Received: 16 January 2021 | Revised: 4 May 2021 | Accepted: 7 May 2021 | Published on: 24 May 2021

DOI: 10.1002/joc.7196

RESEARCH ARTICLE

International Journal of Climatology RMetS

Testing for differences between two distributions in the presence of serial correlation using the Kolmogorov–Smirnov and Kuiper's tests

John R. Lanzante

NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey

Correspondence
John R. Lanzante, NOAA/Geophysical Fluid Dynamics Laboratory, 201 Forrestal Road, Princeton, NJ 08542, USA.
Email: john.lanzante@noaa.gov

Abstract

Testing for differences between two states is a staple of climate research, for example, applying a Student's *t* test to test for the differences in means. A more general approach is to test for differences in the entire distributions. Increasingly, this latter approach is being used in the context of climate change research where some societal impacts may be more sensitive to changes further from the centre of the distribution. The Kolmogorov–Smirnov (KS) test, probably the most widely-used method in distributional testing, along with the closely related, but lesser known Kuiper's (KU) test are examined here. These, like most common statistical tests, assume that the data to which they are applied consist of independent observations. Unfortunately, commonly used data such as daily time series of temperature typically violate this assumption due to day-to-day autocorrelation. This work explores the consequences of this. Three variants of the KS and KU tests are explored: the traditional approach ignoring autocorrelation, use of an 'effective sample size' based on the lag-1 autocorrelation, and Monte Carlo simulations employing a first order autoregressive model appropriate to a variety of data commonly used in climate science. Results indicate that large errors in inferences are possible when the temporal coherence is ignored. The guidance and materials provided here can be used to anticipate the magnitude of the errors. Bias caused by the errors can be mitigated via easy to use 'look-up' tables or more broadly through application of polynomial coefficients fit to the simulation results.

KEYWORDS

distributional testing, effective sample size, Kolmogorov–Smirnov test, Kuiper's test, Monte Carlo simulation, serial correlation, temporal coherence

1 | INTRODUCTION

In climate science, one of the most fundamental pursuits is determination of the significance of differences between two states or sets of conditions. For example, the two states may be characterized by opposite phases of the North Atlantic Oscillation (NAO), by dry versus wet phases of the Asian Monsoon, or by historical and future

climate states as simulated by a Global Climate Model (GCM). Typically, a statistical test is applied to a limited sample of data derived from each of the two states. The Student's t test is perhaps the most commonly applied test and can be used to infer differences between the means of the two states. Although quite useful, sometimes an analyst wants information regarding higher order differences beyond the means. For example, in a climate change context an increase in the upper tail of a temperature or precipitation distribution may be more impactful for some applications than a change in the mean. Thus, in some situations there is a desire to test for a general difference in distributions.

The Kolmogorov–Smirnov (KS) test has been used widely to perform such distributional testing. The KS test operates by quantifying the distance between the empirical distribution functions derived from two different samples of data. Since it is nonparametric it makes no assumptions regarding the underlying distributions from which the samples are drawn. However, like most commonly used statistical tests, there is an underlying assumption that the values within each sample are statistically independent. Violation of this assumption typically leads to an excessive rate of rejections of the null hypothesis that there is no difference. Because most common physical variables of interest to a climate scientist exhibit non-trivial correlation spatially (i.e., between nearby gridpoints or stations) and temporally (i.e., from one observation in time to the next) vigilance in hypothesis testing is warranted. It should be noted that although this work deals only with the effects of temporal coherence, once local (e.g., gridpoint or station) significance has been established by appropriate means, recent advances allow for addressing the problem of spatial coherence in a straightforward manner (Wilks, 2016).

With regard to addressing serial correlation in significance testing, it appears that Laurmann and Gates (1977) introduced to the atmospheric sciences community the notion of an 'effective sample size', n_{eff}, in relation to the actual sample size, n. They proposed that an estimate of n_{eff}, based on the lag-1 autocorrelation in the data, could be used to adjust an estimate of the variance in testing the difference in means between two samples, thereby accounting for serial correlation. Thereafter, Thiebaux and Zwiers (1984) demonstrated that the concept of an 'effective sample size' is nebulous with no unique way to estimate it and cautioned against its use in the general application of statistical tests. Later work by Zwiers and von Storch (1995) explored various ways in which serial correlation could be taken into account in testing for the difference between the means of two samples.

In spite of the complexities and cautions raised, substitution of n_{eff} for n in various statistical tests to account for autocorrelation in time series has proliferated. In particular, the simplified expression for n_{eff} based on the assumption of a first-order autoregressive process (Laurmann and Gates, 1977) has become the de facto means for dealing with serial correlation in significance testing. Surprisingly however, accounting for serial correlation in application of the KS test is often lacking. While several studies have examined this issue (Weiss, 1978; Durilleul and Legendre, 1992; Xu, 2013) they have not resulted in a general approach that can be readily implemented. The purpose of this work is to fill that void.

This work examines three competing strategies in application of the KS test, and the closely related Kuiper's (KU) test, both of which are detailed below:

- the *traditional* approach of ignoring serial correlation
- substitution of an *effective sample size*, n_{eff}, for n
- use of Monte Carlo *simulations* to account for serial correlation.

The third approach, which is the main focus, is analogous to that of Zwiers and von Storch (1995) who provided revised critical values for use in application of the Student's t test based on the lag-1 autocorrelation. As a shorthand, subscripts 't' for traditional, 'e' for effective, and 's' for simulation (using a polynomial fit to smooth the results) are affixed to KS, KU, or simply K when referring to both tests generically.

In what follows, before delving into the details of the methodology, Section 2 presents examples based on both actual observations as well as realistic synthetic data in order to motivate this work. The details of the methodology for the three approaches are given in Section 3 along with thorough instructions for implementation. Section 4 explores some of the properties of the simulation approach in comparison with the other two approaches. Finally, Section 5 concludes with a summary and outlines several different ways in which the results from this study can be applied in practice. Supplementary material contains the coefficients and translation tables that can be used to implement the K_s distributional testing procedures.

2 | MOTIVATIONAL EXAMPLES

Before delving into the details of the various methodologies some examples, based on both real-world observations as well as synthetic data, are presented to motivate this work. Daily maximum (T_{max}) and minimum (T_{min}) surface air temperature from Central Park in New York City, New York were obtained from the National Centers for Environmental Information (NCEI) (https://www.

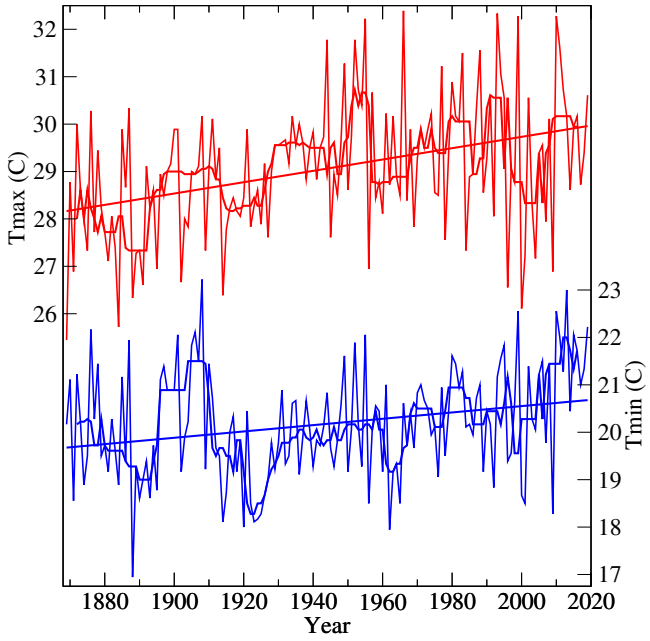

FIGURE 1 Monthly averaged maximum (T_{max}, red) and minimum (T_{min}, blue) temperature for Central Park, New York from 1869 to 2019 with 7-point running median (bold curve) and linear trend (bold line) [Colour figure can be viewed at wileyonlinelibrary.com]

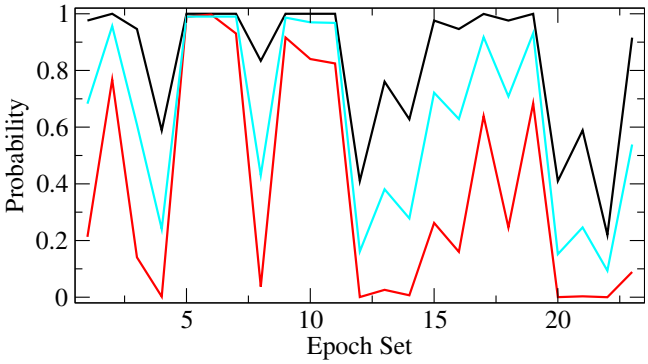

FIGURE 2 Probabilities (1 minus one-tailed significance) from KS test applied to Central Park July daily minimum temperature. Results are shown for three variants of the KS test: traditional (black), simulation (cyan) and use of n_{eff} (red). There are 23 epoch sets consisting of adjacent, non-overlapping 20-year periods such that each set is offset by 5 years. For example, set 1 tests 1870–1889 vs. 1890–1909, set 2 tests 1875–1894 vs. 1895–1914, and set 23 tests 1980–1999 vs. 2000–2019 [Colour figure can be viewed at wileyonlinelibrary.com]

ncdc.noaa.gov/cdo-web/datatools/findstation) spanning the time period 1869 to 2019. These data are from the daily Global Historical Climatology Network (GHCN-Daily) dataset prepared by Menne *et al*. (2012).

Figure 1 displays the time series of monthly averaged Central Park T_{max} and T_{min}. Although both series exhibit distinct warming trends, the smoothed curves highlight the considerable interdecadal variability. This low-frequency variability provides a convenient testbed for application of the KS and KU tests. A series of tests are applied to adjacent 20-year periods, separately for T_{max} and T_{min}, for January and July, and for daily values and monthly means. The first set consists of testing 1870–1889 versus 1890–1909. Shifting the starting points of the sets by 5 years per set produces a total of 23 sets terminating with 1980–1999 versus 2000–2019.

An example of the test results is show in Figure 2 for the KS test applied to Central Park July daily T_{min} with probabilities given for three variants of the tests (KS_t, KS_e, KS_s) for each of the 23 sets. Note that the probabilities here, and throughout the rest of this paper, are expressed as the cumulative probability, from minus infinity to the point in question. For example, a probability of .95 is equivalent to a one-tailed significance level of .05, typically interpreted as reason to reject the null hypothesis that the two samples are drawn from the same distribution. The K tests are one-tailed because a probability in the lower tail is indicative of a very close agreement between the two distributions, hence irrelevant to rejection of the null hypothesis of no difference.

The results in Figure 2 show a consistent pattern in which K_t (K_e) indicates the most (least) significant results with K_s in between, however the separation between the three tests varies considerably. That K_t is consistently the most significant is expected because any degree of autocorrelation, which should be considerable for daily data, reduces the amount of independent information available. Since K_t neglects this it is overly confident. However, these examples do not explicitly display the amount of dependence due to serial correlation.

In order to provide a perspective from which to interpret these results in a more controlled environment, synthetic daily data have been generated. Four sets of data consisting of 10,000 months each are derived from a first-order autoregressive process (AR1) with lag-1 autocorrelations of 0.0, 0.3, 0.6, and 0.9. While the details of the data generation are given below, the relevant point to note here is that since these data were generated in the same fashion as the data used in the Monte Carlo simulations to derive K_s, by construction results for KS_s and KU_s represent the 'truth'.

In evaluating K_t and K_e in the virtual world of synthetic data the customary probability of .95, corresponding to a significance level of .05, is adopted as an example. Thus, a K_t or K_e probability exceeding .95 defines a significant result. If the corresponding K_s probability falls below .95 this is denoted as a false positive (FP). Similarly, when K_s is significant but K_t or K_e are not, this represents a false

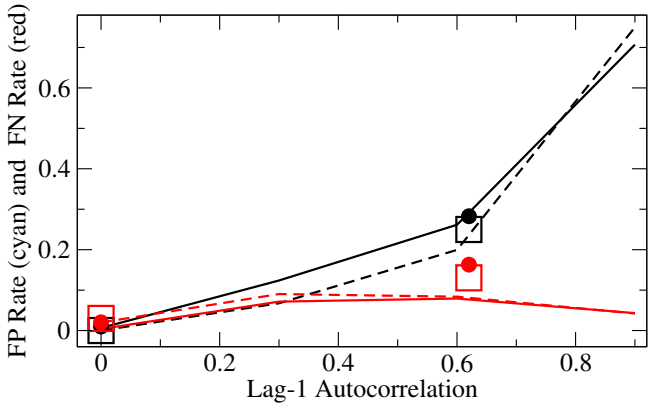

FIGURE 3 Average false positive (FP, K_t, black) and false negative (FN, K_e, red) rates for KS (solid) and KU (dashed) tests as a function of lag-1 autocorrelation as applied to 10,000 years of synthetic AR1 data having lag-1 autocorrelations of 0.0, 0.3, 0.6 and 0.9. Filled circles (open squares) represent averages for KS (KU) tests, as exemplified in Figure 2, applied to Central Park January and July maximum and minimum temperatures with left-most (right-most) cluster of symbols for monthly (daily) data. FP (FN) is the rate at which the traditional (n_{eff}) test attains (fails to attain) significance at the 5% level when the simulation test does not (does) [Colour figure can be viewed at wileyonlinelibrary.com]

negative (FN). Because of the predictable nature of the biases of K_t and K_e as seen in Figure 2, virtually all errors for K_t (K_e) are of the FP (FN) type; round off error for very close results leads to the rare exceptions.

Figure 3 displays the FP rates for K_t (black) and FN rates for K_e (red) as solid (KS) and dashed (KU) lines. For small values of serial correlation the FP and FN rates are modest. However, as the serial correlation increases the FP rate for K_t increases nonlinearly, reaching error rates ~30 to 50% for values of the lag-1 autocorrelation that may not be uncommon in many physical variables of interest, such as daily temperature. The K_e approach performs much better, especially for greater serial correlation but still has error rates ~5 to 10%. For comparison the values averaged over all of the Central Park cases are plotted as symbols. For monthly data, which typically have negligible correlation from year to year, the error rates are near zero. However, much larger error rates ~25% (15%) are seen for K_t (K_e). Note that choosing a fixed value (.95) to denote significance yields a simple, but narrow framework in which to compare the three approaches. Below the differences between the three tests are explored more fully and it is shown that the much higher error rate for K_t compared to K_e seen here is not universal—in other circumstances the roles may be reversed.

3 | DISTRIBUTIONAL TESTING METHODOLOGY

3.1 | Introduction

Distributional testing involves the comparison of empirical Cumulative Distribution Functions (CDFs) derived from two separate samples and can be divided into two classes: supremum and quadratic (Stephens, 1986). The former are based on the maximum difference between the two CDFs while the latter are based on the squared differences between the two CDFs. The most widely used is the KS test which utilizes the supremum approach. Note that the two-sample KS test is sometimes referred to as the Smirnov test (Wilks, 2006). Closely related to it is the lesser-known KU test. Two popular quadratic tests are the Cramer von Mises (CM) and Anderson-Darling (AD). While the KS test is more sensitive to differences near the middle to the distributions, with the CM often yielding similar results, the KU test is equally sensitive across the distribution while the AD test is more sensitive in the tails (Stephens, 1970, 1986). The AD test suffers from the fact that critical values are not as readily available since they depend on sample size (Pettitt, 1976).

This work focuses on the KS test because of its widespread use and also the KU test because it is so closely related to the KS test and is complementary with regards to its sensitivity. Details for the implementation of each of these tests are given below in three ways: the traditional approach, use of an effective sample size, and results derived from Monte Carlo simulations.

It is important to note that all of the results herein pertain to tests applied to two distinct samples. In some contexts it is desirable to compare the CDF from a sample with the CDF from a parametric fit to the same sample. Unfortunately, this violates the basic assumption that the two CDFs are derived from independent samples of data. In such instances, another approach, such as the Lilliefors test (Lilliefors, 1967) or Monte Carlo simulation are required.

3.2 | The traditional approach (K_t)

The first step in performing the K tests involves creating empirical CDFs, $F_1(x)$ and $F_2(x)$, from the two samples of data (Press *et al.*, 1992) and defining:

$$D^+ = \max[F_2(x) - F_1(x)] \quad (1)$$

$$D^- = \max[F_1(x) - F_2(x)] \quad (2)$$

For the KS test define:

$$D=\max(|D^+|,|D^-|) \tag{3}$$

and for the KU test define:

$$V=|D^+|+|D^-| \tag{4}$$

In summary, the KS test is based on the maximum distance between the two CDFs while the KU test is based on the sum of the largest distance above and largest distance below.

Next, if n_1 and n_2 are the two sample sizes compute:

$$N=[(n_1 n_2)/(n_1+n_2)]^{1/2} \tag{5}$$

Finally, define the KS (λ_{KS}) and KU (λ_{KU}) test statistics:

$$\lambda_{KS}=[N+0.12+(0.11/N)]D \tag{6}$$

$$\lambda_{KU}=[N+0.155+(0.24/N)]V \tag{7}$$

While the significance levels corresponding to the λ_K test statistics are available from software packages, they can be estimated via summation series. First, define the significance level (α):

$$\alpha=1-p \tag{8}$$

where p is the probability summed to the upper tail (i.e., the CDF value).

For the KS test:

$$\alpha_{KS}=2\sum_{j=1}^{\infty}(-1^{j-1}\exp(-2j^2{\lambda_{KS}}^2) \tag{9}$$

and for the KU test:

$$\alpha_{KU}=2\sum_{j=1}^{\infty}(4j^2{\lambda_{KU}}^2-1)\exp(-2j^2{\lambda_{KU}}^2) \tag{10}$$

Although these are infinite series they generally converge rapidly and no more than 100 terms are necessary (Press *et al.*, 1992). The summations can be terminated when either of two conditions are met:

- when the absolute value of the current term in the summation is $\leq 10^{-3}$ times the absolute value of the previous term or
- when the absolute value of the term is $\leq 10^{-8}$ times the current sum.

3.3 | Use of an effective sample size (K_e)

The approach introduced here based on an effective sample size is a simple extension of the traditional approach employing an assumption used widely in climate science. Following Laurmann and Gates (1977), if one assumes an AR1 process an effective sample size (n_{eff}) can be defined in terms of the actual sample size (n) and the lag-1 autocorrelation (r):

$$n_{eff}=[(1-r)/(1+r)]n \tag{11}$$

Note that when $r<0$ here we set $n_{eff}=n$. This is a conservative approach which prevents n_{eff} from exceeding n. Some may prefer a less conservative approach in which n_{eff} is allowed to exceed n, reflecting the reduced sampling variability that would be encountered. Once effective sample sizes have been estimated they can be substituted for n_1 and n_2 in (5), proceeding with the traditional approach as outlined above.

3.4 | Monte Carlo simulation and polynomial approximation (K_s)

The simulation procedure employed here is based on one critical assumption, namely that the data to be tested follow that of an AR1 process:

$$X_t=rX_{t-1}+e_t \tag{12}$$

where X_t is the value at time t, r is the lag-1 autocorrelation, and e_t is a zero mean Gaussian random variable with constant variance. Note that while (12) represents a zero-mean process, this does not limit the generality of the results. Strictly speaking, if the AR1 model is not appropriate for the data in question the results of this work cannot be applied. However, in geophysics many common physical variables can be approximated by an AR1 process.

The procedure for the simulations is straightforward, relying on brute-force computing power. A series of simulations, each consisting of 1,000 trials, were performed in which two samples were generated independently from the same AR1 model. For each trial the λ values from (6) or (7) were computed and saved. Based on the distribution of 1,000 λ values quantiles were extracted corresponding to the following percentiles in the CDF: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.96, 0.97, 0.98, and 0.99. Separate trials were performed for the following values of the lag-1 autocorrelation: 0.0, 0.1, 0.2, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95.

Although all of the trials were based on large sample sizes of $n_1 = n_2 = 20{,}000$ in order to reduce sampling error, this was not crucial. One of the simplifying features of the K tests is that the sample size dependence is captured in the λ values via (5)–(7) so that λ critical values are essentially independent of sample size. This was verified by running some simulations for sample sizes of 10, 100, 1,000, and 10,000. This is consistent with the statement by Press *et al*. (1992) that the approximation in (9) is quite good even for n as small as 8, noting as well that (9) and (10) have no sample size dependence. In addition, the fact that the two samples had equal sizes does not diminish the wide applicability of the results since differing sample sizes are also taken into account by (5).

In order to make the simulation results more readily available for application a series of polynomials were fit to the raw quantile values described above. Each CDF for a given lag-1 value was subdivided into three segments, with a separate polynomial fit to each segment. The three segments correspond to CDF ranges of 0.0–0.10, 0.10–0.50, and 0.50–0.99 for KU. For KS the cut-off of 0.5 is mostly replaced by 0.6, the choice being dictated by a better fit. For the lower two segments the polynomial is linear while the upper one is cubic. All of the coefficients and ranges for which they apply are given in Tables S1 and S2 (supporting information), along with an example for their use. These can easily be used to create computer code that returns significance levels in the user's programming language of choice. Note that in application the user would apply (1)–(7) as for the traditional approach, but use probabilities from the polynomials in lieu of (9) and (10). Throughout this paper, results referred to as 'simulation' (K_s) are based on the values derived from the polynomial fits.

There is one additional consideration in application of the polynomial fits, namely how to handle the lag-1 autocorrelations. All of the simulations use the same lag-1 for both samples as it would not have been practical to simulate all of the 126 combinations of the 17 levels used. In most applications one would anticipate that the autocorrelations for the two samples would not be too dissimilar. Given lag-1 autocorrelations of r_1 and r_2 there would seem to be three reasonable choices, two of which involve a consensus value r being applied to the polynomials:

a. $r = \max(r_1, r_2)$
b. $p = \text{average}(p_1, p_2)$
c. $r = \text{average}(r_1, r_2)$

A conservative approach (a) would be to apply the larger of the two autocorrelations to the polynomial fit. This would tend to underestimate the level of significance. The second approach (b) would be to apply r_1 and r_2 separately and average the resulting probabilities. The third approach (c), the recommended one here, is to apply the average of the two autocorrelations to the polynomial fit, using Fisher's z transformation (Zar, 2010) for the averaging:

$$z_1 = 0.5\ln[(1+r_1)/(1-r_1)) \tag{13}$$

$$z_2 = 0.5\ln[(1+r_2)/(1-r_2)) \tag{14}$$

$$z_{\text{avg}} = (z_1+z_2)/2 \tag{15}$$

$$r = [(\exp(2z_{\text{avg}}) - 1]/[(\exp(2z_{\text{avg}}) + 1)] \tag{16}$$

As further incentive for use of the results produced herein a series of 'look-up tables' have been produced, which although less accurate, are easier to apply than the full set of polynomials. Tables S3–S6 (Supporting Information) allow the user to enter the probability arrived at based on either the traditional or n_{eff} approaches, along with the lag-1 autocorrelation, to yield the probability one would obtain based on the polynomial fits. An example illustrating their use is included in the tables. A drawback of this approach is the ability to discriminate between levels of significance varies by probability and lag-1 as discussed in Section 5.

4 | PROPERTIES OF THE SIMULATION RESULTS AND POLYNOMIAL FITS

4.1 | Polynomial fits

Curves displaying the polynomial fits to the Monte Carlo simulation results are shown in Figure 4. Both the lower and upper portions of the CDF exhibit highly nonlinear behaviour. Although the linear fits in the lowest segment are far from ideal (see below), it was decided that the considerable effort needed (i.e., more simulations and much higher order curve fitting) for a proper rendering is not justified since this is a region of extreme non-significance; the small errors incurred will not change any inferences. At the high end of the distribution the user would be advised not to extend results beyond a CDF of 0.99. When a λ exceeds the upper limit (as indicated in Tables S1 and S2) for the given polynomial the probability should be censored to a value of 0.99. Again, in this region results that are highly significant will not be changed by this approach.

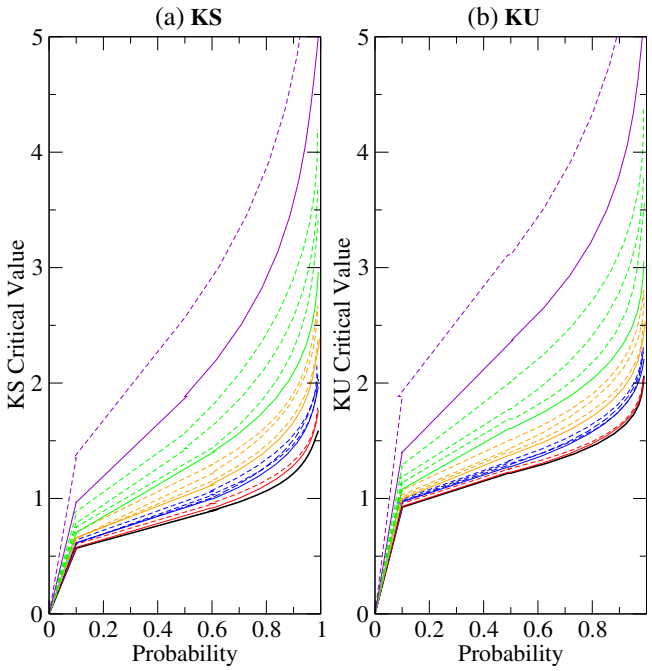

FIGURE 4 Critical values of the test statistic for (a) KS and (b) KU from the polynomial fits to the Monte Carlo results as a function of probability. Coloured curves represent different values of the lag-1 autocorrelation: black (0.0), red (0.1, 0.2), blue (0.3, 0.35, 0.4, 0.45), orange (0.50, 0.55, 0.6, 0.65), green (0.7, 0.75, 0.8, 0.85), and violet (0.9 0.95), with solid lines for 0.0, 0.1, 0.3, 0.5, 0.7, and 0.9, and dashed for others [Colour figure can be viewed at wileyonlinelibrary.com]

The other noteworthy nonlinearity concerns the spacing between the curves representing different levels of autocorrelation. As autocorrelation increases the curves are farther apart indicating a greater sensitivity to the lag-1 value. Taking a big-picture look at the results, consider the fact that the bottom-most curve (lag-1 = 0) represents critical values used in the traditional approach. It is clear that for lag-1 values typical of daily surface temperature for example (i.e., ~0.6 to 0.7, the upper orange and lower green curves) the departure is considerable.

4.2 | Polynomial fits vs. numerical recipes for lag-1 = 0

As a means of verifying the validity of the Monte Carlo and curve-fitting approach, simulations were performed for lag-1 = 0. In theory, these results should be the same as those from the traditional approach, which in this case was carried out using Numerical Recipes in FORTRAN (Press *et al*., 1992). As seen in Figure 5, for the bulk of the distributions (0.10–0.99) the results are virtually indistinguishable. At the lower end (0.0–0.1) the error incurred by the linear approximation is obvious, although

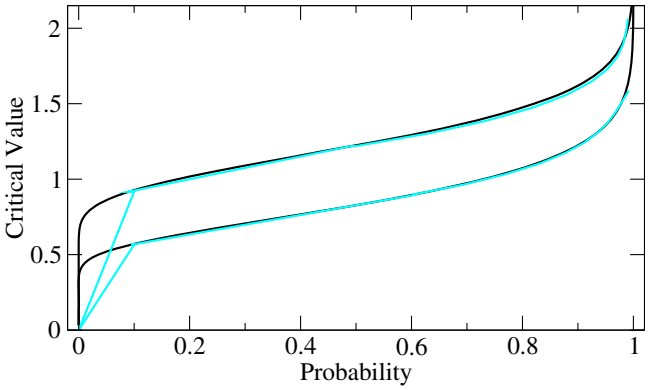

FIGURE 5 Critical values of the test statistic for KU (upper) and KS (lower) curves with black based on numerical recipes and cyan based on fitted values from Monte Carlo simulation for lag-1 autocorrelation = 0. Fitted values do not exceed a probability of 0.99 by design [Colour figure can be viewed at wileyonlinelibrary.com]

significance at this end is usually of little practical importance. However, Stephens (1986) points out that very small probabilities indicate that the small differences between the two CDFs are unlikely to be random. Such results are called superuniform and sometimes indicate that the data have been tampered with.

4.3 | Raw Monte Carlo vs. fitted

Figure 6 displays the raw Monte Carlo values along with the polynomial fitted curves for some select values of autocorrelation. It can be seen that the fitted curves represent the raw values quite well. In the far right tail, particularly for higher values of autocorrelation, the fitted curves smooth out noise expected at the end of the distribution.

4.4 | Comparisons of distributions of K_t, K_e, and K_s critical values

The CDFs for the K_t, K_e and K_s critical values are shown in Figure 7 for four values of autocorrelation. The three approaches yield essentially identical values when lag-1 = 0, except for $p < .10$ where the poorer linear fit is used for K_s. As autocorrelation increases, the curves for K_e and K_s move farther to the right, indicating a greater disparity with the traditional approach which does not take into account temporal coherence.

Note also how the relative positions involving the three approaches vary as a function of autocorrelation, especially for higher autocorrelation. For large probabilities (i.e., more

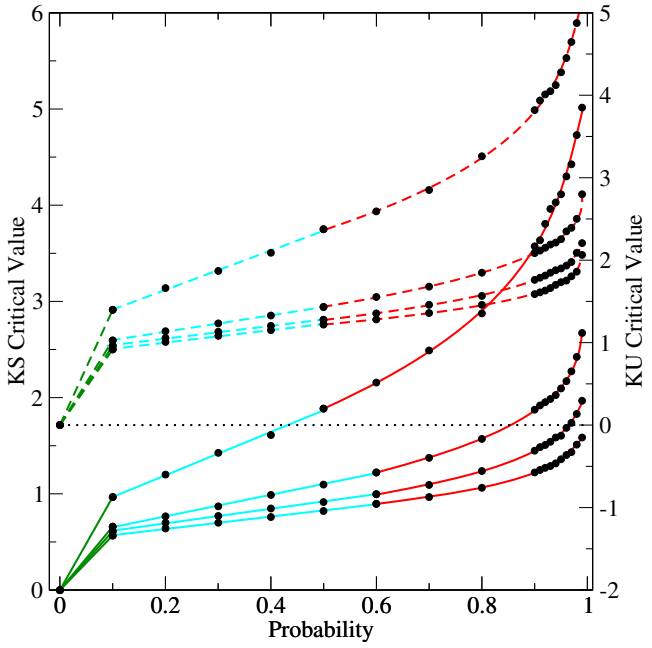

FIGURE 6 Critical values of the test statistic for KS (solid curves and left axis) and KU (dashed curves and right axis) as a function of probability. Each set of four curves, from lower to higher, correspond to lag-1 autocorrelations of 0.0, 0.3, 0.6, and 0.9. For each curve, filled circles represent values derived from 1,000 Monte Carlo simulations with sample sizes of 20,000. Fitting was done using three polynomials, linear for the left (dark green) and middle (cyan) segments, and cubic for the right segment (red). Note the vertical offset of the KS and KU curves (for clarity) with the dashed horizontal line as the zero axis for KU [Colour figure can be viewed at wileyonlinelibrary.com]

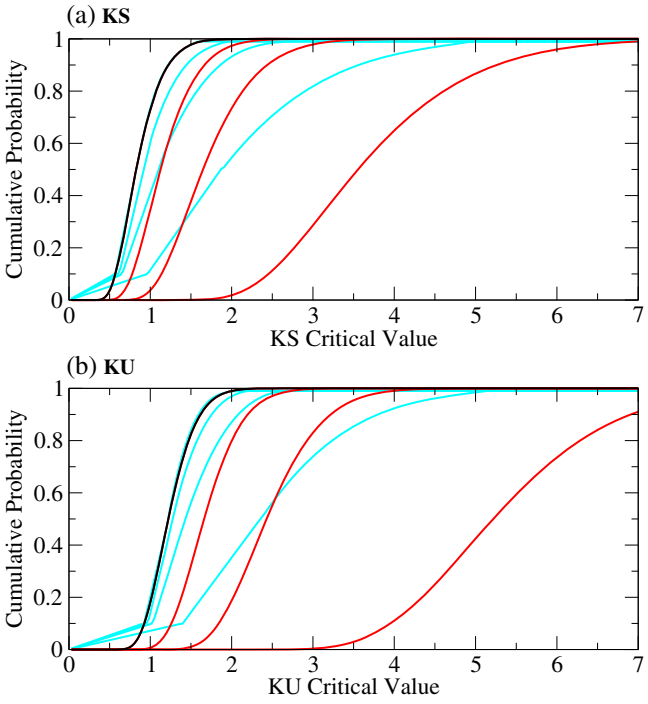

FIGURE 7 Cumulative distribution functions (CDFs) of critical values for three variants of the (a) KS test and (b) KU test: traditional (black), simulation (cyan) and use of n_{eff} (red). For the simulation and n_{eff} sets, from left to right, the four curves correspond to lag-1 autocorrelations of 0.0, 0.3, 0.6, and 0.9. By definition, there is only one curve for the traditional approach, corresponding to lag-1 = 0. Note that for lag-1 = 0, by definition the n_{eff} approach is identical to the traditional approach and the simulation CDF is obscured for most of the range since it is nearly identical to the traditional CDF [Colour figure can be viewed at wileyonlinelibrary.com]

significant results) K_s is closer to K_e than K_t. This is consistent with the result from Figure 3 indicating smaller errors for the K_e than the K_t approach. However, results from Figure 3 are biased towards the high end of the distribution since a value of $p = .95$ was used in determining the FP and FN rates. In Figure 7 it can be seen that for lower probabilities K_s is closer to K_t than K_e. Thus, the reader should not be misled by the specific example in Figure 3—in fact which is better, K_t or K_e is a function of the position in the CDF and is modulated by the amount of autocorrelation.

As a compliment to Figure 7, Figure 8 shows the corresponding probability distribution functions (PDFs). Determination of a PDF is not as straightforward as a CDF because it first requires the estimation of the local slope of the CDF followed by application of a somewhat arbitrary kernel density smoothing operation. Here the smoothing was chosen to present a reasonable appearance. The step-function at the low end for K_s curves is due to the use of a linear fit at the low end. As in Figure 7, K_t and K_e are identical for lag-1 = 0 with K_s nearly the same except for $p < .1$. This figure makes it clear how the three approaches diverge as autocorrelation increases. For the largest autocorrelation, there is not a lot of overlap between the three distributions.

5 | DISCUSSION AND CONCLUSIONS

This work has examined the problem of testing for differences between distributions based on two samples of data using both the widely used KS test as well as the closely related but lesser known KU test. While the former is more sensitive to differences near the middle of the distributions the latter is equally sensitive over the entire distribution. The question explored here is to what extent does lack of independence of the values within each sample, as quantified by the lag-1 autocorrelation, affect the outcomes of the tests? Three approaches to application of these tests were examined to address this question: the traditional approach ignoring the autocorrelation, a

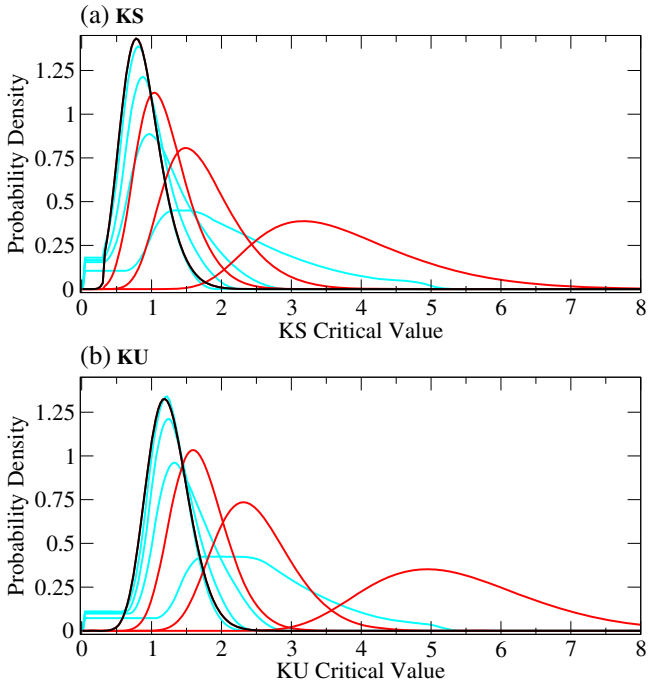

FIGURE 8 Same as Figure 7 except for probability density functions (PDFs) [Colour figure can be viewed at wileyonlinelibrary.com]

modified version using a simple estimate of the effective sample size, and Monte Carlo simulation.

Some examples using both real as well as synthetic data demonstrated that substantial errors can arise when the temporal coherence is ignored. Further diagnosis showed that the differences between the three approaches vary depending on the amount of autocorrelation and the degree to which the two distributions under consideration differ.

The test based on the simulation results is recommended for use. To facilitate implementation, coefficients of polynomials fit to the raw simulation results are provided (Supporting Information). As long as the data samples being tested conform to a reasonable extent to the assumed model used in the simulations the new test will eliminate the bias found in the traditional and effective sample size implementations. Also provided are look-up tables that are easier to use than the full implementation based on the polynomials.

The results from this work can be used in a tiered fashion, depending on how much effort a user is willing to invest. At the lowest tier, the information can provide guidance as to how seriously results will be affected by applying the traditional approach. In some cases, the user may decide that any adverse effects would be minimal. At the second level the user can, with only a little additional effort, gain considerable mitigation of the bias by way of look-up tables. At the highest level, which is the recommended route, the user can fully implement the simulation results via the polynomials.

Given the results shown in Figure 3 one might be tempted to use the K_e approach as a quick-fix alternative to K_t. However, the efficacy of such a strategy varies and may be more appropriate when the goal is to identify the most highly significant differences. As seen in Figure 7, K_e is closer to K_s than K_t only for high probabilities (at least ~0.9 to 0.95). In instances in which accuracy for less significant cases is important, for example when using a set of significance levels in assessing field significance (Wilks, 2016), K_t might be a better choice. In either circumstance, augmenting with use of the look-up tables will render more accurate results, but again results will vary.

If one examines the look-up tables for K_t (Tables S3 and S4) it is possible to achieve a high level of significance only for low values of lag-1 autocorrelation; much larger K_t probabilities, beyond the limits of the tables, would be needed. Conversely, for K_e (Tables S5 and S6), while high significance is possible, probabilities are censored at .99; however, the smallest probabilities are rather high in most cases. To achieve a more realistic range of probabilities in these two contrasting cases would require extensions into the more highly nonlinear regions of the relationships. In summary, while the look-up tables can be useful both diagnostically as well for some applications, they are not a panacea. More generally, the full implementation, utilizing the polynomial fits is recommended.

In closing, there is one further but important consideration for the potential user. Specifically, how appropriate is the statistical model used in the simulations, namely an AR1 process, for the data at hand? The AR1 assumption is appropriate for a broad range of meteorological and climate processes (Thiebaux and Zwiers, 1984; Zwiers and von Storch, 1995; von Storch and Zwiers, 2001). On the other hand some phenomena, particularly those of a quasi-periodic nature, such as the El Nino Southern Oscillation (ENSO), the stratospheric Quasi-Biennial Oscillation (QBO), and Madden and Julian Oscillation (MJO), to name a few, are better characterized as AR2. In other instances, the Gaussian assumption may not be valid. Ultimately, a judgement needs to be made as to whether any perceived violations of the assumptions are outweighed by the benefits of accounting for the autocorrelation effect, which, as we have seen, can at times be quite large. If the simulation model cannot be accepted and there is considerable autocorrelation then it would seem that the final course of action would be a set of Monte Carlo simulations appropriate for the data on hand. Ultimately the user should keep in mind the famous quote by statistician George Box: 'all models are wrong, but some are useful' (Box, 1976).

ACKNOWLEDGEMENTS

Comments on an earlier draft of this manuscript were kindly provided by Dennis Adams-Smith and Nat Johnson, as well as by the anonymous reviewers. Thanks also to Jeff Anderson for introducing me to the Kuiper's test and to Bill Stern for providing some of the FORTRAN code.

AUTHOR CONTRIBUTIONS

John Lanzante: Conceptualization; formal analysis; investigation; methodology; software; visualization; writing-original draft; writing-review & editing.

ORCID

John R. Lanzante https://orcid.org/0000-0002-1736-7170

REFERENCES

Box, G.E.P. (1976) Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949.

Durilleul, P. and Legendre, P. (1992) Lack of robustness in two tests of normality against autocorrelation in sample data. *Journal of Statistical Computation and Simulation*, 42(1–2), 79–91. https://doi.org/10.1080/00949659208811412.

Laurmann, J. and Gates, L. (1977) Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models. *Journal of the Atmospheric Sciences*, 34(8), 1187–1199. https://doi.org/10.1175/1520-0469(1977)034<1187:SCITEO>2.0.CO;2.

Lilliefors, H.W. (1967) On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318), 399–402. https://doi.org/10.1080/01621459.1967.10482916.

Menne, M.J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R.S., Gleason, B.E. and Houston, T.G. (2012) *Global historical climatology network – daily (GHCN-daily), Version 3.26*. Silver Spring, MA: NOAA National Climatic Data Center. https://doi.org/10.7289/V5D21VHZ.

Pettitt, A.N. (1976) A two-sample Anderson-Darling rank statistic. *Biometrika*, 63(1), 161–168. https://doi.org/10.2307/2335097.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1992) *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd edition. New York, NY: Cambridge University Press.

Stephens, M.A. (1970) Use of the Kolmogorov–Smirnov, Cramer-Von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society*, B32(1), 115–122. https://www.jstor.org/stable/2984408.

Stephens, M.A. (1986) In: D'Agostino, R.B. and Stephens, M.A. (Eds.) *Tests Based on EDF Statistics. Chapter 4 in Goodness-of-Fit Techniques*. New York: Marcel Dekker, pp. 97–194.

Thiebaux, H. and Zwiers, F.W. (1984) The interpretation and estimation of effective sample size. *Journal of Climate and Applied Meteorology*, 23(5), 800–811. https://www.jstor.org/stable/26181354.

von Storch, H. and Zwiers, F.W. (2001) *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press 484 pp.

Weiss, M.S. (1978) Modification of the Kolmogorov–Smirnov statistic for use with correlated data. *Journal of the American Statistical Association*, 73(364), 872–875. https://doi.org/10.2307/2286297.

Wilks, D.S. (2006) *Statistical Methods in the Atmospheric Sciences*, 2nd edition. San Diego, CA: Academic Press.

Wilks, D.S. (2016) "The stippling shows statistically significant gridpoints". How research results are routinely overstated and over-interpreted, and what to do about it. *Bulletin of the American Meteorological Society*, 97(12), 2263–2273. https://doi.org/10.1175/BAMS-D-15-00267.1.

Xu, X. (2013) *Methods in Hypothesis Testing, Markov Chain Monte Carlo and Neuroimaging Data Analysis*. Ph.D. dissertation. Cambridge, MA: Harvard University, p. 119. http://nrs.harvard.edu/urn-3:HUL.InstRepos:11108711.

Zar, J.H. (2010) *Biostatistical Analysis*, 5th edition. Upper Saddle River, NJ: Pearson Prentice Hall, p. 947.

Zwiers, F.W. and von Storch, H. (1995) Taking serial correlation into account in tests of the mean. *Journal of Climate*, 8(2), 336–351. https://doi.org/10.1175/1520-0442(1995)008<0336:TSCIAI>2.0.CO;2.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lanzante, J. R. (2021). Testing for differences between two distributions in the presence of serial correlation using the Kolmogorov–Smirnov and Kuiper's tests. *International Journal of Climatology*, *41*(14), 6314–6323. https://doi.org/10.1002/joc.7196