Environmental Modelling & Software 25 (2010) 1252–1255

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Short communication

A simple stochastic weather generator for ecological modeling

A.G. Birt [a,*], M.R. Valdez-Vivas [b], R.M. Feldman [b], C.W. Lafon [c], D. Cairns [c], R.N. Coulson [a], M. Tchakerian [a], W. Xi [a], J.M. Guldin [d]

[a] Department of Entomology, Texas A&M University, College Station, TX 77843, United States
[b] Department of Industrial Engineering, Texas A&M University, College Station, TX 77843, United States
[c] Department of Geography, Texas A&M University, College Station, TX 77843, United States
[d] Southern Research Station, USDA Forest Service, Hot Springs, AR 71901, United States

ARTICLE INFO

Article history:
Received 15 December 2009
Received in revised form
7 March 2010
Accepted 9 March 2010
Available online 15 April 2010

Keywords:
Weather generator
Climate
Stochastic model
Ecology

ABSTRACT

Stochastic weather generators are useful tools for exploring the relationship between organisms and their environment. This paper describes a simple weather generator that can be used in ecological modeling projects. We provide a detailed description of methodology, and links to full C++ source code (http://weathergen.sourceforge.net) required to implement or modify the generator. We argue that understanding the principles of weather generation will allow ecologists to tailor a solution for their own requirements. The detailed, repeatable methodology we present demonstrates that weather generation is relatively straightforward for ecologists to implement and modify.

1. Introduction

Stochastic weather generators are probabilistic models that are used to simulate weather data at a specific site or region by analyzing historical weather data and then generating a time-series of weather variables with statistical properties identical to the historical data. Weather is an important driver of many ecological models and although historical data can be used in such studies, Table 1 highlights a number of features and advantages of using weather generators that make them useful for many applications.

Stochastic weather generation is a four-step process:

1. Develop a model structure (the inter-relationships between parameters) capable of reproducing realistic weather patterns.
2. Parameterize the model for a location or region using historical weather observations.
3. Statistically analyze model outputs to ensure *adequate* representation of the historical weather record.
4. Generate new sets of data that represent weather patterns for the location or region of interest that can be used in an ecological model.

Ecologists are most likely to be directly interested in step 4 of the above process. However, there are a number of reasons why a more detailed understanding of weather generation may be beneficial to the ecologist. Firstly, for weather data to be seamlessly used in an ecological model (step 4) the generator needs to output the appropriate variables at an appropriate temporal and spatial resolution and in a specific format. Secondly an *adequate* representation of real weather patterns (step 3) depends upon the context of an ecological study. Although statistical methods can be used to assess similarity between observed and simulated weather, ultimately the sensitivity of a model to weather data dictates the practical interpretation of statistics. Thirdly, it may be necessary to develop a generator that maximizes the utility of available historical weather data (for example in data poor locations).

Although a number of weather generators already exist, increased knowledge of weather generation will help users choose (or modify) a solution most appropriate for a specific study. The goal of this paper is to present the methodology of a simple, single site stochastic weather generator. We make no claims about the novelty (the basic elements have been adapted from those described by Semenov et al., 1998; Semenov and Barrow, 2002); or relative superiority of the generator; or its applicability to various ecological studies. Instead we aim to disambiguate methodology and encourage users to move from a 'black box' approach towards more customized, ecologically driven applications of weather generation whenever they may be more applicable.

* Corresponding author.
E-mail address: abirt@tamu.edu (A.G. Birt).

1364-8152/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2010.03.006

Table 1
Potential features of weather generators and the advantages to ecological modeling. Note the weather generator we describe here does not demonstrate all these features or benefits.

Feature of weather generators	Potential benefits
Production of unlimited amounts of weather data with statistical properties realistic to historical pattern	–Models can be run for long periods of time and using many permutations of weather
Can be 'trained' using incomplete historical weather data (i.e. records with segments of missing data)	–Weather data is complete, ready to use and without missing values –Allows models to be run for locations with limited historical data
Large historical weather database can be condensed into a much smaller number of weather generator parameters	–Allows small, portable weather databases to be created –Parameter sets can be sent over the internet and weather time-series created on client machines to reduce server load –Potential for rapid generation of 'synthetic' weather data versus retrieval of stored historic data
Weather generator leads to understandable, mechanistic parameters	–Construction or modification of a generator leads to a greater understanding of the inter-relationship between weather variables –Parameters can be manipulated to simulate realistic changes in climate (e.g. using climate change scenarios)
Utilization of historical weather data from multiple sources. For example different data collection methods (satellite measured surface temperatures, station measurements, radar) or spatial or temporal scales of measurement	–Data can be generated at spatial and temporal scales most important to a study
Repeatable, mathematical and scientific formulation	–Weather generator methodology can be modified and improved by users of weather data –Open source code can be provided to provide detailed, unequivocal methodology –Software can be distributed that can be seamlessly integrated into ecological models, streamlining the process of simulating and analyzing systems

2. Methods

2.1. The rain model

In this generator (and many others), we assume that daily temperatures depend on the occurrence of rain. The first step in the weather generator is to develop a rain model capable of describing the length of wet and dry series and for wet days, the amount of rainfall that occurs. Semenov et al. (1998) suggest the use of semi-empirical distributions to model these quantities. A semi-empirical distribution is based on a histogram consisting of a given number of intervals (ten in the example below) denoted by:

$$H = \{a_0, a_1, \ldots a_{10}, h_1, \ldots h_{10}\} \quad (1)$$

where, for $i = 1, \ldots, 10$, h_i denotes the number of data points that are observed within the interval (a_{i-1}, a_i). The generation of a random variate from a semi-empirical distribution based on H occurs in two steps: first one of the ten intervals is generated according to the probabilities representing the proportion of events in each interval and then a value within the selected interval is generated using a uniform distribution. Each month has its own semi-empirical distribution that describes the length of wet and dry series and the amount of rain falling on each wet day. Note then that each semi-empirical distribution requires 21 parameters that describe 10 intervals (and therefore 11 interval boundaries) with a probability associated with each interval. In our case, interval size is smallest for small values because there are more days with lighter rainfall and series of shorter lengths.

Interval widths for cycle lengths are determined by the maximum length of a cycle for a given month using the following rules. Assume that the maximum number of days for a cycle is denoted by d^*, and let the width of the ith interval be w_i; in other words, $w_i = a_i - a_{i-1}$.

If $d^* \leq 10$, then $w_i = 1$ for all i. If $10 < d^* > 55$ then $w_1 = 1$, $w_i = w_{i-1}$ for $i = 2, \ldots, i^*$ and $w_i = w_{i-1} + 1$ for $i = i^* + 1, \ldots, 10$ where i^* is chosen so that a_{10} is as small as possible such that $a_{10} \leq d^*$. (Note that in this case widths are constant until the index i^*, at which point the widths increase by 1. The index i^* is chosen to postpone the increase in widths as long as possible.) If $d > 55$ then $w_i = w_{i-1} + 1$ for $i = 2, \ldots, 10$ with w_1 being chosen so that a_{10} is as small as possible such that $a_{10} \leq d^*$.Interval widths for rainfall quantities are described using the following ten fractions: $f_1 = f_2 = 0.25$, $f_3 = f_4 = 0.5$, $f_5 = f_6 = 1.0$, $f_7 = f_8 = 1.5$, $f_9 = f_{10} = 1.75$. Let the maximum quantity of rain for a day within the given month be denoted by q^*, then the width of the ith interval is given as $w_i = q^* f_i / 10.0$ for $i = 1, \ldots, 10$.

Given this model structure, the next steps are to parameterize the distributions. Then new data can be generated from these distributions. Note that the procedures for parameterizing and creating rainfall data are essentially the opposite of each other. To parameterize the semi-empirical distributions (daily) historical weather data (comprising minimum and maximum temperatures and rainfall) are input into the program. The program searches for wet and dry series which are attributed to the semi-empirical histogram for the month in which that series originates. Similarly, daily rainfall amounts are added to the semi-empirical distribution for a given month. To generate rainfall data from the parameterized model, a series begins by first choosing the length of the series from the semi-empirical distribution of the appropriate month. If a day falls within a wet series, a positive nonzero value for daily rain is generated from the semi-empirical distribution of the month in which the day occurs.

The amount of rainfall each day form an independent sequence given the cycle length, and the amounts are also independent of the cycle length. The lengths of adjacent series are also modeled independently. For example, suppose a wet series begins on June 30 and lasts five days. The length of the wet series is considered data for June and the amount of rain on the first day of the series is also considered June data; whereas, the data for the amount of rain on the next four days would be considered July data. All climate variables determined by a semi-empirical distribution are solely dependent on the month to which they correspond.

The histogram-dependent structure of this distribution allows it to take on a variety of shapes thus providing a high degree of flexibility. A disadvantage, acknowledged by Semenov et al. (1998), is that for a histogram with 10 intervals each distribution requires 21 parameters. Much more than the 3 required for earlier serial rain models (e.g. Racsko et al., 1991). Modifications to the weather generator could be made by changing the number of intervals or the size of each interval or even by replacing the semi-empirical distribution with some other function capable of adequately representing the length of wet series and the amount of rain falling on each wet day.

2.2. The temperature model

Minimum and maximum temperatures are stochastic processes with daily means and standard deviations conditioned on the wet and dry status of the day. Temperature means are modeled by a third order Fourier series fitted to historical daily minimum and maximum temperatures. In other words, four models are developed for the mean temperatures (minimum-dry, minimum-wet, maximum-dry, maximum-wet).

The procedure for modeling each of these quantities is the same: For example, a model of daily minimum-dry means is developed by finding minimum temperature records for days with no rainfall. Each instance is described by an ordered pair (d_i, y_i) where d_i is the Julian date of the dry day and y_i is the minimum temperature for that day (thus if the historical data contains 24 years and half were dry, then this would yield $24 \times 365 \times 0.5$ such data pairs). The following function is then fitted (using least squares) to these data points:

$$y = a_0 + \sum_{n=1}^{3} [a_n \cos(2n\pi d/365) + b_n \sin(2n\pi d/365)] \quad (2)$$

Where d represents the day of the year, y represents the temperature, and the constants a_0, a_1, a_2, a_3, b_1, b_2, b_3 are fitted parameters.

Four models are also developed to describe the standard deviations associated with the data (again for minimum-dry, minimum-wet, maximum-dry, maximum-wet). For example, to obtain the regression model for standard deviation associated with minimum-dry, the standard deviation for each day of the year is calculated

(using the (d_i, s_i) pairs used for modeling the means) before the regression (using Eq. (2)) is performed. If a given day of the year has less than 2 dry (or wet) days, no standard deviation for that date is calculated (note that the only difference in performing the regression for means versus standard deviations is that the former involves thousands of data points but the latter will use at most only 365).

In real weather patterns, daily temperatures are highly correlated. The most common formula for determining an auto-correlation coefficient for a series of random variables makes the assumption that the series is a covariance-stationary series (i.e., the daily means, daily variances, and lag-n correlations coefficients are constant). Thus, before estimating correlation coefficients, we standardized temperatures based on the regression fits, and the ordered pair made up of daily minimum and maximum temperatures are assumed to be a covariance-stationary sequence of normally distributed random vectors with mean zero and variance one. Notice that the standardization procedure uses the regression curves which helps to remove the time-of-year bias from the temperature sequence; thus, we assume that the correlation coefficients are constant throughout the year. Consider the following definitions:

$\tau_{\max,i}$	max temperature for day i
$\tau_{\min,i}$	min temperature for day i
$\widehat{y}_{\max,i}$	mean of max temperature for day i based on Eq. (2)
$\widehat{y}_{\min,i}$	mean of min temperature for day i based on Eq. (2)
$\widehat{s}_{\max,i}$	standard deviation of max temperature for day i based on Eq. (2)
$\widehat{s}_{\min,i}$	standard deviation of min temperature for day i based on Eq. (2)
$Z_{\max,i}$	transformed max temperature: $Z_{\max,i} = (\tau_{\max,i} - \widehat{y}_{\max,i})/\widehat{s}_{\max,i}$
$Z_{\min,i}$	transformed min temperature: $Z_{\min,i} = (\tau_{\min,i} - \widehat{y}_{\min,i})/\widehat{s}_{\min,i}$
$r_{\max,\max}$	lag-1 auto correlation for max temperatures
$r_{\min,\min}$	lag-1 auto correlation for min temperatures
r_{daily}	correlation between min and max temperatures within a day
$r_{\max,\min}$	lag-1 auto correlation between max temperature to next day's min temperature
$r_{\min,\max}$	lag-1 auto correlation between min temperature to next day's max temperature
$ssq_{\max}$	sum of squares of max transformed temperatures (hopefully close to n)
$ssq_{\min}$	sum of squares of min transformed temperatures (hopefully close to n)
n	total number of days of data (n equals the number of years of data times 365)

There are five correlation coefficients to estimate, obtained from the transformed data as follows:

$$r_{\max,\max} \quad \left(\sum_{i=1}^{n-1} Z_{\max,i} \times Z_{\max,i+1}\right) \Big/ ssq_{\max} \tag{3}$$

$$r_{\min,\min} \quad \left(\sum_{i=1}^{n-1} Z_{\min,i} \times Z_{\min,i+1}\right) \Big/ ssq_{\min} \tag{4}$$

$$r_{daily} \quad \left(\sum_{i=1}^{n} Z_{\max,i} \times Z_{\min,i}\right) \Big/ \sqrt{ssq_{\max} \times ssq_{\min}} \tag{5}$$

$$r_{\min,\max} \quad \left(\sum_{i=1}^{n-1} Z_{\min,i} \times Z_{\max,i+1}\right) \Big/ \sqrt{ssq_{\max} \times ssq_{\min}} \tag{6}$$

$$r_{\max,\min} \quad \left(\sum_{i=1}^{n-1} Z_{\max,i} \times Z_{\min,i+1}\right) \Big/ \sqrt{ssq_{\max} \times ssq_{\min}} \tag{7}$$

Once the models have been parameterized using the historic data, the models of mean and standard deviations of daily temperatures (minimum-dry, minimum-wet, maximum-dry, maximum-wet) and the five correlation coefficients ($r_{\max,\max}$, $r_{\min,\min}$, r_{daily}, $r_{\min,\max}$, $r_{\max,\min}$) are used to generate temperature data (daily minimums and maximums). Scheuer and Stoller (1962) describe a method to generate correlated normal vectors that we adapt to apply an auto-correlated sequence of ordered pairs. First generate a sequence of independent standard (i.e. mean zero and variance one) normal random variates: for the standard variate T, then the daily temperature (either min or max) is:

$$\tau = \mu + \sigma T \tag{8}$$

where μ and σ are the mean and standard deviation from the appropriate fitted regression models. Daily temperatures are generated sequentially, first we determine whether the day is wet or dry (from the rain model), then generate the minimum and maximum temperatures are for each day. Assuming that rain has already been determined using the rain model so that $y_{\max,i}$, $y_{\min,i}$, $s_{\max,i}$, and $s_{\min,i}$ are known; and that a correlated sequence of standard normal variates have been generated, the following algorithm is used to generate data.

1. Set day index $i = 1$.
2. Min temperature, day 1: $\mu = \widehat{y}_{\min,1}$ and $\sigma^2 = \widehat{s}^2_{\min,1}$
3. Max temperature, day 1

$$\mu = \widehat{y}_{\max,1} + r_{daily} \times \left(\widehat{s}_{\max,1}/\widehat{s}_{\min,1}\right) \times \left(\tau_{\min,1} - \widehat{y}_{\min,1}\right)$$

$$\sigma^2 = \widehat{s}^2_{\max,1} \times \left(1 - r^2_{daily}\right)$$

4. Min temperature, day $i+1$:

$$\mu = \widehat{y}_{\min,i+1} + \left(r_{\max,\min} - r_{daily} r_{\min}\right) \times \left(\widehat{s}_{\min,i+1}/\widehat{s}_{\max,i}\right) \times \left(\tau_{\max,i} - \widehat{y}_{\max,i}\right) \Big/ \left(1 - r^2_{daily}\right) + \left(r_{\min} - r_{daily} r_{\max,\min}\right) \times \left(\widehat{s}_{\min,i+1}/\widehat{s}_{\min,i}\right) \times \left(\tau_{\min,i} - \widehat{y}_{\min,i}\right) \Big/ \left(1 - r^2_{daily}\right)$$

$$\sigma^2 = \widehat{s}^2_{\min,i+1} - r_{\max,\min}\left(r_{\max,\min} - r_{daily} r_{\min}\right) \times \widehat{s}^2_{\min,i+1} \Big/ \left(1 - r^2_{daily}\right) - r_{\min}\left(r_{\min} - r_{daily} r_{\max,\min}\right) \times \widehat{s}^2_{\min,i+1} \Big/ \left(1 - r^2_{daily}\right)$$

5. Max temperature, day $i+1$:

$$T = \left(\tau_{\min,i+1} - \widehat{y}_{\min,i+1}, \tau_{\max,i} - \widehat{y}_{\max,i}, \tau_{\min,i} - \widehat{y}_{\min,i}\right)^T$$

$$\upsilon = \left(r_{daily}\widehat{s}_{\min,i+1}\widehat{s}_{\max,i+1}, r_{\max}\widehat{s}_{\max,i}\widehat{s}_{\max,i+1}, r_{\min,\max}\widehat{s}_{\min,i}\widehat{s}_{\max,i+1}\right)$$

$$C = \begin{bmatrix} \widehat{s}^2_{\min,i+1} & r_{\max,\min}\widehat{s}_{\max,i}\widehat{s}_{\min,i+1} & r_{\min}\widehat{s}_{\min,i}\widehat{s}_{\min,i+1} \\ r_{\max,\min}\widehat{s}_{\max,i}\widehat{s}_{\min,i+1} & \widehat{s}^2_{\max,i} & r_{daily}\widehat{s}_{\max,i}\widehat{s}_{\min,i} \\ r_{\min}\widehat{s}_{\min,i}\widehat{s}_{\min,i+1} & r_{daily}\widehat{s}_{\max,i}\widehat{s}_{\min,i} & \widehat{s}^2_{\min,i} \end{bmatrix}$$

$$\mu = \widehat{y}_{\max,i+1} + \upsilon C^{-1} T$$

$$\sigma^2 = \widehat{s}^2_{\max,i+1} - \upsilon C^{-1} \upsilon^T$$

6. Increment day index and return to step 4

Occasionally, a simulated day's maximum temperature is less than the minimum. If this occurs, both values are set to the average value of the two temperatures.

2.3. Testing the weather generator

We have chosen five locations (representative of the study and data availability for which the generator was developed), each within the south eastern US, to test the weather generator. We use a two step process for determining the adequacy of the model for our application:

Table 2
Locations used to test weather generator. The locations are representative of the broad range of climates relevant to the ecological model for which the generator was originally designed.

Location	Latitude, longitude	Longitude	Elevation (m)	Mean annual temp (°C)	Mean annual low (°C)	Mean annual high (°C)	Mean annual precipitation (cm)
Asheville, NC	35.57	−82.57	649	12.9	−4	28.3	120.9
Lufkin, TX	31.34	−94.72	95	19.2	3	34	132.1
Nacogdoches, TX	31.6	−94.65	91	19.8	3.2	33.8	122.8
Nashville, TN	36.17	−86.8	202	15.2	−2	32.1	122.2
Richmond, VA	37.54	−77.46	63	14.3	−2.4	29.8	109.7

Table 3
Results of statistical analyses comparing historical data to generated weather variables. Included are *t*-tests and *f*-tests for the difference between means and variances of maximum and minimum temperatures on each Julian day and *t*-tests and *z*-test for average rainfall and the probability of rain on each Julian day. The proportions represent the number of days (out of 365) that the *t*-, *f*- or *z*-tests rejected the null hypothesis (at 5% level) of equal means, variances or probabilities between historic and simulated data sets.

Location	Years of data (*N*)	Maximum temperature		Minimum temperature		Precipitation	
		t-test	*f*-test	*t*-test	*f*-test	*t*-test	*z*-test
Asheville, NC							
Daymet	24	0.030	0.049	0.038	0.058	0.088	0.033
NCDC	100	0.055	0.036	0.036	0.055	0.055	0.060
Lufkin, TX							
Daymet	24	0.044	0.085	0.049	0.101	0.096	0.071
NCDC	68	0.038	0.068	0.063	0.077	0.082	0.038
Nacogdoches, TX							
Daymet	24	0.058	0.088	0.038	0.074	0.115	0.041
NCDC	33	0.049	0.063	0.066	0.068	0.118	0.079
Nashville, TN							
Daymet	24	0.036	0.033	0.049	0.047	0.071	0.044
NCDC	61	0.055	0.052	0.060	0.066	0.066	0.055
Richmond, VA							
Daymet	24	0.049	0.055	0.038	0.047	0.121	0.044
NCDC	60	0.066	0.060	0.044	0.063	0.074	0.058

1) Statistically assess the similarity of weather patterns produced by the generator compared to real world weather patterns.
2) Evaluate the strengths and weaknesses of the model in light of the ecological application for which it was intended.

The five locations we have chosen are shown in Table 2. Daily weather station data for each location was obtained from the National Climatic Data Center (www.ncdc.noaa.gov). However, the study (for which this weather generator was designed) requires weather data for remote locations across the US for which there is no available weather station data. Therefore, we also use a spatially interpolated weather resource maintained by the National Center for Atmospheric Research (http://DayMet.org) as a data source (see Thornton et al., 1997).

Three standard statistical tests (Miller and Freund, 1985) were used to assess the similarity of generated versus real data (other tests could easily be incorporated into the generator code driven by the context of its use):

1) A *t*-test to determine if a significant difference exists between daily means of the generated versus actual data.
2) An *F*-test to determine if a significant difference exists between the daily variances of the generated versus actual data (temperatures only).
3) A *z*-test to determine if a significant difference between the daily probability of rain for the generated versus actual data (Eq. (9)).

$$z = \frac{p1 - p2}{\sqrt{p(1-p)\left(\frac{1}{n1} + \frac{1}{n2}\right)}} \tag{9}$$

where z is the z score, p is the pooled probability of rain, $p1$ and $p2$ are the proportions of rainy days in the generated data and historical data respectively, and $n1$ and $n2$ are the sample sizes used to calculate these proportions. At each location and for each day we recorded how many days out of 365 the test results indicated that the means or variances were not-equal (i.e. different) using a Type-I error of 5%.

3. Results and discussion

Table 3 shows the statistical comparison between actual weather data and those created by the weather generator. The weather generator faithfully reproduces the historical data from weather stations in the National Climatic Data Center database and those available via daymet.org.

Statistical adequacies of the weather generator are however only one measure of its success or utility. In our case, the overriding motivation for the development of a weather generator was to produce a tool to be seamlessly incorporated into ecological models simulating temperature and precipitation driven processes in the Southern United States. We argue that weather generation is an endeavor driven by the needs of a particular study. As such, ecologists should play a role designing such tools. Other ecological studies may require more (or different) weather variables (for example direct or diffuse radiation, wind speed and direction; or even lightning), or data for different spatial or temporal scales (for example spatially correlated regional instead of point data or hourly instead of daily data). The variety and specificity of ecological models; the underlying importance of weather as a driver of ecological processes; and the availability of different kinds of historical weather data should drive the design and use of weather generators. We hope that the methodology and code presented here will encourage the open, transparent development of weather generators and their application to ecological studies.

Acknowledgements

This research was supported in part by the USDA Forest Service Cooperative agreement SRS 06-CA-11330124-196 and through the Knowledge Engineering Laboratory, College of Agriculture, Texas A&M University.

References

Miller, I., Freund, J., 1985. Probability and Statistics for Engineers, third ed. Prentice Hall, Englewood Cliffs, NJ.
Racsko, P., Szeidl, L., Semenov, M., 1991. A serial approach to local stochastic weather models. Ecological Modelling 57, 27–41.
Scheuer, E., Stoller, D., 1962. On the generation of normal random vectors. Technometrics, 278–281.
Semenov, M., Barrow, E., 2002. LARS-WG: a stochastic weather generator for use in climate impact studies. User Manual Version 3.
Semenov, M., Brooks, R., Barrow, E., Richardson, C., 1998. Comparison of the WGEN and LARS-WG stochastic weather generators for diverse climates. Climate Research 10, 95–107.
Thornton, P., Running, S., White, M., 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. Journal of Hydrology 190, 214–251.