Vol. 28: 83–84, 2004

CLIMATE RESEARCH
Clim Res

Published December 31

COMMENT

Comments on the use of statistical tests in the comparison of stochastic weather generators by Qian et al. (2004)

Mikhail A. Semenov*, Sue Welham

Biomathematics & Bioinformatics, Rothamsted Research, Harpenden AL5 2JQ, UK

More and more statistical packages that can be run on a simple PC are becoming available. These packages incorporate a variety of statistical tests and functions (Payne 2003), and by providing a user-friendly interface, they allow scientists without a mathematical or statistical background to perform complex statistical tests by pressing a button. Statistical analysis of data or hypotheses in scientific publications has become a necessity, and a manuscript containing data without statistical analysis is unlikely to be accepted for publication. Hence, there is a temptation for scientists to use well-developed statistical packages without proper knowledge. The danger is that application of complex statistical tests without checking that the conditions for their use are satisfied can lead to flawed conclusions.

In a recent publication, Qian et al. (2004) compared the ability of 2 stochastic weather generators to produce valid distributions of daily and summary weather statistics for various climatic variables at various sites in Canada. They applied several statistical tests, including the Kolmogorov-Smirnov (K-S) test, in their comparison of daily data. On the basis of these test results, they concluded that one of the generators performed better than the other.

Unfortunately, Qian et al. (2004) used some statistical tests inappropriately, so that their conclusions need to be re-assessed. The K-S 2-sample test is unsuitable for comparison of the distributions of (generated and observed) daily temperatures, because the underlying assumptions of the test are not satisfied. The K-S statistic is a test of the null hypothesis that a sample comes from a known distribution, or that 2 populations arise from the same underlying distribution. Let $x_1, \dots x_n$ be observations on continuous independent identically distributed random variables $X_1, \dots X_n$ with a cumulative distribution function (CDF) F. Using these observations, we want to compare F with a known CDF F_0. To test the hypothesis $H_0{:}F = F_0$ Kolmogorov introduced a test statistic D_n, defined as $D_n = \sup_{x \in R} \left| F_n(x) - F_0(x) \right|$, where F_n is an empirical cumulative distribution defined as $F_n(x) = \dfrac{\sum_{i \le n} \chi(x_i \le x)}{n}$ (Kolmogorov 1933). He proved that if H_0 is true, then the distribution for statistics D_n does not depend on F_0 and

$$\lim_{n \to \infty} P\left\{ \sqrt{n} D_n < \lambda \right\} \to K(\lambda) \qquad (1)$$

where $$K(\lambda) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2\lambda^2}$$

In 1948, Smirnov tabulated $K(\lambda)$ (Bolshev & Smirnov 1968). The K-S test has been extended for comparison of 2 empirical cumulative distribution functions (Siegel 1956, Stephens 1970).

Qian et al. (2004) applied the 2-sample K-S test to compare generated and observed daily temperature for selected months. They generated 300 years of synthetic daily weather and compared it with 30 years of observed daily weather. For each month, e.g. January, they compared a generated sample $T^{gen} = \left\{ T^{gen}_{day,year},\ \text{day} = 1,\dots,31,\ \text{year} = 1,\dots,300 \right\}$, where $T^{gen}_{day,year}$ is generated temperature (maximum or minimum) for the day = *day* in January of the year = *year,* with an observed sample $T^{obs} = \{ T^{obs}_{day,year},\ \text{day} = 1,\dots,31,\ \text{year} = 1971,\dots 2000 \}$. They computed the Kolmogorov statistic

$$D_{N_{gen}, N_{obs}} = \sup_{x \in R} \left| F^{gen}_{N_{gen}}(x) - F^{obs}_{N_{obs}}(x) \right|$$

where $N_{gen} = 9300$ and $N_{obs} = 930$, and p values using

$$P\left\{ \sqrt{\frac{N_{gen} N_{obs}}{N_{gen} + N_{obs}}} D_{N_{gen}, N_{obs}} < \lambda \right\}$$

Based on the results of this test they concluded that one generator performed better than the other.

In this case, it is inappropriate to use the 2-sample K-S test to compare the distributions of T^{obs} and T^{gen}, because 2 important assumptions of the test are violated. Values in the sample $T^{obs}_{day,year}$ are not identically distributed, because of the underlying yearly trend in mean temperature, which may differ substantially between the beginning and end of a month. Furthermore, within a year, sample values $T^{obs}_{day,year}$ are not independent, because of strong autocorrelation in temperature values, which is also built into both weather

*Email: mikhail.semenov@bbsrc.ac.uk

Table 1. p values computed using the Kolmogorov-Smirnov 2-sample test for daily maximum temperatures in January in Toronto

	1953	1955	1956	1960
1953	1	0.01	0.01	0.01
1955		1	0.13	0.01
1956			1	0.02
1960				1

generators. For example, the lag-1 autocorrelation for maximum temperature in January in Toronto is 0.58. To illustrate this flaw, we 'compared' daily maximum temperatures in January in Toronto for several years using the K-S 2-sample test available from GenStat 7.2 (Payne 2003). p values from the tests (Table 1) show that for any pair of years, with the exception of {1955, 1956}, we should reject the null hypothesis that maximum January temperatures in Toronto for different years arise from the same distribution. However in this case we know the null hypothesis is true.

Qian et al. (2004) also used a quantile–quantile (Q–Q) plot to compare distributions of daily observed and generated temperature for selected months (see Fig. 4 in Qian et al. 2004). This is not a formal statistical test, but it can provide some insight into differences between distributions. Again, a Q–Q plot assumes independent, identically distributed data, and so is an unsuitable technique to compare samples of correlated data.

As an illustration, we generated 2 sets of 2 samples. In the first set (Fig. 1a), the elements of each sample were independent and came from a normal distribution $N(0,1)$. The Q–Q plot for this set is close to the 1:1 line, confirming that both samples are likely to come from the same distribution. In the second set (Fig. 1b), we generated both samples using a normal distribution, but in this case elements of each sample were generated with a lag-1 correlation of 0.75. In this case the Q–Q plot is systematically lower than the 1:1 line, suggesting systematic differences in the underlying distribution. This happened not because the underlying distributions were different, but simply because the within-sample correlation makes the Q–Q plot unsuitable.

Methods of statistics are a powerful tool in applied research, assisting scientists in making educated guesses and testing hypotheses. However, when applied incorrectly, statistical methods may direct a scientist to flawed conclusions. There is a clear responsibility for producers of statistical packages to provide sufficient information (help) to enable scientists to make informed decisions about the statistical tests they use. Conversely, scientists should also ensure that they are using appropriate tests.

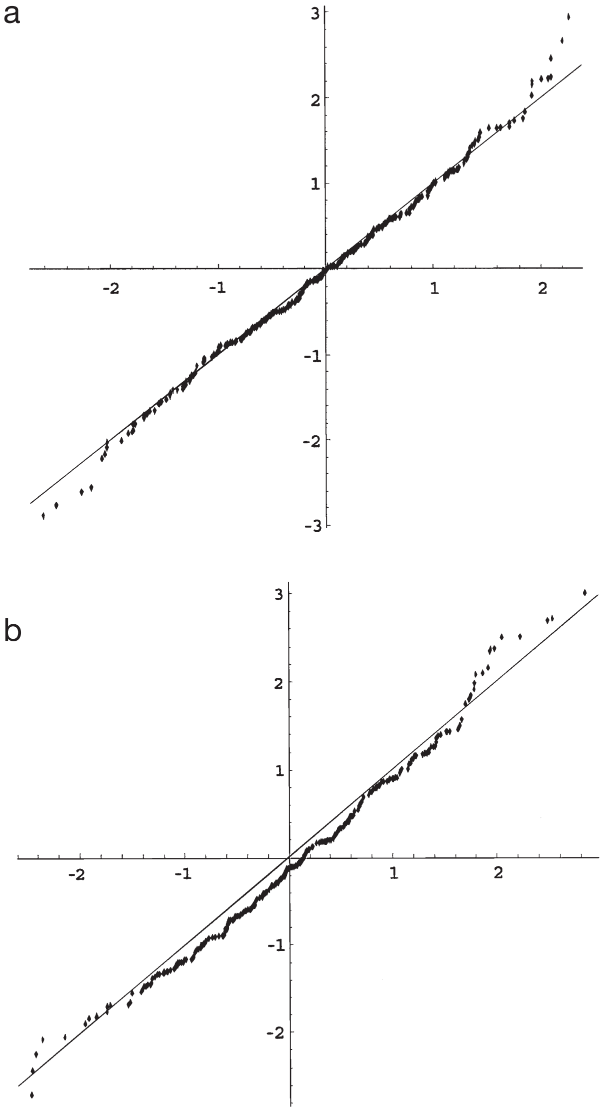

Fig. 1. Q–Q plot of 2 samples with 300 elements each, generated using a normal distribution; (a) each element in each sample generated independently, (b) 2 samples generated with a lag 1 autocorrelation coefficient = 0.75

LITERATURE CITED

Bolshev LN, Smirnov NV (1968) Tables of mathematical statistics. Nauka, Moscow

Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. G Instit Ital Atuari 4:83–91

Payne RW (ed) (2003) The guide to GenStat, Release 7.1 — Part 2: statistics. VSN International, Hemel Hempstead

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

Siegel S (1956) Nonparametric statistics in behavioural sciences. McGraw-Hill, New York

Stephens MA (1970) Use of Kolmogorov, von Mises, and periodogram statistics without extensive tables. J R Stat Soc B 32:115–122

Editorial responsibility: Otto Kinne, Oldendorf/Luhe, Germany

Submitted: November 2, 2004; Accepted: December 2, 2004
Proofs received from author(s): December 10, 2004