NOTES AND CORRESPONDENCE

A Cautionary Note on the Use of the Kolmogorov–Smirnov Test for Normality

DAG J. STEINSKOG

Nansen Environmental and Remote Sensing Center, and Bjerknes Centre for Climate Research, and Geophysical Institute, University of Bergen, Bergen, Norway

DAG B. TJØSTHEIM

Department of Mathematics, University of Bergen, Bergen, Norway

NILS G. KVAMSTØ

Geophysical Institute, University of Bergen, and Bjerknes Centre for Climate Research, Bergen, Norway

(Manuscript received and in final form 28 April 2006)

ABSTRACT

The Kolmogorov–Smirnov goodness-of-fit test is used in many applications for testing normality in climate research. This note shows that the test usually leads to systematic and drastic errors. When the mean and the standard deviation are estimated, it is much too conservative in the sense that its p values are strongly biased upward. One may think that this is a small sample problem, but it is not. There is a correction of the Kolmogorov–Smirnov test by Lilliefors, which is in fact sometimes confused with the original Kolmogorov–Smirnov test. Both the Jarque–Bera and the Shapiro–Wilk tests for normality are good alternatives to the Kolmogorov–Smirnov test. A power comparison of eight different tests has been undertaken, favoring the Jarque–Bera and the Shapiro–Wilk tests. The Jarque–Bera and the Kolmogorov–Smirnov tests are also applied to a monthly mean dataset of geopotential height at 500 hPa. The two tests give very different results and illustrate the danger of using the Kolmogorov–Smirnov test.

1. Introduction

The Kolmogorov–Smirnov test (hereafter the KS test) is a much used goodness-of-fit test. In particular, it is often employed to test normality, also in climate research. Normality tests are important for at least two reasons. First, nonlinearity and interacting physical processes usually lead to non-Gaussian distributions, and the generating mechanism of the processes can therefore be better understood by examining the distribution of selected variables. For instance, Burgers and Stephenson (1999) test for the normality of the amplitude of the El Niño–Southern Oscillation (ENSO) using moment estimates of skewness and kurtosis. Such moments can be used to diagnose nonlinear processes and to provide powerful tools for validating models (see also Gershunov et al. 2001). Other examples are Branstator and Berner (2005) and Berner (2005) analyzing planetary waves and Stephenson et al. (2004) investigating multiple climate regimes.

A second reason for implementing normality tests is that many statistical procedures require or are optimal under the assumption of normality, and it is therefore of interest to know whether or not this assumption is fulfilled. A recent example of such a use of a normality test is given in Sanders and Lea (2005) in their study of hurricane activity. Of course it may also be of interest to test for the presence of other specific distributions, as in Mohymont et al. (2004), who study extreme value distributions.

The KS test is employed in some of these papers, and

Corresponding author address: Dr. Dag Johan Steinskog, Nansen Environmental and Remote Sensing Center, Thormøhlensgt. 47, N-5006 Bergen, Norway.
E-mail: dag.johan.steinskog@nersc.no

DOI: 10.1175/MWR3326.1

in other papers as well, despite the fact that it probably should not be used. In particular, conclusions based on not rejecting normality could be very misleading. D'Agostino and Stephens (1986, p. 406) are strongly stressing that the test should not be applied if parameters have to be estimated (usually the case) and that it is of historical interest only. What makes this even worse is that the description of the KS test in standard software packages is quite confusing. In the S-Plus (Krause and Olson 2002) package, for example, even though it appears under the appellation of the KS test, it is not really this test that is used when parameters are estimated but rather the Lilliefors correction (LI) (Lilliefors 1967). In *R* (R Development Core Team 2005) and Matlab (Hunt et al. 2001), the appellation "Kolmogorov–Smirnov" is reserved for the KS test only and there is a separate Lilliefors test, the latter one being recommended as an alternative in Matlab even though it is not implemented to return higher p values than 0.20. For p values higher than 0.20, two approaches are possible—a time consuming Monte Carlo simulation of critical values or an approximation suggested by Stephens (1974) that is used in *R* and S-Plus (see also Dallal and Wilkinson 1986).

We have looked at cited publications in the geophysical literature, but they do not make clear which implementation has been used or if adjustments have been made to take into account the use of estimated parameters. If the KS test in R and Matlab has been used, there are reasons to believe that the results may be misleading. However, it is quite safe to use the S-Plus KS test for estimated parameters, which is really the Lilliefors test.

There is thus a need for clarification, and the purpose of this paper is to demonstrate how large the differences are between the incorrect and correct use of these tests. We will limit ourselves to the testing of normality. We will revisit the Lilliefors correction and, in addition, we will compare it to a number of alternative tests, all available in R. Both the size and the power of the tests will be examined, and a simple example with geophysical data is provided.

2. Test results

For comparison of the tests, simulated standard normal variables are used. In addition, the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalyzed (Kalnay et al. 1996) monthly mean geopotential heights at 500 hPa for the months of December, January, February, and March (DJFM) of the period 1950–2001 are explored. Altogether, 204 time points are analyzed.

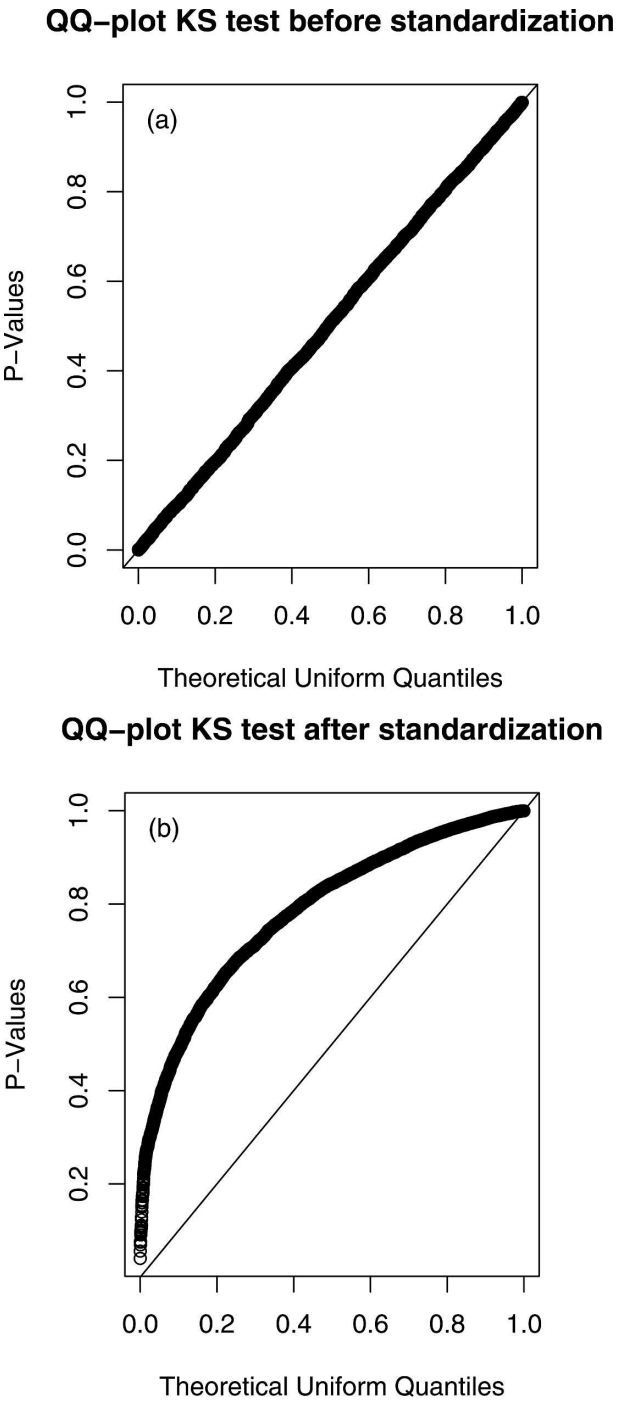

FIG. 1. Quantile–quantile plots of p values from the KS test on simulated data against a theoretical uniform distribution (a) before and (b) after standardization with $\hat{\mu}$ and $\hat{\sigma}$. The lines in the plots are 45° reference lines.

These data come from 4176 grid points in the Northern Hemisphere, and we would like to test the time series for normality at each site.

In the simulation experiment, 5000 independent stan-

TABLE 1. The number of resulting p values from the KS test with estimated parameters in seven intervals for different sample sizes under the null hypothesis of normality.

Sample size	0–0.05	0.05–0.10	0.10–0.20	0.20–0.40	0.40–0.60	0.60–0.80	0.80–1.00
1000	0	13	31	201	501	1029	2401
2000	2	5	26	222	486	1003	2432
3000	0	7	40	195	522	981	2431
5000	0	8	34	212	540	1045	2337
7000	1	5	35	223	539	1033	2340
8000	0	7	34	217	538	1036	2344
9000	3	7	36	186	533	1023	2388

dard normally distributed variables were generated at 4176 positions (i.e., 4176 tests each with sample size 5000), corresponding to the number of grid point locations of the Northern Hemisphere dataset. In practice, one will have to test for goodness of fit for a normal distribution with a nonzero (and unknown) mean and a nonunity (and unknown) standard deviation. We have chosen to use the standard normal in our simulations for matters of convenience. The estimation aspect is taken care of by pretending that we do not know the mean μ and the standard deviation σ and hence these quantities have been estimated; for given simulated data X_i, $i = 1, \ldots, n$, the normalized data points $(X_i - \hat{\mu})/\hat{\sigma}$ have been compared with the standard normal distribution. This is equivalent to comparing the X_i-data to a normal distribution with an estimated mean $\hat{\mu}$ and an estimated standard deviation $\hat{\sigma}$. There is one major difference between the real data and the simulated data. In the former dataset there is both some temporal and spatial dependence, and in the latter there is not. This is not important for the point we want to make. In fact, the simulation experiments were repeated for dependent data with very similar results (not shown).

Let $S_n(x)$ be the empirical cumulative distribution function and $F_0(x)$ the population cumulative distribution under the null hypothesis H_0. The KS statistic can be written as

$$d = \sup_x[|F_0(x) - S_n(x)|]. \quad (1)$$

The table of critical points $d_\alpha(n)$ of the distribution of d is presented for various sample sizes n and significance levels α in Massey (1967). In this paper, the cumulative distribution $F_0(x)$ is standard Gaussian. The test works well when the parameters are known, in our case $\mu = 0$, $\sigma = 1$. It is seen from the quantile–quantile (qq) plot in Fig. 1a that the p values of the 4176 tests are uniformly distributed, as they should be, but after standardization of the same data with $\hat{\mu}$ and $\hat{\sigma}$ (Fig. 1b), the test utterly fails. The p values of the KS test are much too large. This is due to overfitting of the data caused by the estimation of the mean and standard deviation. Massey (1967) mentioned this, but his warning and that of D'Agostino and Stephens (1986) seem to have gone largely unheeded, or at least, it has not induced authors to state precisely what they are doing when using the KS test. This tendency would also appear for nonnormal distributions with estimated parameters, and there are reasons to believe that it also persists in much more complicated situations. For instance, An and Cheng (1996) use the KS statistic to test linearity and get zero rejections in a simulation experiment when the null hypothesis H_0 is true, even though the nominal level of the test is 5%.

As a much too low proportion of small p values is observed (Table 1) for any sample size up to 9000, this is surely not a small sample effect. In fact, the distribution of p values is fairly stable as the sample size increases from 1000 to 9000.

The Lilliefors correction (Lilliefors 1967) uses the same test statistic as the KS test but adjusted critical values. A table of critical values can be obtained by Monte Carlo approximation. The Lilliefors test is not affected by the estimation of parameters, as is seen from Fig. 2. The test could be very time consuming if the Monte Carlo generation of exact critical values is chosen instead of an analytical approximation (Stephens 1974; Dallal and Wilkinson 1986), and given the existence of other tests, would only be retained if it has better power properties than its rivals.

There are several alternative tests for normality. We have used the Jarque–Bera (JB), Shapiro–Wilk (SW), Anderson–Darling (AD), Cramer–von Mises (CVM), Pearson chi-square (PCH), and Shapiro–Francio (SF) [for references see Thode (2002)]. The results of all of these tests under the null hypothesis of normality and with estimated parameters are very similar to that of the Lilliefors test of Fig. 2 and are therefore not plotted separately.

A simple study of the power of the tests has been undertaken. The nonnormal alternative models are similar to those used in the paper by Lilliefors (1967),

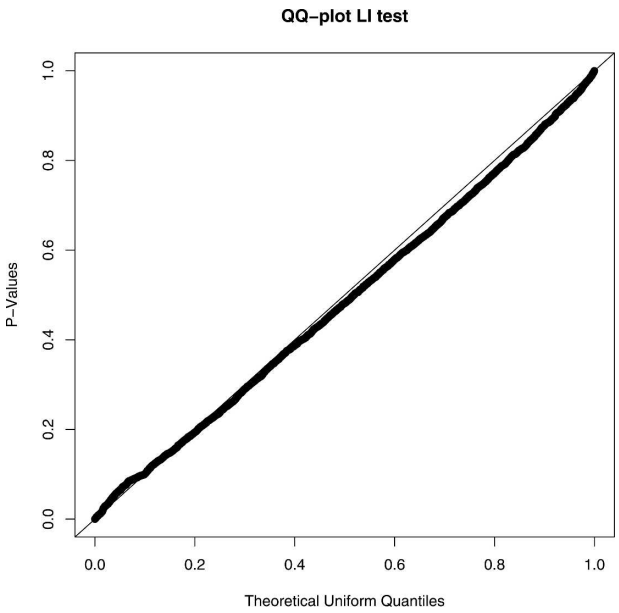

FIG. 2. Quantile–quantile plots of p values from the LI test against a theoretical uniform distribution using estimated parameters. The line in the plot is a 45° reference line.

and a sample size of $n = 204$ and 4176 realizations have been used to compare with the real data example. The simulated data have been standardized with $\hat{\mu}$ and $\hat{\sigma}$ before the tests were applied. Table 2 shows that with the exception of the KS statistic, all of the tests provide

TABLE 2. The power of the eight goodness-of-fit tests for normality applied to five distributions. The significance level is taken to be α, and the proportion of rejections is given in the table. NORM, TDIST, CHI, EXP, and UNI represent standard normal distribution, Student's t distribution (with 20 degrees of freedom), chi-square distribution (with 20 degrees of freedom), exponential distribution, and uniform distribution (with minimum equal to −1 and maximum equal to 1), respectively.

Test	α	NORM	TDIST	CHI	EXP	UNI
KS	0.05	0.0002	0.0005	0.0491	1	0.1305
	0.10	0.0012	0.0010	0.1365	1	0.3582
LI	0.05	0.0500	0.0711	0.6137	1	0.9540
	0.10	0.1056	0.1466	0.7426	1	0.9875
SW	0.05	0.0580	0.1458	0.9215	1	1
	0.10	0.1049	0.2186	0.9564	1	1
JB	0.05	0.0467	0.2031	0.8606	1	1
	0.10	0.0843	0.2560	0.9286	1	1
AD	0.05	0.0568	0.0934	0.8252	1	1
	0.10	0.1080	0.1739	0.8968	1	1
CVM	0.05	0.0558	0.0836	0.7522	0.9852	0.9981
	0.10	0.1070	0.1559	0.8427	0.9861	0.9990
PCH	0.05	0.0520	0.0568	0.3491	1	0.9131
	0.10	0.1032	0.1133	0.4859	1	0.9492
SF	0.05	0.0553	0.1882	0.9006	1	1
	0.10	0.1080	0.2763	0.9456	1	1

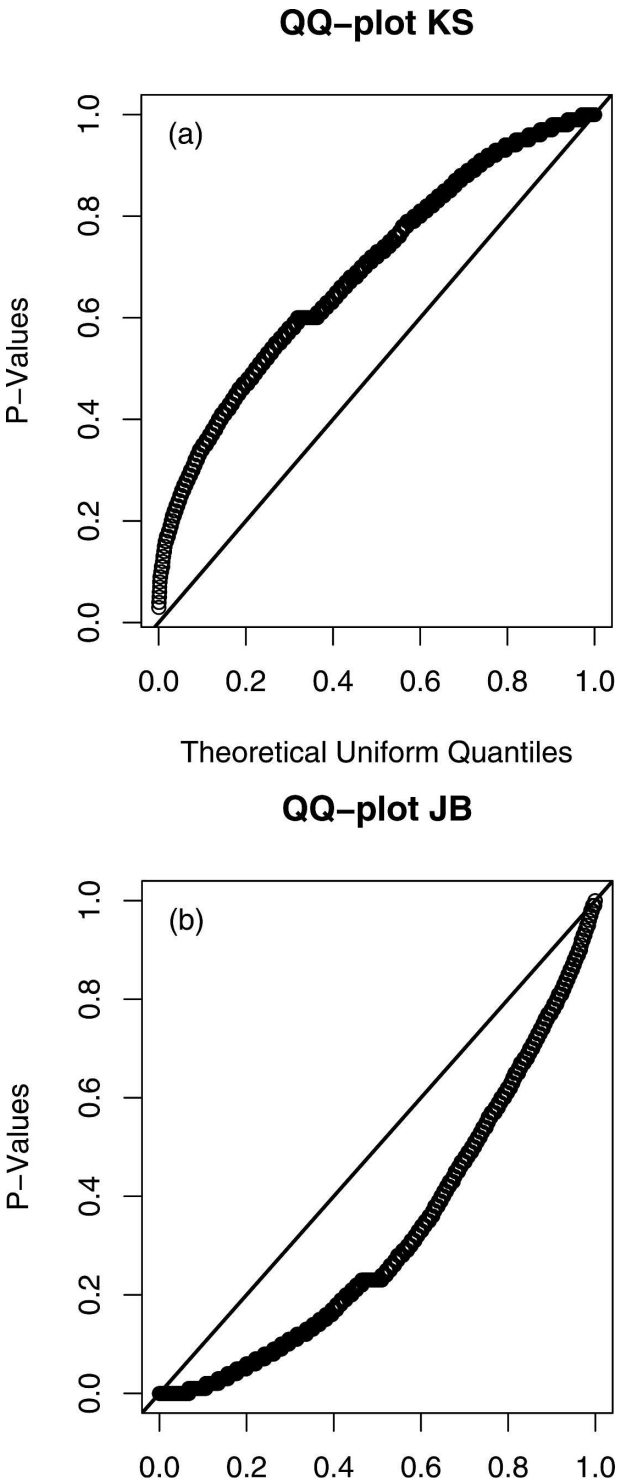

FIG. 3. Quantile–quantile plots of p values from the (a) KS and (b) JB test for a real dataset against a theoretical uniform distribution using estimated parameters. The lines in the plots are 45° reference lines.

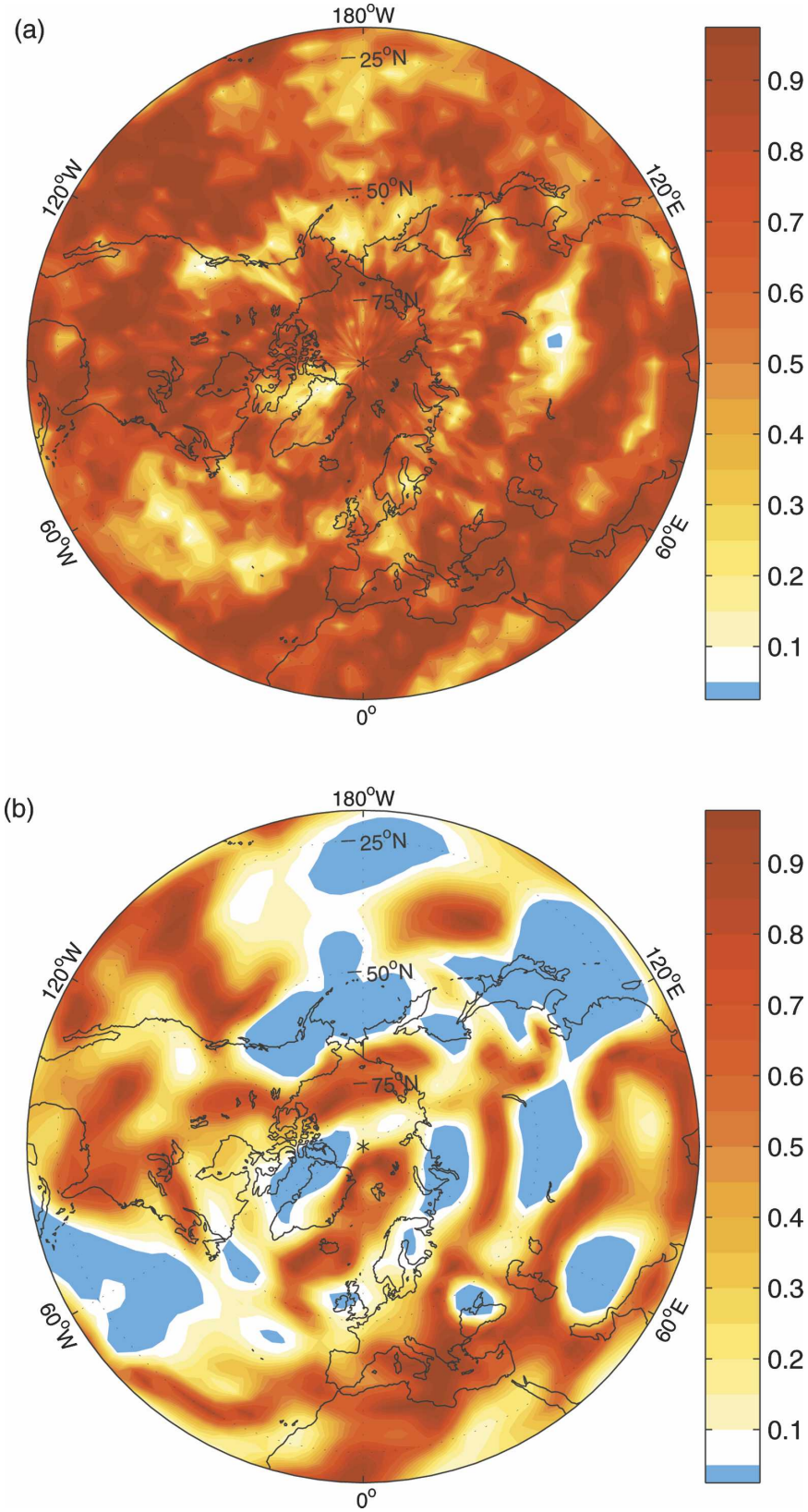

FIG. 4. The spatial structure of p values for the (a) KS and (b) JB test for DJFM at 500-hPa geopotential height. Blue regions indicate p values less than 0.05, locations where normality is rejected assuming $\alpha = 0.05$ as a significance level. For values above $\alpha = 0.05$, shading from yellow to dark red is used, where dark red indicates high p values.

small p values (large power) for three of the nonnormal alternatives. The Student's t distribution with 20 degrees of freedom is close to the normal distribution, and all of the tests have trouble with that alternative. Overall the Jarque–Bera and the Shapiro–Wilk test have the best power for these examples, and they are clearly better than the Lilliefors test. The KS test fares very poorly. Its level under the null hypothesis is much too low, and with the exception of the exponential distribution, its power is disastrous. Again it is demonstrated that if the KS test does not reject normality for a given dataset, this result carries almost no informative value.

For the atmospheric data, we have tested normality on monthly mean geopotential height data at 500 hPa to highlight the pronounced differences between the KS and the JB tests at a nominal significance level of 5%. As can be observed in Fig. 3a, the KS test hardly rejects normality anywhere. On the other hand, the JB test (Fig. 3b) demonstrates clearly that the data are nonnormally distributed at a number of locations. The different results have geophysical implications in that different interpretations and different analyses would be optimal for normal as opposed to nonnormally distributed data.

Moreover, by investigating this example closer, we observe from Fig. 4b that areas of low p values when using the JB test (blue areas) coincide with or are situated near regions with local maxima in quasi-stationary wave amplitudes. One such example is near the East Coast trough. An examination of the marginal time series in the region (not shown) indicates that the low p values are caused by the fact that the negative anomalies have higher amplitudes than the positive ones. This is consistent with strong eddy activity, frequent baroclinity, and thus widespread regional cyclogenesis. These phenomena largely determine the mean location of the East Coast trough, while regional sea surface temperatures (SSTs) and the North Atlantic Oscillation (NAO) are associated with the intraseasonal and multiyear variability (Lau 1988; Harman 1991; Colucci 1976; Zishka and Smith 1980). None of these effects were detected by the KS test (Fig. 4a).

3. Concluding remarks

The standard KS goodness-of-fit test is not recommended when parameters are estimated from the data (usually the case), regardless of sample size. It should also be noted that the term "KS test" does not necessarily mean the same test in different software packages. The correction made by Lilliefors (1967) and later updated by Dallal and Wilkinson (1986) and Stephens (1974) is a better alternative, but it seems to have less power than, for instance, the SW and JB tests.

While the nonnormal distributions detected by the JB test in the real data example can be interpreted physically, the KS test fails to detect these phenomena. The problems of the KS test persist under other distributions under H_0, when the parameters of these distributions have to be estimated.

Acknowledgments. We wish to thank David Stephenson and Chris Ferro for their constructive remarks on a draft of this paper. This work has received support from the University of Bergen through a grant from the Board of Marine Sciences at the Faculty of Mathematical Sciences. The work has also received support from the EU project ENSEMBLES (GOCE-CT-2003-505539). Moreover, we are very grateful to three reviewers for their useful comments and suggestions.

REFERENCES

An, H. Z., and B. Cheng, 1996: A Kolmogorov–Smirnov type statistic with application to test for normality of time series. *Int. Stat. Rev.,* **59,** 45–61.

Berner, J., 2005: Linking nonlinearity and non-Gaussianity of planetary wave behavior by the Fokker–Planck equation. *J. Atmos. Sci.,* **62,** 2098–2117.

Branstator, G., and J. Berner, 2005: Linear and nonlinear signatures in the planetary wave dynamics of an AGCM: Phase space tendencies. *J. Atmos. Sci.,* **62,** 1792–1811.

Burgers, G., and D. B. Stephenson, 1999: The "normality" of El Niño. *Geophys. Res. Lett.,* **26,** 1027–1030.

Colucci, S. J., 1976: Winter cyclone frequencies over the eastern United States and adjacent western Atlantic, 1964–1973. *Bull. Amer. Meteor. Soc.,* **57,** 548–553.

D'Agostino, R. B., and D. A. Stephens, 1986: *Goodness-of-Fit Techniques.* Marcel Dekker, 576 pp.

Dallal, G. E., and L. Wilkinson, 1986: An analytic approximation to the distribution of Lilliefors's test statistic for normality. *Amer. Stat.,* **40,** 294–296.

Gershunov, A., N. Schneider, and T. Barnett, 2001: Low-frequency modulation of the ENSO–Indian monsoon rainfall relationship: Signal or noise? *J. Climate,* **14,** 2486–2492.

Harman, J. R., 1991: *Synoptic Climatology of the Westerlies: Process and Pattern.* Association of American Geographers, 80 pp.

Hunt, B. R., R. L. Lipsman, and J. M. Rosenberg, 2001: *A Guide to MATLAB: For Beginners and Experienced Users.* Cambridge University Press, 416 pp.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

Krause, A., and M. Olson, 2002: *The Basics of S-PLUS.* 3d ed. Springer, 448 pp.

Lau, N.-C., 1988: Variability of the observed midlatitude storm tracks in relation to low-frequency changes in the circulation pattern. *J. Atmos. Sci.,* **45,** 2718–2743.

Lilliefors, H. W., 1967: On the Kolmogorov–Smirnov test for normality with mean and variance unknown. *J. Amer. Stat. Assoc.,* **62,** 399–402.

Massey, F. J., Jr., 1967: The Kolmogorov–Smirnov test for goodness of fit. *J. Amer. Stat. Assoc.,* **46,** 68–78.

Mohymont, B., G. R. Demarée, and D. N. Faka, 2004: Establishment of IDF-curves for precipitation in the tropical area of Central Africa—Comparison of technique and results. *Nat. Hazards Earth Syst. Sci.,* **4,** 375–387.

R Development Core Team, 2005: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2535 pp. [Available online at http://www.R-project.org.]

Sanders, M. A., and A. S. Lea, 2005: Seasonal prediction of hurricane activity reaching the coast of the United States. *Nature,* **434,** 1005–1008.

Stephens, M. A., 1974: EDF statistics for goodness of fit and some comparisons. *J. Amer. Stat. Assoc.,* **69,** 730–737.

Stephenson, D. B., A. Hannachi, and A. O'Neill, 2004: On the existence of multiple climate regimes. *Quart. J. Roy. Meteor. Soc.,* **130,** 583–605.

Thode, H. C., 2002: *Testing for Normality.* Marcel Dekker, 368 pp.

Zishka, K. M., and P. J. Smith, 1980: The climatology of cyclones and anticyclones over North America and surrounding ocean environs for January and July, 1950–77. *Mon. Wea. Rev.,* **108,** 387–401.