

2.2. Model with aggregated covariates

As mentioned in the introduction, downscaling techniques have emerged as an efficient means of generating more realistic weather scenarios for impact assessments. How to link the larger scale to the smaller scale provides the opportunity to account for our understanding of the relationships and interaction between the two disparate scale processes. Downscaling makes use of the relationship between the meso- or larger scale atmospheric or oceanic predictor variables and station-scale meteorological variables (e.g., Benestad *et al.*, 2008; Katz and Parlange, 1998; Mehrotra *et al.*, 2004). Statistical downscaling relies on the principle that there is necessarily a close relationship between the climate at the large scale and the weather at the local scale.

Kim *et al.* (2012) considered LOESS smoothed seasonal climate statistics as covariates in the GLM weather generator to introduce enough noise into the daily weather statistics and result in "underdispersion". The basis of their approach is to relate long-term temporal scale predictor variables to short-term temporal scale predictands. For example, indices of large-scale atmospheric or oceanic circulation, such as the El Niño-Southern Oscillation (ENSO) phenomenon or global temperature, can be used as covariates in the daily precipitation model. Instead, here we incorporate time series of seasonal climate statistics in the GLM weather generator as covariates in the manner of disaggregation. That is, seasonal total precipitation is used as predictor variable for the model predictands consisting of daily precipitation. Retaining global temperature as a covariate as well would make the interpretation of the model more difficult, as well as complicating the use of the model in the statistical downscaling of seasonal forecasts. Our approach still indirectly takes into account the effects of global temperature on daily weather statistics, because of the well established global temperature in these aggregated climate statistics in the South Korea.

As will be seen, 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 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 is descriptively known as locally weighted polynomial regression. LOESS 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 easily represented by a mathematical formula. Nevertheless, it is a very simple and flexible procedure and resistant to outliers (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 value and the degree of the local polynomial: here we use 0.8 for the degree of smoothing and two for the degree of the local polynomials, see Figure 2.1).

Formally, our approach involves introducing LOESS smoothed seasonally aggregated climate statistics into the basic GLM weather generators specified by (2.1)-(2.2) as follows:

$$\ln\left(\frac{p_t}{1-p_t}\right) = \mu+\alpha J_{t-1}+\beta_1 C_t+\beta_2 S_t+\gamma_1 C_t J_{t-1}+\gamma_2 S_t J_{t-1}+\beta_s I_t P_t^S+\beta_w(1-I_t)P_t^W \quad (2.3)$$

$$\ln(\mu_t) = \mu + \beta_1 C_t + \beta_2 S_t + \beta_s I_t P_t^S + \beta_w(1-I_t)P_t^W \quad (2.4)$$

where I_t is a seasonal indicator (i.e., $I_t = 1$ in summer (April-September) and $I_t = 0$ in winter (October-March)) and P^S and P^W LOESS smoothed summer and winter seasonal total precipitation. Note that the summer and winter time series are smoothed separately. The seasonal indicators in (2.3)-(2.4) allows for different relationships with the aggregated covariates depending on the season. The degree of LOESS smoothing is determined by the criterion based on minimizing the overdispersion phenomenon, through trial and error ranging from the case of no smoothing to as smooth as possible. Note that the overdispersion results are not very sensitive to the value of this smoothing parameter.

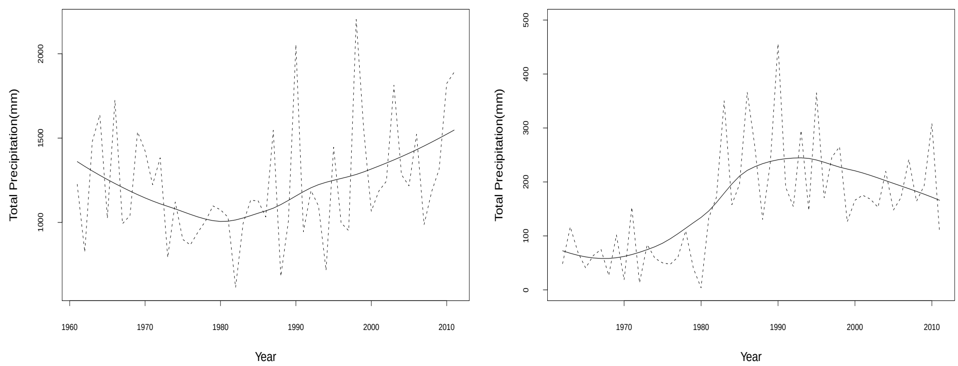

Figure 2.1 Optimal smoothed aggregated covariates of total precipitation during summer (left) and winter (right) seasons. Dashed lines: corresponding observed values of the data series.

3. Fit of GLM weather generator to data

The data used in this study are the daily observations of the precipitation at Seoul for the period 1961-2011 given by Korea Meteorological Administration (KMA). The annual precipitation cycle in this region has a clear maximum in late spring and summer and a marked winter minimum.

Table 3.1 lists the parameter estimates and standard errors, obtained through repeated application of the"glm" function in the open source software R. Each covariate category is statistically significant except for interaction terms (dispersion parameter for binomial family and Gamma family are taken to be 1 and 2.572, respectively). Note that 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 at Seoul.

Furrer and Katz (2007) already mentioned that this form of GLM stochastic weather generator underestimates the observed standard deviation of annual and summer (i.e., October through March in South America) total precipitation, notwithstanding their inclusion of El Niño (in our case, global temperature) as a covariate.

Table 3.2 summarizes the estimated coefficients and associated standard errors for all of the components of the stochastic weather generator fitted to precipitation data at Seoul. Comparing this table with the corresponding models in Table 3.1, the AIC and BIC always select the one with the aggregated covariates as being a better fit. The estimated coefficients of the remaining covariates do not change very much (especially those for autocorrelation and dependence) when the aggregated covariates are included (dispersion parameter for binomial family and Gamma family are taken to be 1 and 2.518, respectively).

Table 3.1 Estimated coefficients (Coef.) and standard error (SE) values for all components of the basic stochastic weather generator at Seoul.

Covariate category	Precipitation Occurrence			Precipitation intensity (mm)		
	Term	Coef.	SE	Term	Coef.	SE
Constant	μ	-1.67	0.032	μ	2.18	0.025
Autocorrelation	J_{t-1}	1.17	(0.040)	–	–	–
Global Temp	G_t	0.677	0.076	–	–	–
Seasonality	C_t	0.69	0.033	C_t	0.91	0.034
	S_t	0.25	0.031	S_t	0.30	0.034
AIC		19097			32881	
BIC		19150			32903	

Table 3.2 Estimated coefficients (Coef.) and standard error (SE) values for all components of the stochastic weather generator with aggregated climate statistics as covariates at Seoul.

Covariate category	Precipitation Occurrence			Precipitation intensity (mm)		
	Term	Coef.	SE	Term	Coef.	SE
Constant	μ	-2.68	0.079	μ	1.74	0.118
Summer	$I_tP_t^S$	0.18	0.014	$I_tP_t^S$	0.09	0.019
Winter	$(1-I_t)P_t^W$	0.32	0.083	$(1-I_t)P_t^W$	0.28	0.115
Autocorrelation	J_{t-1}	1.06	0.041	–	–	–
Seasonality	C_t	0.65	0.059	C_t	0.74	0.073
	S_t	0.21	0.031	S_t	0.31	0.034
Interaction	C_tJ_{t-1}	0.18	0.057	–	–	–
	S_tJ_{t-1}	0.11	0.054	–	–	–
AIC		18859			32858	
BIC		18907			32895	

Figure 3.1 illustrate how our proposed model performs in reproducing variances of annual, as well as summer and winter, total precipitation. Time series of daily weather were simulated over a 51 year period and aggregated statistics calculated, with the simulation exercise being repeated 500 times. Shown are boxplots (showing the minimum, lower quartile, median, upper quartile, and maximum) of the standard deviation (SD) of the aggregated

statistics along with the corresponding values for the data series over the 51 years. Boxplots are a preferred method of data analysis in many applications, as they show the range of variation in the statistics of simulations and provide a straightforward method of comparing these statistics with the historical data. The proposed model virtually eliminates the overdispersion phenomenon in nearly all cases, with the results (not shown) not being very sensitive to the precise value of the parameter governing the degree of smoothing in LOESS (for example, the overdispersion in the annual total precipitation is still virtually eliminated if the degree of smoothing is 0.9 instead of 0.8). The GLM weather generator with less smoothed aggregated covariates tends to overestimate inter-annual variances (i.e., underdispersion). Note that precipitation in the winter season is simply not as variable as in the summer season.

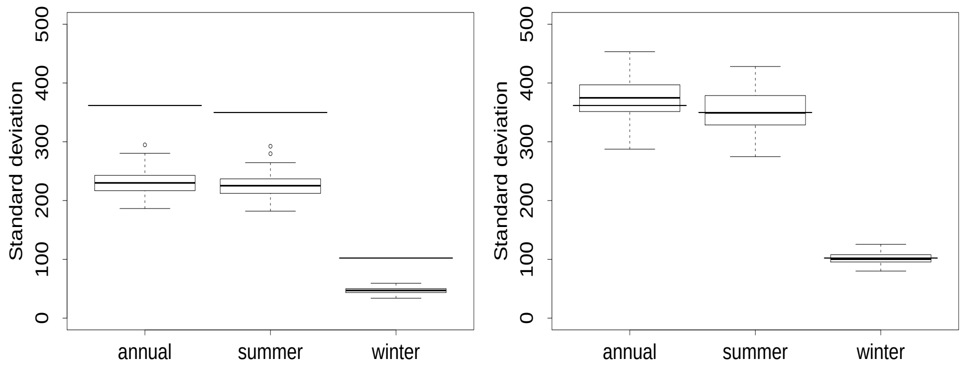

Figure 3.1 Boxplots of simulated Sd of annual, summer and winter total precipitation(mm) based on the original model (left) and with smoothed seasonal covariates (right). Horizontal solid lines: corresponding observed values of the data series.

To demonstrate performance of the GLM weather generator, we look into more meaningful daily statistics such as dry spells. The distribution of observed dry spells is compared to that of simulated dry spells during summer season (see Figure 3.2). The Markov chain model for daily precipitation occurrence can be fully characterized by the 2 transition probabilities $p_{11}(t) = Pr\{J_t = 1|J_{t-1} = 1\}$, the conditional probability of a wet day given the previous day was wet, and $p_{01}(t) = Pr\{J_t = 1|J_{t-1} = 0\}$, the conditional probability of a wet day given the previous day was dry. As a function of the time of year, Figure 3.3 shows $p_{11}(t)$ and $p_{01}(t)$, respectively. In each case, the curves for the GLM weather generator are included along with the observed daily statistics.

4. Concluding remarks

It has been 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 re-

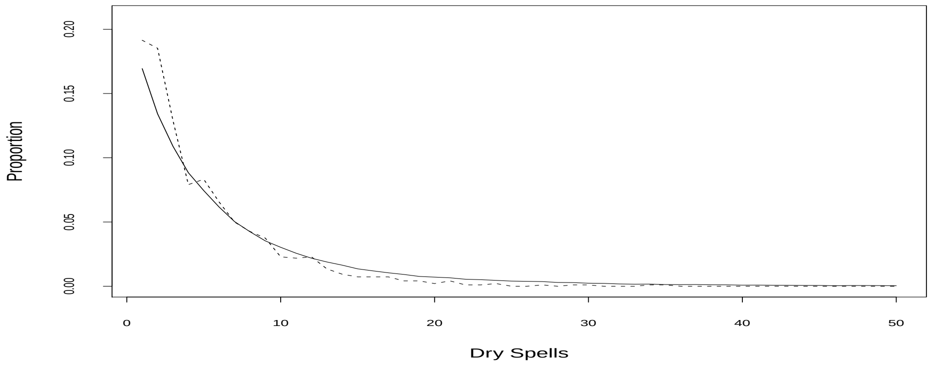

Figure 3.2 Distributions of observed dry spells (dashed line) and simulated dry spells (solid line) during summer season over the 51 years.

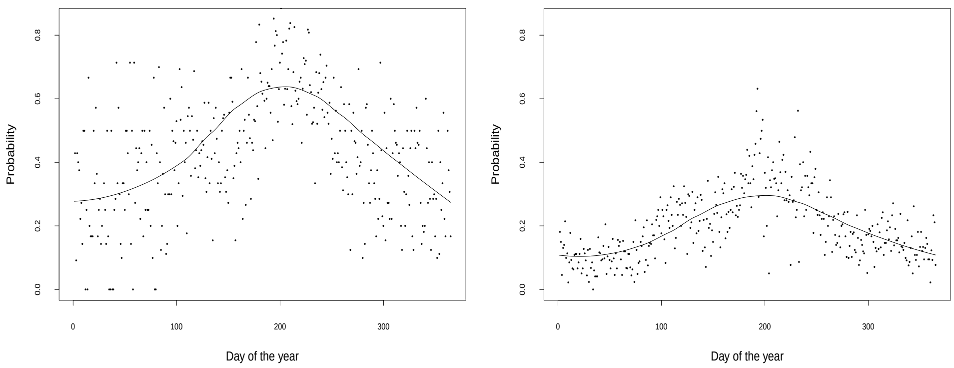

Figure 3.3 Modeled transition probabilities $p_{11}(t)$ (left) and $p_{01}(t)$ (right) with smoothed aggregated covariates. Dots: empirical transition probabilities, i.e. frequencies of observed transitions calculated separately on each day of the year.

alistic statistical properties resulting in better risk assessments. This extension involves the incorporation of smoothed seasonally aggregated climate statistics into the GLM model 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 degree of smoothing. However, it is still somewhat uninformative to remove overdispersion through explicit use of seasonal aggregated climate statistics as covariates in the GLM weather generator.

Basically, our approach can be extended to other cliamte variables such as daily minimum and maximum temperature. Let (X_t, Y_t) denote the minimum and maximum temperature on day t of a given year, jointly modeled as a bivariate first-order autoregressive AR(1) process (as in Richardson, 1981; Furrer and Katz, 2007). This bivariate process is modeled indirectly through two univariate linear models of the form:

$$X_t = \mu_{X,0} + \mu_{X,1} J_t + \varphi_X X_{t-1} + \psi_X Y_{t-1} + \beta_{X,1} C_t + \beta_{X,2} S_t + \beta_{X,S} I_t N_t^S + \beta_{X,W} (1 - I_t) N_t^W + \varepsilon_{X,t} \quad (4.1)$$

$$Y_t = \mu_{Y,0} + \mu_{Y,1} J_t + \varphi_Y Y_{t-1} + \psi_Y X_t + \beta_{Y,1} C_t + \beta_{Y,2} S_t + \beta_{Y,S} I_t M_t^S + \beta_{Y,W} (1 - I_t) M_t^W + \varepsilon_{Y,t} \quad (4.2)$$

where N^S and N^W (M^S and M^W) are LOESS smoothed summer and winter seasonal mean minimum (maximum) temperatures. Note that the summer and winter time series are smoothed separately. This approach can also be extended to multisite weather generate by using a latent Gaussian process to drive precipitation occurrence and a probability integral transformed Gaussian process for intensity.

Statistical downscaling is based on the view that the regional climate is conditioned by two factors: the large scale climatic state and regional/local physiographic features. For a climate projection perspective, it is also interesting how the incorporation of such seasonally aggregated climate statistics facilitates statistical downscaling of seasonal climate forecasts. These results are encouraging in that the methodology provides a robust tool to generate weather sequences consistent with any seasonal climate forecast of potential use in resources planning and management. In the case of seasonal forecasts, the GLM weather generator makes it straightforward to translate the uncertainty in the seasonal forecast product into that for the corresponding conditional daily weather statistics.

References

Benestad, R. E., Hanssen-Bauer, I. and Chen, D. (2008). *Empirical-statistical downscaling*, World Scientific, Singapore.

Buishand, T. A. (1978). Some remarks on the use of daily rainfall models. *Journal of Hydrology*, **47**, 235-249.

Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829-836.

Furrer, E. M. and Katz, R. W. (2007). Generalized linear modeling approach to stochastic weather generators. *Climate Research*, **34**, 129-144.

Hansen, J. W. and Mavromatis, T. (2001). Correcting low-frequency variability bias in stochastic weather generators. *Agricultural and Forest Meteorology*, **109**, 297-310.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, Chapman and Hall, New York.

Katz, R. W. and Parlange, M. B. (1998). Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, **11**, 591-601.

Katz, R. W. and Zheng, X. (1999). Mixture model for overdispersion of precipitation. *Journal of Climate*, **12**, 2528-2537.

Kim, Y., Katz, R. W., Rajagopalanc, B., Furrer, E. M. and Podestá, G. (2012). Reducing overdispersion in stochastic weather generators using a generalized linear modeling approach. *Climate Research*, **53**, 13-24.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R.E., Kendon, E. J., Widmann, M., Brienen, N., Rust, H. W., Sauter, T., Themeβl, M. *et al.* (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, **48**, doi:10.1029/2009RG000314.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, 2d ed., Chapman and Hall.

Mehrotra, R., Sharma, A. and Cordery, I. (2004). Comparison of two approaches for downscaling synoptic atmospheric patterns to multisite precipitation occurrence. *Journal of Geophysical Research D: Atmospheres*, **109**, 1-15.

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, **17**, 182-190.

Stern, R. D. and Coe, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society A*, **147**, 1-34.

Wilks, D. S. and Wilby, R. L. (1999). The weather generator game: A review of stochastic weather models. *Progress in Physical Geography*, **23**, 329-357.

Wilks, D. S. (2010) Use of stochastic weather generators for precipitation downscaling. *Wiley Interdisciplinary Reviews: Climate Change*, **1**, doi:10.1002/wcc.85

Wilks, D. S. (1989). Conditioning stochastic daily precipitation models on total monthly precipitation. *Water Resources Research*, **25**, 1429-1439.