Vol. 28: 85–87, 2004 | CLIMATE RESEARCH Clim Res | Published December 31

REPLY COMMENT

Resampling tests — a possible alternative to the standard statistical tests with caution: Reply to Semenov & Welham (2004)

Budong Qian*, Sam Gameda, Henry Hayhoe, Reinder De Jong, Andy Bootsma

Eastern Cereal and Oilseed Research Centre, Agriculture and Agri-Food Canada, 960 Carling Ave., Ottawa K1A 0C6, Canada

Semenov & Welham (2004, this issue) raised a very important point on the use of statistical tests in applied research. Misuse of statistical tests is one aspect of the misuse of statistical analysis in climate research, as demonstrated by von Storch (1995). Proper application of statistical techniques becomes more important as statistical packages that are easy to use on PCs become increasingly available. This is fundamentally important, because almost all data in climate research are intercorrelated in both space and time. These correlations result in difficulties, since most standard statistical techniques are based upon the assumption that the data are derived in independent experiments.

The 2-sample Kolmogorov-Smirnov (K-S) test was applied in our recent paper (Qian et al. 2004) to verify whether the largest difference ($D_{mn} = \max_x |F_m(x_1) - F_n(x_2)|$) between the 2 empirical cumulative distribution functions (CDFs), estimated from the observed weather series (x_{11}, x_{12}, ……, x_{1m}) and from the synthetic series (x_{21}, x_{22}, ……, x_{2n}) generated by stochastic weather generators, is small enough to not reject the null hypothesis that the synthetic series comes from the same probability distribution as the observed series. If the test statistic D_{mn} is larger than the critical value, i.e. $D_{mn} > \left[\frac{1}{2}\left(\frac{1}{m}+\frac{1}{n}\right)\ln\left(\frac{\alpha}{2}\right)\right]^{1/2}$, the null hypothesis is rejected at the $\alpha \times 100\%$ significance level. Although the K-S test is a nonparametric approach, basic assumptions are still applied, as Semenov & Welham (2004) indicated. These assumptions are essential for the test, especially for determining the critical values at the $\alpha \times 100\%$ level by the assumption of independent sampling. When the data involved in the test are serially correlated, the effective sample size or effective number of degrees of freedom is smaller than the data sample size used in the test. Therefore, when the data sample sizes m and n are used to determine the critical value for the test at a given significance level, the critical value is smaller than the one under the independent sampling assumption. This makes the K-S test liberal, i.e. it rejects the null hypothesis more often than expected at a given significance level.

To better estimate the false rejection rate in our tests, a Monte-Carlo experiment was conducted. We took Fredericton as an example, since the rejection rate was relatively high at this station for both weather generators. We performed the K-S test 1000 times on randomly resampled daily temperature data from a 300 yr synthetic dataset generated by the 2 weather generators. The test was conducted for each month, and separately for daily maximum and minimum temperatures and for the 2 weather generators. For each test, two 30 yr independent samples were formed by independently resampling the 300 yr synthetic daily maximum (or minimum) temperature without replacement from the same weather generator. The resampling process was used to select years, rather than days, so that the serial structure would be maintained in the resampled series to be tested. The null hypothesis is true, as the 2 samples in all tests were sampled from the same distribution. The rejection rate (Table 1) varied from 11 to 24% when the 5% level was applied to the 2-sided probability, instead of the single-sided one shown above in the formula for the critical value. No significant difference was observed between weather generators, implying that both weather generators reproduced the serial structure of daily temperatures adequately, and that the relative performance of the 2 weather generators was not affected by the higher rejection rate. We used the 2-sided probability as a compromise to reduce the risk of rejecting the null hypothesis when it is true, as we were aware of the

*Email: qianb@agr.gc.ca

Table 1. Rejection rate (%) in 1000 times performing the Kolmogorov-Smirnov test at the 5% level on 2 independent 30 yr samples resampled from a 300 yr synthetic dataset of daily maximum (T_x) and minimum (T_n) temperatures generated by LARS-WG and AAFC-WG for the station of Fredericton

	LARS-WG		AAFC-WG	
	T_x	T_n	T_x	T_n
Jan	17	22	21	19
Feb	20	15	21	19
Mar	19	23	22	20
Apr	19	18	19	14
May	17	17	18	16
Jun	21	16	18	17
Jul	20	22	19	24
Aug	22	16	19	19
Sep	20	17	15	11
Oct	21	14	18	16
Nov	21	15	21	19
Dec	22	18	21	20

bias that could result from performing the K-S test on data which did not fully satisfy the assumptions of the test. We also assumed that the bias of the tests resulting from autocorrelations might be smaller when a large number of data samples was used, rather than a small number.

In addition, we employed the quantile–quantile (Q–Q) plots to help visually in the assessment of the goodness-of-fit of the 2 samples based on their empirical distributions. As Q–Q plots only compare the empirical distributions of the data samples, we do not think that autocorrelation will have significant effects on the comparison when a large number of data samples is used. To demonstrate this, we generated several sets of 2 random samples from the standard normal distribution with lag-1 correlations of 0.00, 0.25, 0.50 and 0.75. For each set, Sample 1 was generated for 900 elements and Sample 2 for 9000 elements, giving similar sample sizes to those used in Qian et al. (2004). The Q–Q plots (Fig. 1) show that a comparison of the empirical distributions of 2 samples generated from the same probability distribution will indicate a good fit regardless of the magnitude of the autocorrelations between the elements of the samples. We also conducted the same experiment with the sample size of 300 elements used by Semenov & Welham (2004). No difference was observed, except that both tails deviate slightly from the 1:1 line because of the smaller sample size. This implies that a larger sample size is needed for precisely estimating the tails of the distributions when strong autocorrelation exists in the samples, as opposed to cases of weak autocorrelation. This bias at the tails did not affect our analysis.

In applying the K-S tests, we intended to reduce the risk of accepting the null hypothesis when it is false at the cost of increasing the risk of rejecting the null hypothesis when it is true. A better assessment through improved statistical tests may be still better than our compromise measures. As has been indicated, the problem arises from the critical values of the test statistic for a given significance level when the K-S test is performed on data samples that do not fully satisfy the test assumptions, especially if the data samples are serially correlated. Therefore, determining valid critical values of the test statistic may be a possible solution to the problem, instead of using the estimated values from the assumed probability distribution of the statistic in the K-S test. Since the advent of inexpensive and fast computing, resampling tests or Monte-Carlo tests (Wilks 1995) have become practical for this purpose.

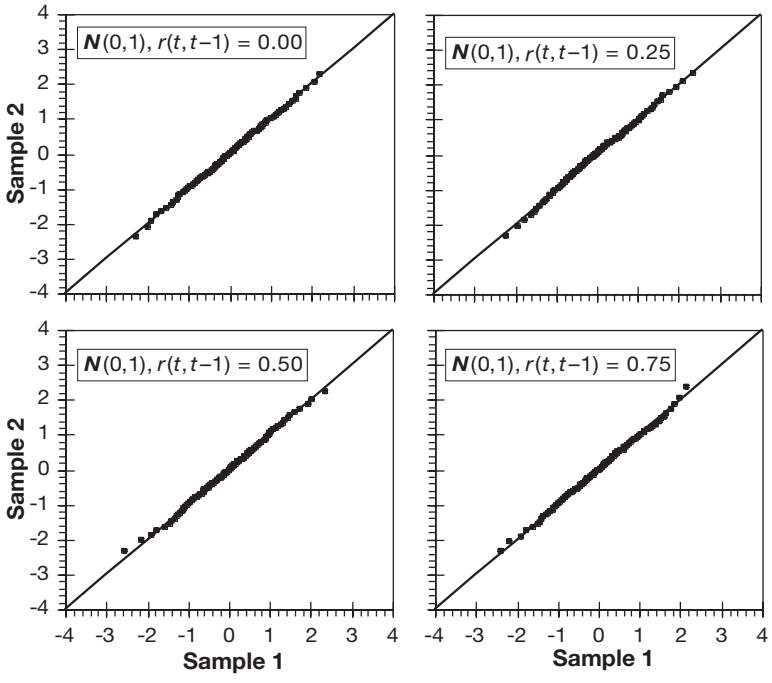

Fig. 1. Q–Q plot of 2 samples (900 elements for Sample 1 and 9000 elements for Sample 2), generated using the standard normal distribution $\boldsymbol{N}(0,1)$ with a lag-1 autocorrelation $r(t, t-1)$

The critical values of the K-S test statistic D_{mn} were estimated through resampling tests for each test. The resampling procedure was as follows: (1) pooling the 30 yr observations of daily maximum (or minimum) temperature and the corresponding 300 yr synthetic data generated by LARS-WG and AAFC-WG, to form a 330 yr data pool; (2) random selection of a year from the 330 yr data pool without replacement for 30 times, forming a 30 yr sample; (3) random selection of a year from

the remaining 300 yr data pool without replacement for 300 times, forming a 300 yr sample; (4) computation of the test statistic D_{mn} from the empirical CDFs estimated from the 30 yr sample and the 300 yr sample; (5) repetition of Steps 2 to 4 for 1000 times to obtain 1000 values of D_{mn}; (6) taking the 95th percentile of the 1000 values of D_{mn} as the critical value of the test statistic D_{mn} for the test at the 5% significance level. This significance level is the probability of a Type I error, i.e. the probability of falsely rejecting the null hypothesis given that it is true. The order of the data years in the 2 samples does not affect the magnitude of the test statistic in the resampling test. The magnitude of D_{mn} is determined by the years included in each sample. After the critical values of D_{mn} were determined, they were used to reject or accept the null hypothesis in the K-S test, for the corresponding location, month, weather variable and the weather generator. Test results for daily maximum and minimum temperature distributions (corresponding to Table 3 in Qian et al. 2004) are listed in Table 2. As expected, the rejection rate was reduced for the synthetic data generated by both weather generators, while the previous conclusions relating to their relative performance remain valid. Nevertheless, much smaller values of the test statistic D_{mn} were often found for the synthetic data from AAFC-WG, compared to LARS-WG. Alternatively, estimation of effective sample size may be applicable in determining the valid critical values for the tests.

It is always important to use statistical techniques appropriately in applied research, and to take into account the assumptions they require. The use of resampling or Monte-Carlo tests may be an alternative to the standard tests, which often require assumptions that may not apply. The resampling approach should be applicable for testing the means and variances in validations of weather generator simulations where the conventional *t*-test and *F*-test are commonly applied, even though daily weather series may not fully satisfy the assumptions for these tests. However, caution is still required, as basic assumptions are also applied in resampling tests. For example, serial correlation may still have an effect on statistical inferences made with resampling procedures (Zwiers 1990). When observations are serially correlated, inferences will be made relative to incorrectly derived reference or sampling distributions, because the resampling process does not replicate the serial correlation structure of observed climate processes.

In our case, serial correlation of daily temperatures is significant in a month of a given year, rather than between years, e.g. daily temperature T_{i1} on Day 1 in January of Year i may be significantly correlated to daily temperatures T_{it} on Day t ($t = 2, 3, \ldots, 31$) of Year i rather than T_{jt} ($t = 1, 2, \ldots, 31$; $j \neq i$). In our resampling tests, we shuffled the years rather than the days, and therefore the serial correlation structure in the dataset for the test was preserved. It may be still worthwhile to mention that all our statistical analyses were based on the basic assumption that daily values of a weather variable in a month can be treated as a random variable; however, neither these analyses nor others (e.g. χ^2 test for distributions) will be applicable if this assumption is in question. This also applies to the tests or analyses in the comparisons of means, variances and other statistics of the synthetic daily weather data (as well as any weather data from numerical models) in a month with observations.

Table 2. Number of months showing significant differences between observed and simulated daily maximum temperature (T_x), minimum temperature (T_n) by LARS-WG and using the 2-sample Kolmogorov-Smirnov test with critical values from resampling tests at the 5% level

Stn	LARS-WG		AAFC-WG	
	T_x	T_n	T_x	T_n
Beaverlodge	4	5	0	0
Fredericton	5	6	1	1
Goose	3	6	0	0
Ottawa	3	3	0	0
Regina	3	5	0	0
Toronto	1	0	0	0
Truro	3	6	0	1
Vancouver	0	0	0	0
Winnipeg	2	4	0	0

Acknowledgements. B.Q. thanks Dr. Xuebin Zhang of the Climate Research Branch, Meteorological Service of Canada, for helpful discussions.

LITERATURE CITED

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

Semenov MA, Welham S (2004) Comments on the use of statistical tests in the comparison of stochastic weather generators by Qian et al. (2004). Clim Res 28:83–84

von Storch H (1995) Misuses of statistical analysis in climate research. In: von Storch H, Navarra A (eds) Analysis of climate variability: applications of statistical techniques. Springer, Berlin, p 11–26

Wilks DS (1995) Statistical methods in the atmospheric sciences. Academic Press, San Diego

Zwiers FW (1990) The effect of serial correlation on statistical inferences made with resampling procedures. J Clim 3: 1452–1461

Editorial responsibility: Otto Kinne, Oldendorf/Luhe, Germany

Submitted: November 26, 2004; Accepted: December 2, 2004
Proofs received from author(s): December 13, 2004