HYDROLOGICAL PROCESSES
Hydrol. Process. (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/hyp.6668

Implementing quality control on a random number stream to improve a stochastic weather generator[†,‡]

Charles R. Meyer,[1§] Chris S. Renschler[2*] and Roel C. Vining[3]

[1] *National Soil Erosion Research Laboratory, USDA-ARS, West Lafayette, IN 47907, USA*
[2] *Department of Geography/NCGIA, University at Buffalo (SUNY), Buffalo, NY 14261-0023, USA*
[3] *National Air Quality Team, USDA-NRCS, Portland, OR 97232, USA*

Abstract:

For decades, stochastic modellers have used computerized random number generators to produce random numeric sequences fitting a specified statistical distribution. Unfortunately, none of the random number generators we tested satisfactorily produced the target distribution. The result is generated distributions whose mean even diverges from the mean used to generate them, regardless of the length of run. Non-uniform distributions from short sequences of random numbers are a major problem in stochastic climate generation, because truly uniform distributions are required to produce the intended climate parameter distributions. In order to ensure generation of a representative climate with the stochastic weather generator CLIGEN within a 30-year run, we tested the climate output resulting from various random number generators. The resulting distributions of climate parameters showed significant departures from the target distributions in all cases. We traced this failure back to the uniform random number generators themselves. This paper proposes a quality control approach to select only those numbers that conform to the expected distribution being retained for subsequent use. The approach is based on goodness-of-fit analysis applied to the random numbers generated. Normally distributed deviates are further tested with confidence interval tests on their means and standard deviations. The positive effect of the new approach on the climate characteristics generated and the subsequent deterministic process-based hydrology and soil erosion modelling are illustrated for four climatologically diverse sites. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS climate; stochastic model; quality control; random number; weather; weather generator; runoff; soil erosion; deterministic model

Received 10 April 2006; Accepted 15 December 2006

INTRODUCTION

Random numbers in climate generation

For over 40 years stochastic modellers have been using computerized random number generators (RNGs), expecting random numeric sequences that fit the specified statistical distribution. Unfortunately, their trust seems to be misplaced. None of the many RNGs we tested produced the target distribution satisfactorily. This results in distributions whose means do not even match the mean they were generated from, regardless of the number of iterations. In stochastic modelling, two attributes of random number generation are of interest: (a) the 'randomness' of the number stream; (b) its faithful reproduction of the expected statistical distribution. The first characteristic is employed to mimic the apparently random occurrence of natural processes. The second is needed for model equations to perform predictably as designed. This stochastic process begins with a uniform RNG, which, given enough iterations, is theoretically supposed to produce a uniform distribution of random numbers. In this paper we do not examine the quality of the 'randomness' as addressed by Yu (2002); instead, we investigate whether the numbers generated fit the expected distributions, specifically whether a standard normal distribution results from our stochastic process that starts with a uniform RNG, followed by a standard normal deviate generator (SNG). Further, in this paper, comparisons are made between climatic outputs of the model and historically derived climatic inputs to the model (monthly mean, standard deviation (SD), skew), not between model outputs and daily historical observations per se. In other words, we are comparing what climate came out of the model with what climate went in to produce it.

The purposes of this study were:

1. To examine whether it is reasonable to expect the monthly means and SDs of our weather generator's daily output to converge on the monthly means and SDs used to generate them within a typical 30-year simulation run.
2. If the monthly means and SDs are not reproduced, to implement mechanical filtering to ensure this.

[*] Correspondence to: Chris S. Renschler, Department of Geography/National Center for Geographic Information and Analysis (NCGIA), University at Buffalo—The State University of New York (SUNY), 105 Wilkeson Quad, Buffalo, NY 14261-0023, USA.
E-mail: rensch@buffalo.edu

[†] The contributions of Charles R. Meyer and Roel C. Vining to this article were prepared as part of their official duties as United States Federal Government employees.

[‡] The method and source code used in this research is publicly available on the CLIGEN website of the National Soil Erosion Research Laboratory (http://topsoil.nserl.purdue.edu/nserlweb/weppmain/cligen/).

[§] Charles "Chuck" Meyer passed away on 4 February 2007.

3. To measure how well we can generate the actual distributions of daily temperature and precipitation expected.
4. To gauge the effects of these changes on a deterministic process-based erosion model driven by the weather generator outputs.

The Water Erosion Prediction Project (WEPP) model (Flanagan and Nearing, 1995) and its Geospatial Interface (GeoWEPP; Renschler, 2003) are process-based hillslope and watershed models. Pruski and Nearing (2002) performed a climate-change-motivated sensitivity analysis of WEPP by varying the annual precipitation by either changing the number of wet days per year, the amount and intensity of daily rainfall, or combinations of either. Their analysis showed that on an average based on three locations with a diverse climate, each 1% change in average annual precipitation induced 1·28–2·50% change in runoff and 0·85–2·38% change in soil loss. We used the CLIGEN stochastic weather generator (Nicks *et al.*, 1995) to test modifications on the RNG and their effect on the distribution of daily temperature and precipitation and we used the deterministic WEPP to monitor the impact on simulated runoff and sediment discharges. Two versions of CLIGEN are used: version 4.x, which does not utilize the quality control (QC) method, and version 5.x, which uses QC on its RNG and SNG.

WEPP users typically like to make model runs of 30 years, or at most 100 years to account for the stochastic nature of the climate input; however, Baffaut *et al.* (1996) found when using CLIGEN V-4·2 to drive WEPP that 'to obtain a stable running average of the annual soil loss, 30 years of simulation was not enough'. In most cases the minimum simulation period varied between 50 and 100 years, with some locations requiring even more than 100 years (Baffaut *et al.*, 1996). It is our expectation that if the stochastic CLIGEN V-5.x converges to its input values more quickly, then the deterministic WEPP modelling will converge more quickly too.

CLIGEN produces time-series of daily climate parameters from static monthly values derived from daily values observed at the site for some period of record. These values include monthly mean, SD, and skewness. This approach permits generation of representative weather patterns for user-selectable time intervals, using a relatively small amount of input data. The quality of results produced by CLIGEN and other stochastic models, like GEM (Johnson *et al.*, 1996), WINDGEN (Wagner, 1999), USCLIMATE (Hanson *et al.*, 1994), WGEN (Richardson and Wright, 1984), and also models that use their outputs, like SWAT (Arnold *et al.*, 1995), SWRRB (Arnold and Williams, 1994), GLEAMS (Knisel, 1993), EPIC (Sharpley and Williams, 1990), WEPS (Hagen, 1991), and CREAMS (Knisel, 1980), depends directly upon the quality of the distributions produced by the RNGs. The issues discussed in this paper are part of a much bigger problem, the propagation of errors though a system of data processing in an environmental model.

Renschler (2003) suggests a theory in this regard that summarizes the paradigm of appropriately applying an environmental process model and offers guidance to a solution: it requires the careful consideration of all steps involved in integrating observed data, processing/generation of model input parameters, modelling, and decision-making based on these model results. He describes that each step in such a 'scaling sequence' must be assessed in terms of how data are being transformed (scaled). The most basic scaling step is represented by the transformation of a true pattern of a natural process to a representative pattern described in the measured data (Blöschl, 1999). The RNG of a weather generator is part of such a scaling sequence to generate natural patterns of various climate variables that are potentially model input to other environmental process models. In this paper we will discuss the steps that describe the effect of the RNG for CLIGEN on predicting soil loss with the WEPP model (Figure 1).

CLIGEN generates eight parameter distributions on a daily time step. For normal distributions, like temperature, this is accomplished by feeding the output from an RNG (Figure 1, step 1) into an SNG (Figure 1, step 2)

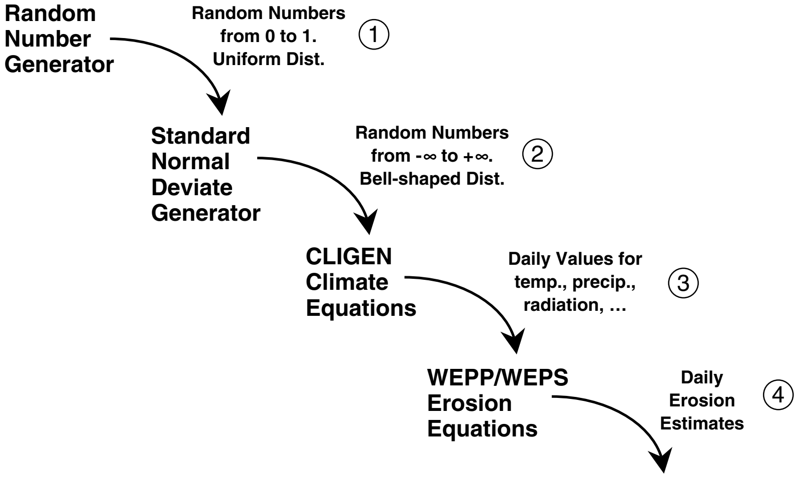
Random Number Generator
Random Numbers from 0 to 1. Uniform Dist.
1
Standard Normal Deviate Generator
Random Numbers from -∞ to +∞. Bell-shaped Dist.
2
CLIGEN Climate Equations
Daily Values for temp., precip., radiation, ...
3
WEPP/WEPS Erosion Equations
Daily Erosion Estimates
4

Figure 1. Overview of the four computational steps to process information from the generation of random numbers to generate climate parameters with CLIGEN and using them to predict runoff and soil erosion with WEPP (compare steps with Renschler (2003: figure 5))

DOI: 10.1002/hyp

to produce sets of 'normal deviates' that approximate the standard normal distribution. The values of each standard normal distribution are then scaled by the corresponding observed historical monthly mean and SD to produce daily climate values (Figure 1, step 3). Thus, the quality of the daily climate parameter distribution produced depends directly upon the quality of the distributions produced by the RNG. Lack of quality assurance for these distributions has potentially serious implications for CLIGEN and for simulation models depending on it (Figure 1, step 4). As mentioned in Press *et al.* (1992), 'a reliable source of random uniform deviates is an essential building block for any sort of stochastic modeling'.

The stochastic weather generator CLIGEN

The equations used in CLIGEN to generate precipitation and maximum temperature are briefly summarized below. All the CLIGEN equations are documented in detail in the WEPP model documentation (Nicks *et al.*, 1995: chapter 2). Generating a uniform random variable, in combination with a two-state Markov chain, using the observed probability for the month of a wet day following a wet day $P(W|W)$, and a wet day following a dry day $P(W|D)$ determines whether precipitation should be generated for the current day (Nicks and Harp, 1980). A skewed normal Pearson Type III equation (Equation (1)), provides the daily precipitation amount P:

$$x = \frac{6}{g}\left\{\left[\frac{g}{2}\left(\frac{P-u}{s}\right)+1\right]^{1/3}-1\right\}+\frac{g}{6} \qquad (1)$$

where x is the generated standard normal deviate (Figure 1, step 2), u is the observed monthly mean of the daily values when precipitation occurs, s is the observed monthly SD, and g is the skew of the monthly precipitation. Daily maximum temperature $T_{\max}$ is generated using a normal distribution:

$$T_{\max} = u_{T_{\max}} + s_{T_{\max}}x \qquad (2)$$

where u is the observed monthly mean and s is the observed SD, for the current month. Both state-of-the-art process-based erosion models in the USA, i.e. the water erosion model WEPP and the wind erosion model WEPS (Hagen, 1991), rely on CLIGEN to generate their climate input (Figure 1, step 3).

CLIGEN performance

Johnson *et al.* (1996) observed that whereas CLIGEN reproduced historical long-term average annual values and extreme values for the simulation period reasonably well, it poorly reproduced year-to-year variance in average annual temperature values and it did a dismal job of reproducing the monthly SDs, failing all 72 tests performed. Our goodness-of-fit tests on CLIGEN outputs indicated at high levels of probability that CLIGEN V-4·2 monthly output distributions often did not match the target distributions. Based on limited 20-year records of precipitation data at two sites in Uganda, Elliot and Arnold (2001) found significant differences between the observed SDs of monthly precipitation and those predicted by CLIGEN. They noted this can be a problem for erosion prediction models, because it is the occurrence of major events that generally causes most of the larger, more drastic soil losses.

If CLIGEN's problems were due primarily to its climate equations, then the long-term historic means probably would not be successfully reproduced for the climatic variables; however, since the problem seems to involve the *distribution* of the outputs relative to the mean, it seems more likely that the distribution of random numbers fed to the equations is suspect.

In this paper we propose to show that there is a problem with the numeric distributions produced by common RNGs in the way they are typically employed: that at least for small lots of numbers appropriate to a 30-year simulation the 'uniform' distributions are not uniform. We show that subsequent calculations are adversely and unacceptably affected. This is likely a problem in any model that employs an RNG. Our method of QC may be applied to the stream of numbers from an RNG in any stochastic model to correct this deficiency. We show comparisons between results of the quality-controlled and the uncontrolled processes at each step.

METHODS AND MATERIALS

Chi-square goodness-of-fit test

The chi-square goodness-of-fit test can be employed to determine the probability that one distribution differs from another. The general approach is that the probability space is divided into an arbitrary number of ranges (bins) of arbitrary size, and the number of observations in a bin is compared with the number expected. The method is reputed to be somewhat sensitive to the number of bins used. In practice we found it to be *highly* sensitive to the number of bins, enough to render it useless for our purposes (Table I).

Kolmogorov–Smirnov goodness-of-fit test

The Kolmogorov–Smirnov (K–S) test is another goodness-of-fit test that compares the cumulative numbers in the bins. For continuous distributions, a very few authors suggest K–S may be more suitable than chi-square. Our tests verified the K–S test to be far less sensitive to bin size (Table II). Consequently, it is also to be used to measure goodness of fit when comparing CLIGEN distributions of daily outputs with the monthly parameters from which they were generated. It is also used for QC of uniform distributions in the new version of CLIGEN (5.x) reported in this paper. It might seem that using the same statistical test for QC in the model and then again to measure outputs of the model is bound to give favourable results; however, this was not the case when we initially used the chi-square test for both purposes.

DOI: 10.1002/hyp

Table I. Chi-square goodness-of-fit test on maximum temperature with varying number of equal-sized bins, showing probability that the month's set of daily values produced is *not* the target distribution. The 30-year run with CLIGEN V-4·2 (no QC) for Indianapolis, IN[a]

Bin no.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
20	—	0·55	—	**0·79**	—	0·68	—	—	0·57	**0·90**	—	—
19	—	0·62	—	**0·76**	—	0·65	***0·93***	0·69	0·53	>***0·99***	—	0·64
18	**0·81**	**0·84**	—	***0·93***	—	**0·78**	—	0·70	**0·76**	>***0·99***	—	0·69
17	0·64	**0·76**	—	—	—	**0·87**	0·73	—	—	0·55	—	—
16	0·51	0·74	—	0·73	—	***0·99***	—	—	—	***0·96***	—	—
15	—	0·69	—	—	—	0·54	—	—	**0·83**	***0·92***	—	—
14	—	0·57	—	**0·81**	—	**0·89**	***0·98***	0·57	**0·80**	—	—	0·65
13	**0·79**	—	—	—	—	**0·78**	0·52	**0·83**	—	***0·96***	—	0·67
12	—	**0·77**	—	—	0·62	0·54	0·71	**0·88**	0·71	**0·88**	—	0·65
11	—	—	—	0·70	—	—	—	**0·86**	0·51	**0·88**	—	—
10	—	—	—	—	—	0·50	—	0·61	0·59	***0·95***	—	0·51

[a] Probabilities below 50% are indicated with a dash, those above 75% appear in bold, and those above 90% are bold italic.

Table II. K–S goodness-of-fit test on maximum temperature with varying number of equal-sized bins, showing probability that the month's set of daily values produced is *not* the target distribution. The 30-year run with CLIGEN V-4·2 (no QC), for Indianapolis, IN[a]

Bin no.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
20	—	—	—	—	—	—	—	—	0·65	**0·80**	—	—
19	—	—	—	—	—	—	—	—	**0·80**	***0·98***	—	—
18	—	—	—	—	—	—	—	—	**0·80**	**0·80**	—	—
17	—	—	—	—	—	—	—	—	0·70	**0·85**	—	—
16	—	—	—	—	—	—	—	—	0·70	**0·80**	—	—
15	—	—	—	—	—	—	—	—	**0·80**	***0·90***	—	—
14	—	—	—	0·50	—	—	0·50	—	0·70	**0·80**	—	—
13	—	—	—	—	—	—	—	—	0·70	***0·90***	—	—
12	—	—	—	—	0·65	—	0·55	—	**0·80**	**0·80**	—	—
11	—	—	—	0·50	—	—	—	—	0·50	**0·75**	—	—
10	—	—	—	—	—	—	—	—	0·65	**0·80**	—	—

[a] Probabilities below 50% are indicated with a dash, those above 75% appear in bold, and those above 90% are bold italic.

Confidence interval tests

The central limit theorem states that sample *means* approach a *normal distribution* as the number of samples becomes large, regardless of the underlying distribution (Ross, 1993). Statisticians generally agree that the number of samples required for a test is around 20 for symmetric distributions that resemble a bell-shaped curve, and around 30 for others. This powerful theorem justifies confidence interval (CI) testing on means, and even on SDs if the underlying distribution is known. We employ CI tests in two different ways: one after a run and one during a run.

To determine whether CLIGEN was generating the expected means of daily climate values, we performed CI tests on CLIGEN's daily climate outputs. Since CLIGEN generates its outputs from monthly parameters, the appropriate comparison is the mean of CLIGEN's daily outputs for each month with the historical monthly means using the historical monthly SDs from which the daily outputs were generated.

Implementing the quality control and correction method

To accomplish the automated random number QC in CLIGEN V-5.x, code was added to V-4.x to perform K–S testing on all uniform distributions generated. Then CI tests are performed for all normal distributions generated (Figure 1, step 2), testing both the means and SDs (Meyer, 2006). For the results reported in this paper, the only difference in CLIGEN V-4.x and V-5.x is the absence of this QC in the former and its presence in the latter. In CLIGEN V-5.x, deviates are generated for each parameter a month at a time and tested. Because our goal is simply to achieve the desired distribution at the end of the run, the numbers for each parameter are examined each month, as a total set from the beginning of the simulation through the current month. If they fail QC, then a new set for the current month is generated and tested. In effect, this permits relaxing the constraints as the run progresses so that we preclude fewer extreme events than if we controlled each month independently. We arbitrarily selected a probability threshold of 50%.

The appropriate CI test statistic for the mean is

$$\frac{\overline{X}-\mu}{\delta/\sqrt{N}} \approx N(0,1) \tag{3}$$

where N is the number of samples, $\overline{X}$ is the sample mean, μ is the population mean (for standard normal, $\mu = 0$);

σ is the population SD (for standard normal, $\sigma = 1$), and N(0,1) is distributed normally (mean: 0; SD: 1).

Since the distribution produced by the SNG is supposed to be standard normal, the SD is also easy to test. The statistical test for the SD is

$$\sum_i \left(\frac{X_i - \mu}{\delta}\right)^2 \approx X^2(N) \qquad (4)$$

where X_i is the value of the i-th sample and $X^2(N)$ is distributed chi-square, with N degrees of freedom. But for a standard normal population this reduces to

$$\sum_i ({X_i}^2) \approx X^2(N) \qquad (5)$$

Graphical convergence tests

We used a check by graphical inspection to judge whether the distribution from each RNG seemed to be converging, how fast, and whether it appeared to be converging to the expected value (Figure 2). For a uniform distribution where x_i is the random number generated and 0·5 is the expected value of x, the *mean* value of $x_i - 0{\cdot}5$ should converge to zero.

Because the expected value of a sum is the sum of the expected values, it follows from Equations (6) and (7) that the *sum* on the left-hand side of Equation (8) should also converge to zero:

$$E\sum_{i=1}^{N}(X_i - 0{\cdot}5) = \sum_{i=1}^{N} E(X_i - 0{\cdot}5) \qquad (6)$$

$$E(X_i - 0{\cdot}5) = 0 \qquad (7)$$

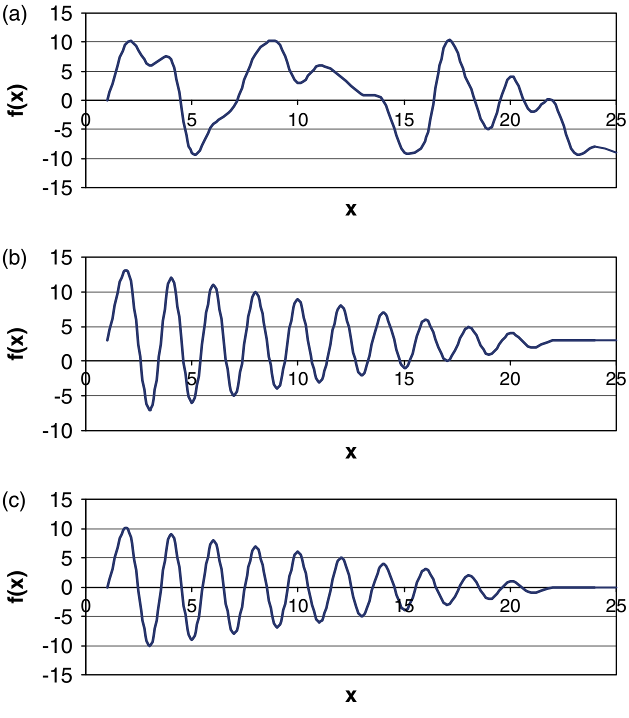

Figure 2. Example of three series: (a) one that does not converge; (b) one that converges, but on the wrong value; (c) one that converges on the value expected

$$E\sum_{i=1}^{N}(X_i - 0{\cdot}5) = 0 \qquad (8)$$

RESULTS

CLIGEN without quality control (version 4.X)

For our initial tests we used CLIGEN V-4.x, which exerts no QC on its random deviates. Because we observed that, for runs up to 1000 years, CLIGEN always seems to converge on its final mean value within 28 years (possibly due to its internal level of numeric precision), we selected 30-year runs as a good compromise between minimum runtime and reaching steady state. The parameters of interest were daily maximum temperature and daily precipitation amount. We used a spreadsheet to compute the probability that the daily outputs were from a population different than the historically observed one represented by the input monthly mean and SD (and skewness in the case of precipitation). (This corresponds to Figure 1, step 3.) Maximum temperature is mathematically simple to test, since CLIGEN uses a normal distribution to produce it. Precipitation is somewhat more complex, since a skewed normal (Pearson Type III) is used. For this reason, initial adequacy checks of the chi-square and K–S tests were performed using daily maximum temperature.

In goodness-of-fit tests, the choice of the number and size of bins is arbitrary: the bins merely must cover the entire probability space. (The chi-square test additionally requires a minimum expected frequency of five per bin.) To simplify our task we chose to make our bin size uniform, having the same number of expected observations. Since there are several goodness-of-fit tests that we might employ, it was necessary to identify one that gave consistent results, regardless of the number of bins used. We ran the chi-square goodness-of-fit test varying the number of bins from 10 to 20. It proved so inconsistent that we deemed it useless for judging whether the desired distribution was produced (Table I). The K–S test was much more consistent. The results shown in Table II for Indianapolis, Indiana, are typical of all locations tested. We concluded that the RNG in CLIGEN produced unacceptable results.

Bringing the generator process back into control

In an industrial setting, when a production line goes out of control one might try to improve the quality of the units produced by improving the *process* that produced them, i.e. replace the RNG and/or SDG in CLIGEN with better ones. We examined a number of RNGs (see Meyer (2006) for RNG names and source code) and concentrated further testing on the most promising ones. Because our users require a reproducible sequence of numbers, RNGs based on the system clock were not considered. Several of the alternate RNGs yielded measurable improvements; however, none consistently produced results that demonstrated the process was under control.

When it is not feasible to improve the quality of the units produced, industry commonly resorts to inspection of the units coming off the production line, and rejection of those not within specifications. Of course, more stringent specifications result in a higher rejection rate, and specifications that are unreasonably rigid can conceivably curtail the acceptance rate to zero. The confidence interval approach is commonly used in manufacturing to ensure that goods meet the desired production quality standards. If CLIGEN's RNG and SNG together are thought of as a "factory" producing random numbers with standard normal distribution for the CLIGEN climate equations to use, the quality of the distribution can be both measured and controlled using K-S testing on uniform distributions, and confidence interval testing of the mean and the SD of normal distributions.

CLIGEN with quality control (version 5.x)

Automated K–S and CI code to measure and control the random deviates was added to CLIGEN V-4.x to create CLIGEN V-5.x. We arbitrarily used a confidence threshold of 50% for each stochastically generated parameter, rejecting monthly lots of uniform deviates if the K–S indicated more than a 50% probability of non-uniformity, and subsequently rejecting standard normal deviates which, when combined with those already accepted, would give either a mean or SD outside the 50% confidence limits.

Table II clearly shows that the daily maximum temperature outputs from CLIGEN are failing to reproduce the original distribution of daily values for October, and there is a high probability they are failing for September as well. However, Table III shows that controlling the uniform deviates with K–S and the means and SDs of the standard normal deviates with CI is sufficient to bring the daily values of the temperature distribution in line with the original historically observed distribution, with a maximum 65% probability of difference, regardless of bin size used for the test.

We chose four climatically diverse sites, listed in Table IV. For these sites, Table V summarizes CI tests comparing the monthly means of precipitation (both with and without QC) with the historically observed means. The original version without QC was outside the 95% confidence limits 7 of 48 times (15%), and outside the 50% confidence limits 25 of 48 times (52%). By comparison, the version with QC was outside the 95% confidence limits 0 of 48 times and outside the 50% confidence limits only twice (4%).

For the same stations, average error in annual mean precipitation for the version with QC was 8·33 mm, and without QC it was 37·82 mm (4·5×). Average error in monthly mean precipitation for the version with QC was 2·97 mm, and without QC it was 7·04 mm (2·4×); and average error in SD by station was respectively 2·36 mm and 6·05 mm (2·6×). The analogous error in precipitation per event was 0·20 mm and 0·74 mm (3·6×), with SD of 0·18 mm and 0·61 mm (3·9×) respectively. Average error in monthly event number was 0·357 and 0·478

Table III. K–S goodness-of-fit test on maximum temperature, with varying number of equal-sized bins, showing probability that the month's set of daily values produced is *not* the target distribution. The 30-year run with CLIGEN V-5·2255 (with QC), for Indianapolis, IN[a]

Bin no.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
20	—	—	—	—	—	—	—	—	—	—	—	—
19	—	—	—	—	—	—	—	—	—	—	—	—
18	—	—	—	—	—	—	—	—	—	—	—	—
17	—	—	—	—	—	—	—	—	—	—	—	—
16	—	—	—	—	—	—	—	—	—	—	—	—
15	—	—	—	—	—	—	0·55	—	—	—	—	—
14	—	—	—	0·50	—	—	—	—	—	—	—	—
13	—	—	—	—	—	—	0·60	0·60	—	—	—	—
12	—	—	—	—	0·65	—	0·50	—	—	—	—	—
11	—	—	—	0·50	—	—	—	—	—	—	—	—
10	—	—	—	—	—	—	—	—	—	—	—	—

[a] Probabilities below 50% are indicated with a dash, those above 75% appear in bold, and those above 90% are bold italic.

Table IV. Climate classification for four different test sites used

Station, state	Köppen climate classification[a]	Annual temperature distribution	Annual precipitation distribution	Avg. annual precipitation (mm)
Indianapolis, IN	Cfb	Warm summer	Uniform	1013·1
College Station, TX	Cfa	Hot summer	Uniform	959·5
Moscow, ID	Dsb	Warm summer	Dry season	621·5
Tucson, AZ	Csa	Hot summer	Dry season	293·2

[a] Griffiths and Driscoll (1982).

Table V. CI test on precipitation per event, comparing the generated monthly mean with the historically *observed values*. The 30-year run with CLIGEN V-4·2 (no QC) and V-5·2255 (with K–S QC)[a]

Station	Source	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Indianapolis, IN	Obs.	*0·23*	*0·24*	*0·27*	*0·29*	*0·34*	*0·40*	*0·45*	*0·39*	*0·36*	*0·32*	*0·33*	*0·26*
	V-4·2	**0·26**	0·24	**0·29**	**0·33**	0·31	***0·48***	**0·42**	**0·35**	***0·43***	0·32	0·35	0·26
	V-5.x	0·24	**0·26**	0·28	0·28	0·34	0·40	0·44	0·38	0·35	0·34	0·34	0·26
College Station, TX	Obs.	*0·30*	*0·38*	*0·33*	*0·53*	*0·61*	*0·57*	*0·43*	*0·42*	*0·60*	*0·60*	*0·42*	*0·37*
	V-4·2	**0·33**	***0·46***	**0·29**	0·54	0·63	0·60	0·46	0·42	0·61	**0·69**	**0·46**	0·39
	V-5.x	0·30	0·38	0·33	0·53	0·62	0·60	0·43	0·41	0·62	0·61	0·44	0·37
Moscow, ID	Obs.	*0·20*	*0·19*	*0·16*	*0·18*	*0·21*	*0·21*	*0·20*	*0·22*	*0·19*	*0·21*	*0·21*	*0·20*
	V-4·2	**0·22**	0·20	0·17	0·18	***0·25***	**0·19**	0·19	0·22	0·20	**0·19**	0·22	***0·24***
	V-5.x	0·21	0·18	0·16	0·18	0·21	0·20	0·21	0·23	0·18	0·20	0·21	0·20
Tucson, AZ	Obs.	*0·19*	*0·19*	*0·17*	*0·15*	*0·11*	*0·13*	*0·22*	*0·25*	*0·30*	*0·29*	*0·20*	*0·21*
	V-4·2	0·20	**0·22**	0·17	**0·17**	0·10	***0·23***	**0·20**	**0·29**	**0·35**	0·28	***0·11***	**0·17**
	V-5.x	0·20	0·18	0·17	0·15	0·10	0·13	**0·20**	0·26	0·28	0·28	0·20	0·22

[a] Probabilities outside the 50% CI appear in bold, and those outside the 95% CI are also in bold italic.

respectively. The results were consistently more uniform with QC; but, because underestimates from individual months can cancel out overestimates when comparing events per year, the average error in events per year was virtually identical for the two versions at 0·7.

Table VI shows a K–S comparison of the daily precipitation distributions at the same four stations with the distribution derived from the historical parameters. It should be noted that, for those months when the quality-controlled version of CLIGEN failed to achieve a good match to the target distribution, the skew values were high (3·4 to 4·1), and we have observed that the skewed normal equation used in CLIGEN performs poorly when the skew exceeds values in the range 2·25–2·50. So, this failure is probably due to lack of robustness in the CLIGEN precipitation equation.

We ran convergence checks for four RNGs: RANDN from CLIGEN (Nicks *et al.*, 1995), RAN2 from Press *et al.* (1992), RAN3 from GEM (Johnson *et al.*, 1996), and RANDN with QC. Figure 3 illustrates to what extent each uncontrolled RNG tested converges to its expected mean value in a 100-year run, and the extent to which our QC improved the situation for the RNG employed in CLIGEN.

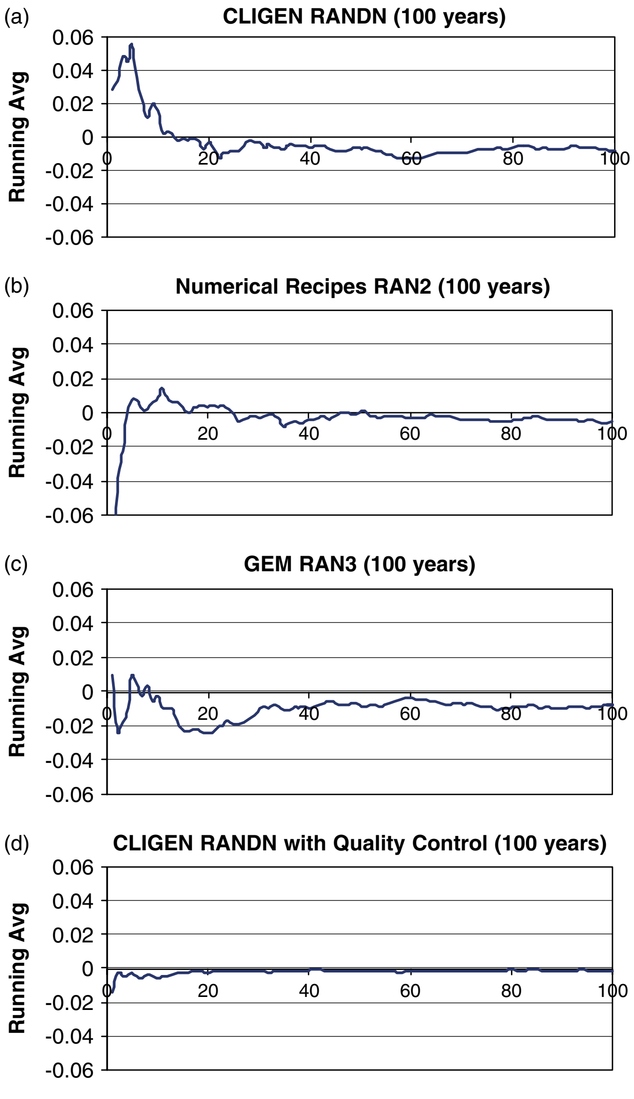

Figure 3. Running means of uniform deviates (minus 0·5, the expected value) for daily values for a single parameter and month from three RNGs without QC (a–c) and one with QC (d), showing their respective convergence upon the proper value (zero)

The running sum convergence check was employed to examine the serial independence of the three RNGs: RANDN, RAN2, and RAN3 (Figure 4). (Serial independence implies that a user can draw a subset of numbers from anywhere in the stream without affecting the distribution of the numbers sampled.) Approximately 3 650 000 iterations, equivalent to a 10 000-year run of CLIGEN, showed each of them to cycle, first showing strong bias in one direction and then in the other. This shows that none of the RNGs tested exhibited serial independence.

CLIGEN quality impact on deterministic soil erosion modelling

The climate data generated were subsequently used to compare results from the WEPP soil erosion model (see also Figure 1, step 4.) We chose the four climatically diverse sites listed in Table IV. In Table VII we compare the annual means of their daily precipitation values with the average annual precipitation values from the historic

DOI: 10.1002/hyp

Table VI. K–S test on precipitation per event, comparing the generated monthly distribution with distribution derived from the historically observed parameters. Using 10 equal-sized bins, showing probability that the month's set of daily values produced is *not* the target distribution. The 30-year run with CLIGEN V-4·2 (no QC) and V-5·2255 (with K–S QC)[a]

Station	Source	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Indianapolis, IN	V-4·2	0·55	—	—	***0·90***	—	0·70	—	0·65	**0·85**	**0·80**	—	0·65
	V-5.x	—	—	—	—	—	—	—	—	—	—	—	—
College Station, TX	V-4·2	—	—	***0·95***	—	—	—	—	—	0·65	—	—	0·70
	V-5.x	—	—	**0·80**	—	—	—	—	—	—	—	—	0·50
Moscow, ID	V-4·2	**0·85**	—	—	—	0·55	0·50	0·50	**0·80**	—	—	—	—
	V-5.x	—	—	—	—	—	—	—	—	—	—	—	—
Tucson, AZ	V-4·2	—	0·60	—	—	—	***0·98***	>***0·99***	0·65	**0·80**	—	—	**0·85**
	V-5.x	—	—	—	—	—	0·70	>***0·99***	**0·85**	—	—	—	—

[a] Probabilities below 50% are indicated with a dash, those above 75% appear in bold, and those above 90% are bold italic.

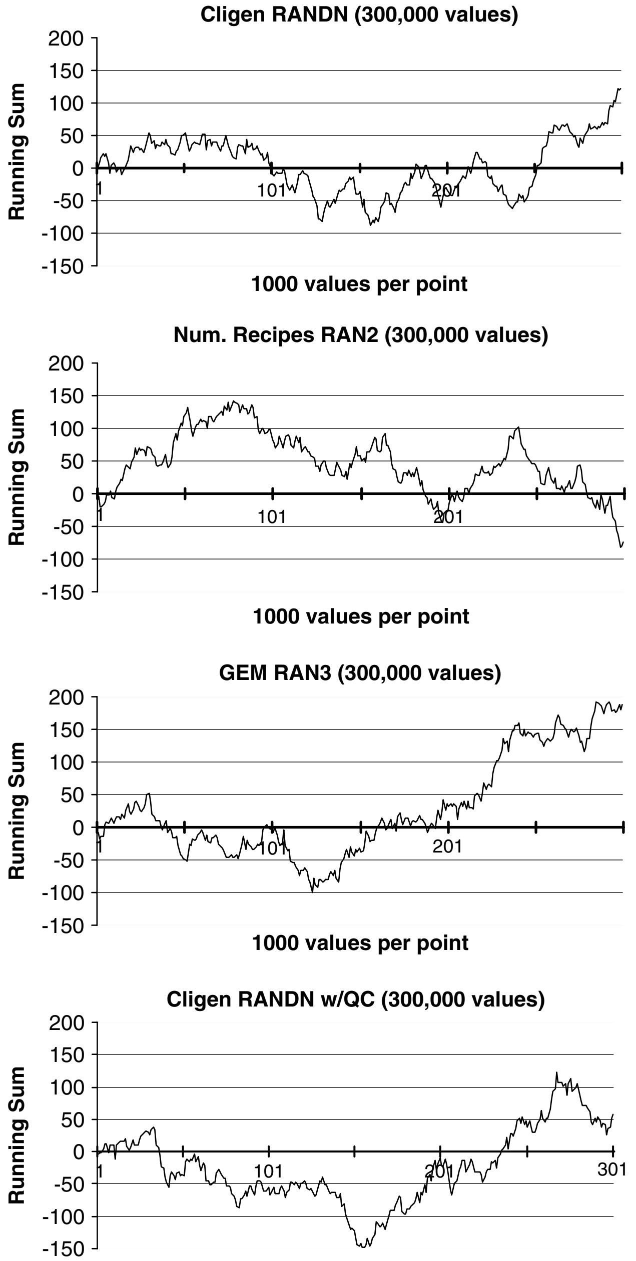

Figure 4. Running sums of uniform deviates (minus 0·5, the expected value) for daily values for a single parameter and month from three RNGs showing lack of convergence after 300 000 iterations. Note that this illustrates their lack of serial independence when used in this manner

record at each site. Since CLIGEN V-4·2 and V-5.x used the historic monthly means to compute their daily values, the computed average annual value should converge back to the historically measured value. The same hillslope topography, soils and land management were used for all four locations.

Since CLIGEN's output reports the historic monthly means for the site, it is easy to determine whether the version with QC is reproducing the historical recorded mean annual precipitation better than the original version. The results of 30-year CLIGEN runs summarized in Table III show that, when CLIGEN V-4·2 differed significantly from the QC version CLIGEN V-5.x, the QC was closer to the observed average annual precipitation. Also note the *similarity* of average annual precipitation and runoff amounts for Moscow, ID (MOS), and the *difference* between the amounts of average annual sediment yield predicted by the WEPP model. This underscores the need to reproduce the historical *distributions* accurately, not just the *total amounts*. A more detailed evaluation of the ability of the CLIGEN model to reproduce daily, monthly, and annual precipitation amounts, extremes, and internal weather or storm patterns (i.e. storm duration, relative peak intensity, and time to peak) and the assessment of the impact of generated storm patterns on WEPP runoff and erosion prediction can be found in Zhang and Garbrecht (2003). They found that CLIGEN-generated precipitation durations were generally too long for small storm events and too short for large storm events, causing WEPP prediction errors as high as 35% for average annual runoff and 47% for annual sediment yield at four sites in Oklahoma.

DISCUSSION

We became interested in using QC on our RNG when we observed events occurring in a 100-year run that would not be expected in 10 000 years. We traced the problem back to our RNG. It was generating a few extremely aberrant values that radically altered the distribution mean, the most elemental statistical metric. Initially we sought to replace our RNG with a better one. Testing every RNG we could find, we discovered they all share the same weakness. Our next idea was to filter the output

Table VII. Stochastic CLIGEN-generated average annual precipitation and deterministic simulated runoff and soil loss (WEPP). The 30-year run with CLIGEN V-4·2 (no QC) and V-5·2255 (with K–S QC)[a]

Station	Data source	Avg. precipitation (mm year^{-1})	Avg. runoff (mm year^{-1})	Avg. soil loss (kg m^{-2} year^{-1})
Indianapolis, IN	Observed[b]	*1013·1*	*n.a.*	*n.a.*
	V-4.x	1034·9	***144·5***	10·1
	V-5.x	1011·9	130·7	9·7
College Station, TX	Observed[c]	*959·5*	*n.a.*	*n.a.*
	V-4.x	**1040·4**	***307·6***	***33·6***
	V-5.x	987·7	265·4	28·1
Moscow, ID	Observed[b]	*621·5*	*n.a.*	*n.a.*
	V-4.x	643·0	21·7	***1·9***
	V-5.x	644·4	21·4	1·5
Tucson, AZ	Observed[b]	*293·2*	*n.a.*	*n.a.*
	V-4.x	**310·1**	***25·5***	**4·2**
	V-5.x	291·1	21·2	3·9

[a] Note that relative changes to CLIGEN V-5.x results above 5% appear in bold, and those above 10% are bold italic; n.a.: not available.
[b] Average based on 45 years of observations.
[c] Average based on 42 years of observations.

of the RNG. (Like adding a filter to the kitchen faucet to remove chlorine, water delivered by the municipal supplier, which is for the most part safe and good, but which has a disagreeable taste and odour, becomes much more palatable after removing that relatively small amount of chemical!) We decided to reject a set of numbers if we were more than 50% sure it was 'bad' (P50). Note that using a higher threshold level does not ensure a more realistic retention of extreme values, since the values generated have more freedom to go lower as well as higher. While Table VIII shows higher extreme values at 95% and 90% rejection threshold, the once at 50% are all lower than observed. It is also interesting to note for the values in Table VIII that the K-S goodness of fit statistics do not change one whit across the three P-values used. From this we infer that any mismatch is due to our assumption that a skewed normal represents the natural population, rather than flaws in our random number generation process.

Clearly, the RNGs we tested are not producing the results we require within the modelling time-frame we need. Although it could be true that our interval is too short to provide the desired distribution, it seems more likely due to other problems, since our runs of 3 650 000 iterations (equivalent to 10 000 years) are also unacceptable. To examine whether the cumulative effect of low numeric precision in the computer algorithms of the RNG was biasing the results, the algorithms were changed from single precision to double precision. This did not correct the problem.

The practical effect of the mechanical screening is to ensure the desired distribution mean and SD at the expense of some randomness. The shorter the run, or the

Table VIII. Differences between observed and generated precipitation with variable rejection threshold (18-year breakpoint precipitation record for Birmingham, AL). A 95% rejection threshold for the random number stream means that only monthly lots of random numbers that are 95% sure that they do not match the target distribution (e.g. uniform or normal) will be rejected and, consequently, another monthly lot will be generated instead. K–S tests for all values generated using 10 bins at natural breaks of observed data showed less than 50% probability that the distribution generated was different from the observed distribution for events greater than 2.5 mm

Observed in 18 years	Rejection threshold (%)	Generated in no. of years				
		18	25	36	50	100
No. of events 87·6	50	91·9	85·7	86·5	86·3	87·3
	90	94·3	87·8	87·8	89·1	87·8
	95	87·6	95·4	89·6	88·7	89·8
Mean event precipitation 13·6 mm	50	13·6	13·6	13·5	13·4	13·6
	90	13·6	13·8	13·7	13·8	13·7
	95	13·6	13·7	13·9	13·9	13·8
Maximum event precipitation[a] 162·6 mm	50	149·1	149·1	149·1	178·3	178·3
	90	162·6	120·4	120·4	150·9	150·9
	95	162·6	132·6	132·6	221·3	221·3
Annual precipitation 1189·9 mm	50	1253·8	1163·7	1168·3	1160·5	1183·0
	90	1305·6	1200·2	1210·4	1216·7	1211·0
	95	1189·9	1309·9	1245·1	1236·1	1241·2

[a] Note: minimum event size was consistently 0·3 mm.

more control exerted by the filter, the more pronounced this trade-off becomes. For our purposes, the desired distribution is more essential than the randomness, so we feel it is an acceptable compromise. There were concerns that filtering might eliminate 'extreme events', which are of the greatest significance in single-event as well as long-term erosion prediction. This does not occur if one runs the simulation sufficiently long enough (see Table XIII). How long is 'long enough'? Operating at a 50% level of probability, i.e. P50, we exceed the largest observed event by running 2·5 times as long as the period of record. With a simulation of the same duration as the historical period of record we get an extreme event within 10% of the maximum historically observed. (P50 means we are allowing only a 50% chance that the distribution generated is not the original distribution.) Using looser control, i.e. P90 or P95, would shorten the run required to duplicate extreme events, but would allow less strict adherence to the mean and SD. This level is easy enough to change. The 'best' probability level depends upon the needs of the user.

One potential problem with this QC approach is that it may skew the results in *time*. Our primary goal was to achieve the desired distribution at the end of the run. The numbers for a given parameter and month of the year are generated and tested in combination with those previously accepted, as each new year is simulated. There are two offsetting factors involved in the CI QC test (the second test applied to normal distributions after the K–S test):

1. When very few numbers are involved, the confidence interval is very wide. There is not enough information to constrain the interval much.
2. As more numbers are added to the distribution tested, the confidence limits get tighter, but adding a really aberrant mean (or SD) does not affect the mean of the overall group very much.

Currently, it is not known how these offsetting factors play out against each other in the early stages of the run. It is conceivable that a divergent number might be rejected early in the run, but accepted later, once the system has enough 'mathematical inertia' not to be grossly affected by it. More investigation is needed to examine this. Of course, if it is determined that the first 5 years of a 30-year run are biased, then one might simply make a 35-year run and discard the first 5 years. However, Figure 4 shows that serial dependence is already a problem in all the uncontrolled RNGs we tested, not just ours with QC.

Because, by definition, two normal distributions are 'equivalent' if they have the same mean and SD, we originally expected that achieving this condition alone would be sufficient to give satisfactory normal distributions. It does seem to be adequate for generating distributions with the correct mean and SD. However, if it is desirable to reproduce the target distribution faithfully over small intervals of the range, then a more stringent QC is needed. Adding the chi-square test to the uniform distribution gave improved but disappointing results. Table I illustrates why: it is inconsistent when bin size is varied. Adding the K–S test on the uniform did a much more satisfactory job, especially where the standard normal distribution is used to generate a skewed normal (Pearson Type III), as in CLIGEN's precipitation outputs. The K–S test using 20 intervals is a much more stringent mathematical test, relying on 20 numbers instead of just two (mean and SD). We now use it to test all uniform distributions generated in the model. For normal distributions we follow the K–S with CI tests on the mean and SD.

SUMMARY AND CONCLUSIONS

We have demonstrated conclusively that, for short-duration runs, the unfiltered RNGs we tested are not producing the uniform distributions we require, or the standard normal distributions we expect, and this problem cascades forward to produce unacceptable results from model equations designed for such distributions, culminating in unacceptable distributions of daily climate outputs. The results of stochastic weather generators, and deterministic hydrology models that depend on them, are sensitive to the quality of their randomly generated distributions. Fortunately, this problem is relatively easy to remedy, as shown in this study, without major changes to the existing program structure using this straightforward QC method. Although we chose 50% for our QC, the level selected is somewhat arbitrary. Users may want to use a higher probability level to apply less stringent control in order to preclude fewer extreme events early in the run. However, the level selected should be sufficient to cause the mean value of the RNG to converge in a reasonable length of time. The number of iterations available before the model run must end and, consequently, the degree of control required to create convergence may vary considerably according to the application. We recommend implementing the QC approach presented rather than searching for, or developing, the perfect RNG. The procedure described can be readily retrofitted into any stochastic model using random number streams.

ACKNOWLEDGEMENTS

We thank Jurgen D. Garbrecht from the USDA–ARS Grazinglands Research Laboratory, El Reno, Oklahoma, for his recommendations and suggestion to include Figure 4.

REFERENCES

Arnold JG, Williams JR. 1994. *SWRRB; a watershed scale model for soil and water resource management.* USDA–ARS Grassland Soil and Water Research Laboratory, Temple, TX.

Arnold JG, Williams JR, Srinivasan R, King KW. 1995. *SWAT: soil and water assessment tool*. USDA–ARS Grassland Soil and Water Research Laboratory, Temple, TX.

DOI: 10.1002/hyp

Baffaut C, Nearing MA, Nicks AD. 1996. Impact of CLIGEN parameters on WEPP-predicted average annual soil loss. *Transactions of the ASAE* **39**: 447–457.

Blöschl G. 1999. Scaling issues in snow hydrology. *Hydrological Processes* **13**: 2149–2175.

Elliot WJ, Arnold CD. 2001. Validation of the weather generator CLIGEN with precipitation data from Uganda. *Transactions of the ASAE* **44**: 53–58.

Flanagan DC, Nearing MA (eds). 1995. *USDA–Water Erosion Prediction Project: hillslope profile and watershed model documentation*. USDA–ARS–NSERL Report No. 10. West Lafayette, IN.

Griffiths JF, Driscoll DM. 1982. *Survey of Climatology*. Charles E. Merrill.

Hagen LJ. 1991. A wind erosion prediction system to meet user needs. *Journal of Soil and Water Conservation* **46**: 106–111.

Hanson CL, Cumming KA, Woolhiser DA, Richardson CW. 1994. *Microcomputer program for daily weather simulation*. USDA–ARS Publication No. ARS-114.

Johnson GL, Hanson CL, Hardegree SP, Ballard EB. 1996. Stochastic weather simulation: overview and analysis of two commonly used models. *Journal of Applied Meteorology* **35**: 1878–1896.

Knisel WG. 1980. *CREAMS, a field scale model for chemicals, runoff and erosion from agricultural management systems*. USDA Conservation Research Report No. 26.

Knisel WG (ed.). 1993. *GLEAMS: groundwater loading effects of agricultural management systems*. USDA–ARS Southeast Watershed Research Laboratory, Tifton, GA.

Meyer CR. 2006. *CLIGEN weather generator, expanded and improved*. http://rizon.nserl.purdue.edu/Cligen/ [9 April 2006].

Nicks AD, Harp JF. 1980. Stochastic generation of temperature and solar-radiation data. *Journal of Hydrology* **48**: 1–17.

Nicks AD, Lane LJ, Gander GA. 1995. Weather generator. In *USDA–Water Erosion Prediction Project: hillslope profile and watershed model documentation*, Flanagan DC, Nearing MA (eds). USDA–ARS National Soil Erosion Research Laboratory; 2·1–2·22.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. *Numerical Recipes in Fortran 77—The Art of Scientific Computing*, 2nd edition, vol. 1. Cambridge University Press. http://lib-www.lanl.gov/numerical/bookfpdf.html [9 April 2006].

Pruski FF, Nearing MA. 2002. Runoff and soil-loss responses to changes in precipitation: a computer simulation study. *Journal of Soil and Water Conservation* **57**: 7–16.

Renschler CS. 2003. Designing geo-spatial interfaces to scale process models: the GeoWEPP approach. *Hydrological Processes* **17**: 1005–1017.

Richardson CW, Wright DA. 1984. *WGEN: a model for generating daily weather variables*. USDA–ARS, ARS-8.

Ross S. 1993. *Introduction to Probability Models*, 5th edition. Academic Press.

Sharpley AN, Williams JR (eds). 1990. *EPIC—Erosion/Productivity Impact Calculator*. USDA–ARS Technical Bulletin 1768, USDA–ARS Grassland, Soil, and Water Research Lab, Temple, TX.

Wagner LE. 1999. *Wind Erosion Prediction System (WEPS): overview*. Wind Erosion: USDA–ARS Wind Erosion Research Unit. An International Workshop. http://www.weru.ksu.edu//proceedings/wagner2.pdf [9 April 2006].

Yu B. 2002. Using CLIGEN to generate RUSLE climate inputs. *Transactions of the ASAE* **45**: 993–1001.

Zhang XC, Garbrecht JD. 2003. Evaluation of CLIGEN precipitation parameters and their implication on WEPP runoff and erosion prediction. *Transactions of the ASAE* **46**: 311–320.

DOI: 10.1002/hyp