Simple Nonparametric Techniques for Exploring Changing Probability Distributions of Weather

CHRISTOPHER A. T. FERRO, ABDELWAHEB HANNACHI, AND DAVID B. STEPHENSON

Department of Meteorology, University of Reading, Reading, United Kingdom

(Manuscript received 11 February 2004, in final form 18 February 2005)

ABSTRACT

Anthropogenic influences are expected to cause the probability distribution of weather variables to change in nontrivial ways. This study presents simple nonparametric methods for exploring and comparing differences in pairs of probability distribution functions. The methods are based on quantiles and allow changes in all parts of the probability distribution to be investigated, including the extreme tails. Adjusted quantiles are used to investigate whether changes are simply due to shifts in location (e.g., mean) and/or scale (e.g., variance). Sampling uncertainty in the quantile differences is assessed using simultaneous confidence intervals calculated using a bootstrap resampling method that takes account of serial (intraseasonal) dependency. The methods are simple enough to be used on large gridded datasets. They are demonstrated here by exploring the changes between European regional climate model simulations of daily minimum temperature and precipitation totals for winters in 1961–90 and 2071–2100. Projected changes in daily precipitation are generally found to be well described by simple increases in scale, whereas minimum temperature exhibits changes in both location and scale.

1. Introduction

The comparison of two time series is a common task in climate research. Daily precipitation, for example, might be compared at two sites, or during two periods at one site. Many comparisons could be made, but here the focus is on differences in the marginal probability distributions of the two series. Our interest and examples are motivated by a desire to explore possible future changes in the distributions of meteorological variables due to climate change. For example, there is much interest in how extremes in the tails of the distribution (e.g., the 90th and higher percentiles) might change in future climates (Watson and Core Writing Team 2001). Changes in such quantities are likely to have greater societal impact than changes in the mean of the distribution (e.g., Beniston et al. 2005, manuscript submitted to *Climatic Change*).

Two distributions can be compared graphically by plotting estimates of the density functions, such as histograms, or by quantile–quantile plots. Inferences about the similarity of distributions of weather variables have generally been made with parametric statistical tests, such as the *t* test for equality of means, or the *F* test for equality of variances; see von Storch and Zwiers (2001) or Wilks (1995) for details. However, such tests rely on strong distributional assumptions to which their performance can be sensitive, they give only a limited view of how the distributions differ when more detail can be useful, and their implementation in the presence of serial dependence (e.g., correlation) can be troublesome. The *t* test for example is unable to detect changes in scale, the *F* test is unable to detect changes in location, and both can have low power if the distributions are not normal (e.g., Wilcox 1997, chapter 5). This study describes a simple technique that depicts how two distributions differ and that can be used to assess whether or not the difference can be characterized by a change in *location* or *scale*, often measured, respectively, by sample means and variances. The technique is flexible and can be tailored to focus on parts of the distribution, such as the extreme lower or upper tails, that are of specific interest, and it takes proper account of possible temporal and spatial dependence within and between the two series.

The technique is described in section 2 and demonstrated with an application in section 3. The example

Corresponding author address: C. Ferro, Department of Meteorology, University of Reading, Earley Gate, P.O. Box 243, Reading RG6 6BB, United Kingdom.
E-mail: c.a.t.ferro@reading.ac.uk

application examines the changes in distributions of daily minimum temperature and daily precipitation throughout Europe over the twenty-first century, as simulated by the Danish Meteorological Institute's high-resolution (50-km grid) regional climate model HIRHAM4 (Christensen et al. 1998) for the European Union project, Prediction of Regional Scenarios and Uncertainties for Defining European Climate Change Risks and Effects (PRUDENCE; Christensen et al. 2002). The simulations comprise a control (1961–90) and a scenario (2071–2100) integration, the latter forced by the Intergovernmental Panel on Climate Change (IPCC) A2 emissions scenario (Nakićenović and Swart 2000). Boundary conditions are supplied by the Hadley Centre's global, atmosphere-only model HadAM3H, which is driven by observed sea ice and sea surface temperature (HadISST1) in the control; sea ice and sea surface temperatures in the scenario are determined from changes simulated by the Third Hadley Centre Coupled Ocean–Atmosphere General Circulation Model (HadCM3). See the PRUDENCE project Web site at http://prudence.dmi.dk for more details.

2. Statistical method

a. Hypotheses

The aim is to understand any differences between the probability distributions of two variables, such as daily maximum temperatures at two sites. Simple characterizations of the possibly complex differences are often able to capture the main features and aid understanding. Two such characterizations are of particular interest: differences in location or scale, for which the distributions are related by a constant translation or scaling, respectively.

Let X and Y denote the two variables, and let their distribution functions be $F(x) = P(X \le x)$ and $G(y) = P(Y \le y)$, where $P(A)$ denotes the probability of an event A. See von Storch and Zwiers (2001) or Wilks (1995) for basic introductions to probability distributions. The following hypotheses are of interest for understanding changing distributions:

$$\begin{aligned} H_0&: \quad F(z) = G(z) \\ H_S&: \quad F(\sigma_X z) = G(\sigma_Y z) \\ H_L&: \quad F(\mu_X + z) = G(\mu_Y + z) \\ H_{LS}&: \quad F(\mu_X + \sigma_X z) = G(\mu_Y + \sigma_Y z) \end{aligned} \tag{1}$$

for all $-\infty < z < \infty$ and unknown constants μ_X, μ_Y, $\sigma_X > 0$, and $\sigma_Y > 0$. Hypothesis H_0 claims no difference between F and G, H_S a difference only in scale, H_L a difference only in location, and H_{LS} a difference only in location and scale. The relative impacts of location and scale changes have been discussed using parametric approaches by Mearns et al. (1984) and Katz and Brown (1992) among others.

In the remainder of this section, functions of quantiles that summarize the differences between F and G are defined, and simple, informative plots are described that support an informal assessment of the legitimacy of hypotheses (1). Methods are presented for computing confidence intervals to represent the variability of the quantile estimators, and formal testing of the hypotheses is discussed.

b. Quantiles

The p quantile ($100p$ percentile) of a continuous distribution is the value below which a proportion p of the probability mass falls. For example, the p quantile, x_p, of F satisfies $F(x_p) = p$. If $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ are samples from distributions F and G, and $X_{(1)} \le \ldots \le X_{(m)}$ and $Y_{(1)} \le \ldots \le Y_{(n)}$ are the order statistics (samples arranged in increasing order), then estimators for the p quantiles of F and G are

$$\hat{x}_p = X_{(\lfloor pm+0.5 \rfloor)} \quad \text{and} \quad \hat{y}_p = Y_{(\lfloor pn+0.5 \rfloor)},$$

where $\lfloor z \rfloor$ denotes the integer part of z. Different quantile estimators might be preferred if the sample sizes are small (Parrish 1990). See Bonsal et al. (2001) for an application examining changes in temperature quantiles and Wilcox (1997) for more material on such nonparametric statistics.

Three useful statistics for summarizing a distribution are the median, interquartile range, and Yule–Kendall skewness measure, which are computed from just three quantiles:

$$\begin{aligned} m_X &= \hat{x}_{0.5}, \\ s_X &= \hat{x}_{0.75} - \hat{x}_{0.25}, \\ a_X &= (\hat{x}_{0.75} - 2\hat{x}_{0.5} + \hat{x}_{0.25})/s_X. \end{aligned} \tag{2}$$

These statistics are resistant measures of the location, scale, and *shape* (asymmetry) of F, and can be compared with the corresponding statistics, m_Y, s_Y, and a_Y, for G. See Lanzante (1996) for examples demonstrating the benefits of using such measures.

For illustration, these statistics are now computed from all of the winter daily minimum temperatures simulated by HIRHAM4 at a single grid point (46.4805°N, 7.9761°E) in both the control (X) and scenario (Y) 30-yr integrations. Winter is defined to cover December–January–February (DJF), yielding sample sizes $m = n = 2700$. In the control, $m_X = -12.2$°C,

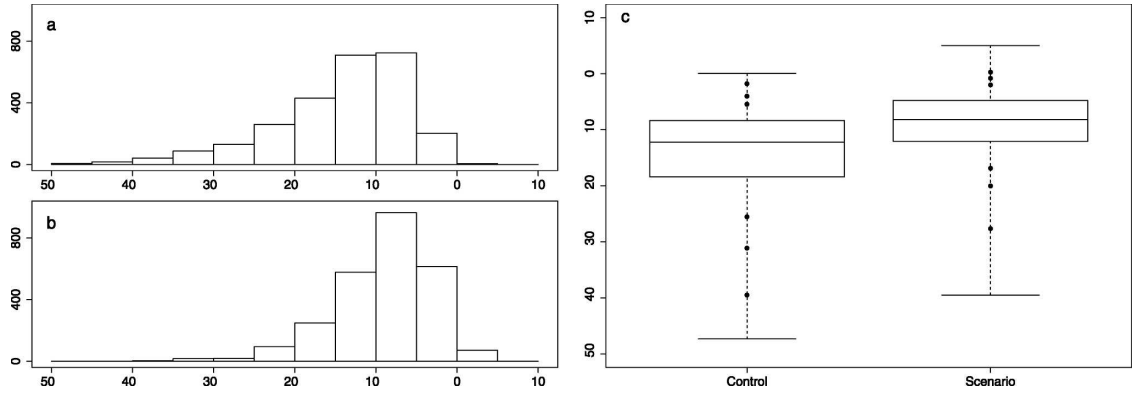

FIG. 1. Histograms of (a) control and (b) scenario DJF daily minimum temperatures (°C); (c) boxplots of control and scenario temperatures. The boxplot whiskers extend over the range of the data; three lower quantiles ($p = 0.01$, 0.05, and 0.1) and three upper quantiles ($p = 0.9$, 0.95, and 0.99) are marked (●).

$s_X = 10.1$°C, and $a_X = -0.23$; in the scenario, $m_Y = -8.2$°C, $s_Y = 7.3$°C, and $a_Y = -0.06$. Histograms and boxplots of the gridpoint temperatures are reproduced in Fig. 1. Note that the skewness measures compare the relative heights of the lower and upper boxes in the boxplots. These statistics and plots indicate a general warming together with a reduction in scale and a change in shape of the distribution: the long, colder tail evident in the control becomes shorter, resulting in a more symmetric distribution in the scenario. Similar behavior has been noted in observations such as the Central England Temperature series (Antoniadou et al. 2001).

Another informative comparison is made by the quantile–quantile plot of $\hat{y}_p$ against $\hat{x}_p$ for $p = 1/N$, $2/N, \ldots, 1$, where $N = \min(m, n)$. The hypotheses (1) correspond to different linear relationships between the two sets of quantiles:

$$\begin{aligned} H_0&: \quad y_p = x_p \\ H_S&: \quad y_p = \sigma_Y(x_p/\sigma_X) \\ H_L&: \quad y_p = \mu_Y + (x_p - \mu_X) \\ H_{LS}&: \quad y_p = \mu_Y + \sigma_Y(x_p - \mu_X)/\sigma_X \end{aligned} \tag{3}$$

for all $0 < p < 1$. The right-hand sides of equalities H_S, H_L, and H_{LS} are the quantiles for the distribution obtained by adjusting F to have, respectively, the same scale, location, and location and scale as G.

The quantile–quantile plot of the scenario versus control gridpoint temperatures is reproduced in Fig. 2a. Estimates of the linear relationships (3) for hypotheses H_0, H_L, and H_{LS} are superimposed, where the location parameters, μ_X and μ_Y, are estimated by the medians, m_X and m_Y, and the scale parameters, σ_X and σ_Y, by the interquartile ranges, s_X and s_Y. The estimated location-scale model for H_{LS} (dotted line) is reasonably close to

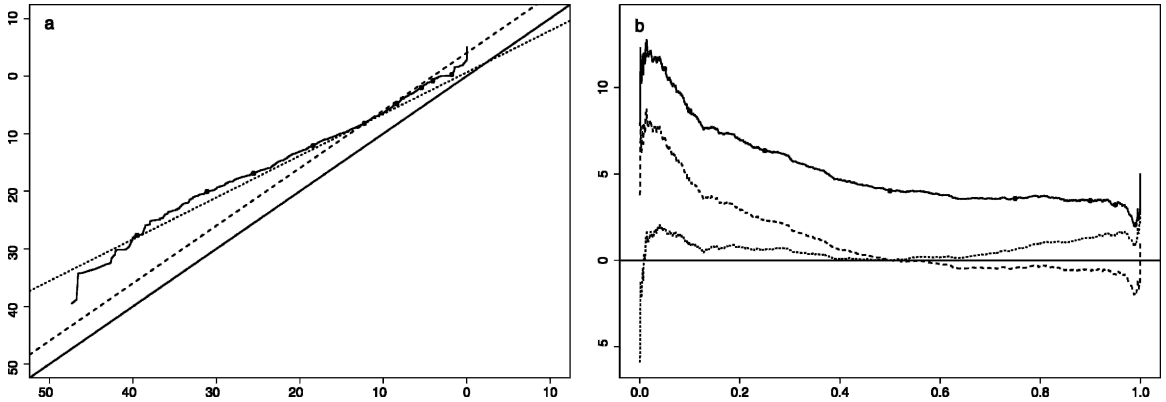

FIG. 2. (a) Quantile–quantile plot (line with dots) of scenario vs control DJF daily minimum temperatures (°C) with straight lines corresponding to hypotheses H_0 (normal line), H_L (dashed line), and H_{LS} (dotted line); (b) quantile differences (line with dots), location adjusted (dashed line), and location and scale adjusted (dotted line) against probability. Nine quantiles ($p = 0.01$, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, and 0.99) are highlighted (●).

the quantile–quantile plot except for some discrepancies at low and high temperatures. This indicates that, while the changes to the body of the temperature distribution are well described by a change in location and scale, a more complex model is required to describe the changes in the extreme tails of the distribution.

c. Exploring changes in gridded fields

The quantile–quantile plot is practicable if only a small number of pairwise comparisons are able to be made. With gridded data, comparing distributions between two time periods at each of several thousand grid points is often of interest. In this case, it is more valuable to plot maps showing the change in a single quantile, such as $\hat{y}_p - \hat{x}_p$, at each grid point. Values of p can be chosen to cover different parts of the distribution: for the center, $p = 0.25$, 0.5, and 0.75 might suffice; for the lower or upper tail, $p = 0.01$, 0.05, and 0.1 or $p = 0.9$, 0.95, and 0.99 could be used. All nine of these quantiles will be examined in our application. The 0.1 and 0.9 quantiles correspond to the Watson and Core Writing Team (2001) definition of an extreme event; the rarer quantiles provide more information about the tails. Sample size will dictate how far into the tails quantile estimators are acceptably precise.

If there is no difference between F and G (hypothesis H_0), then from (3) $\hat{y}_p - \hat{x}_p$ is expected to be approximately zero for each p. If maps of these quantile differences show nonzero values, then a simple explanation could be a change in location. If hypothesis H_L and the corresponding relationship (3) hold, then the estimators

$$\hat{y}_p - \{m_Y + (\hat{x}_p - m_X)\}$$

for the *location-adjusted* quantile differences are expected to be zero, and maps of these differences are useful for diagnosing a location shift. If significant patterns still remain, then it is possible to look for an additional change in scale. If hypothesis H_{LS} and the corresponding relationship (3) hold, then the estimators

$$\hat{y}_p - \left\{m_Y + s_Y\left(\frac{\hat{x}_p - m_X}{s_X}\right)\right\}$$

for the *location- and scale-adjusted* quantile differences are expected to be zero, and maps of these differences are useful for diagnosing location and scale shifts. Careful scrutiny of such maps can lead to a good understanding of how distributions differ at each grid point, and how these differences vary geographically.

For illustration, the quantile differences between the control and scenario temperatures at the example grid point are shown in Fig. 2b. The location- and scale-adjusted differences confirm the earlier finding that model H_{LS} provides a reasonable fit except in the tails. See Beniston and Stephenson (2004) and McGregor et al. (2005) for other applications of this technique.

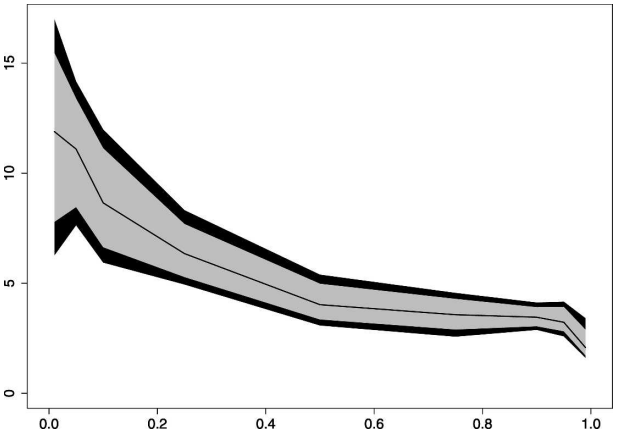

FIG. 3. Quantile differences (thin line) of DJF daily minimum temperatures (°C) against probability, with pointwise (gray) and simultaneous (black) 90% confidence intervals.

d. Confidence intervals

Sample estimates $\hat{d}_p = \hat{y}_p - \hat{x}_p$ differ from the true differences $d_p = y_p - x_p$ due to sampling uncertainty. This can be quantified by calculating confidence intervals for d_p. The interval $[L_p, U_p]$ is a *pointwise* $(1 - \alpha)$ confidence interval for d_p if

$$P(L_p \le d_p \le U_p) = 1 - \alpha. \tag{4}$$

The probability that the interval contains the true quantile difference is $1 - \alpha$. See von Storch and Zwiers (2001) for background on confidence intervals and other aspects of statistical inference.

There are many ways to construct confidence intervals. The approach employed here is based on bootstrap resampling, a popular and effective technique that can be adapted to account for dependence within and between the samples. Dunn (2001), for example, uses the bootstrap to estimate pointwise confidence intervals for rainfall quantiles from a single sample; see also Wilks (1995, 1997). A discussion of bootstrap confidence intervals and their implementation in the current setting is deferred to the appendixes. Pointwise confidence intervals for the quantile differences between the control and scenario temperatures at the example grid point are shown in Fig. 3, where it can be seen that the uncertainty is greater in the relatively long, colder tail.

The hypotheses (1) refer not to single quantiles but to entire distributions: if hypothesis H_0 holds, for ex-

ample, then $d_p = 0$ for all p. Suppose that limits L'_p and U'_p are available for each of M values, $p_1, \ldots, p_M$, of p such that

$$P(L'_p \le d_p \le U'_p \quad \text{for all} \quad p = p_1, \ldots, p_M) = 1 - \alpha. \tag{5}$$

If the pointwise limits L_p and U_p satisfying expression (4) are used to define these *simultaneous* confidence intervals, then the coverage probability $1 - \alpha$ is unlikely to be obtained in (5). For example, if the different quantiles were independent, then the pointwise limits would give

$$P(L_p \le d_p \le U_p \quad \text{for all} \quad p = p_1, \ldots, p_M) = \prod_{k=1}^{M} P(L_{p_k} \le d_{p_k} \le U_{p_k}) = (1 - \alpha)^M.$$

Although independence is unrealistic, it remains true that simultaneous intervals are generally wider than pointwise intervals with the same coverage. The method used here to construct simultaneous intervals is described in appendix A.

e. Hypothesis tests

The simultaneous intervals (5) can be used to test H_0: the hypothesis is rejected at significance level α unless $L'_p \le 0 \le U'_p$ for all p. Rather than use all $p_1 = 1/N$, $p_2 = 2/N, \ldots, p_M = p_N = 1$, attention can be restricted to a subset of quantiles, the choice of which involves a trade-off between statistical power and the strength of conclusions. Power reduces as more quantiles are considered because the simultaneous intervals widen, and a real change in distribution is less likely to be detected. On the other hand, a hypothesis test based on only a few quantiles ignores possible changes in other parts of the distribution. Our selected set of nine quantiles ($p =$ 0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, and 0.99) form a compromise, but other choices may be more appropriate in other applications. A small simulation study of the power of tests based on these quantiles is summarized in appendix C.

Simultaneous confidence intervals for the quantile differences between the control and scenario temperatures at the example grid point are shown in Fig. 3. The line $d_p = 0$ lies outside the shaded simultaneous 90% confidence intervals, which therefore support rejection of H_0 at the 10% level of significance.

One might attempt to test H_L by constructing confidence intervals for $y_p - \{\mu_Y + (x_p - \mu_X)\}$, which is zero for all p if H_L holds, by substituting it for d_p in the previous section. This is successful only if μ_X and μ_Y are specified. For example, choosing $\mu_X = x_{0.5}$ and $\mu_Y = y_{0.5}$ produces confidence intervals for $y_p - \{y_{0.5} + (x_p - x_{0.5})\}$. Such intervals, however, do not support a test of H_L because H_L leaves μ_X and μ_Y unspecified: even if the confidence intervals for $y_p - \{y_{0.5} + (x_p - x_{0.5})\}$ failed to contain zero for all p, other choices for μ_X and μ_Y might yield a different conclusion. Similar considerations apply to confidence intervals for $y_p - \sigma_Y x_p/\sigma_X$ and $y_p - \{\mu_Y + \sigma_Y(x_p - \mu_X)/\sigma_X\}$.

Some authors (e.g., Sun et al. 2001) have proposed rejecting H_L if a horizontal line cannot pass completely through the confidence band for d_p, that is, if $\max_p L'_p > \min_p U'_p$. Such a test is conservative, however: if H_L holds, then the band will contain the true value of $\mu_Y - \mu_X$ with the appropriate probability, but the probability that it contains any constant line is greater, so H_L will be rejected too infrequently. Parametric models can be used to test for specific departures from H_L, H_S, or H_{LS}, but such tests can be sensitive to model assumptions, as mentioned in section 1.

Our preferred approach for testing hypotheses H_0, H_L, H_S, and H_{LS} would be first to test H_{LS} against the general alternative that there are some differences between the distributions that cannot be described by location and scale changes, then H_L or H_S against H_{LS} if the first test were passed, and finally H_0 against H_L or H_S if the second test were passed. Several nonparametric procedures exist for the latter two tests when there is no serial dependence: see Conover et al. (1981) and Lehmann (1975, p. 95) for example. We have failed to find any published nonparametric tests for H_{LS} against the general alternative, so we are investigating elsewhere the use of minimized distance measures, such as the Kolmogorov–Smirnov distance and the quantile distance considered by Zhang and Yu (2002), as test statistics.

3. Temperature and precipitation fields

The use of quantiles and adjusted quantile differences for diagnosing distributional changes was illustrated in the previous section with data at a single grid point from the HIRHAM4 integrations. In this section, the methods are applied across the entire spatial domain, first for winter daily minimum temperatures and then for winter daily total precipitation.

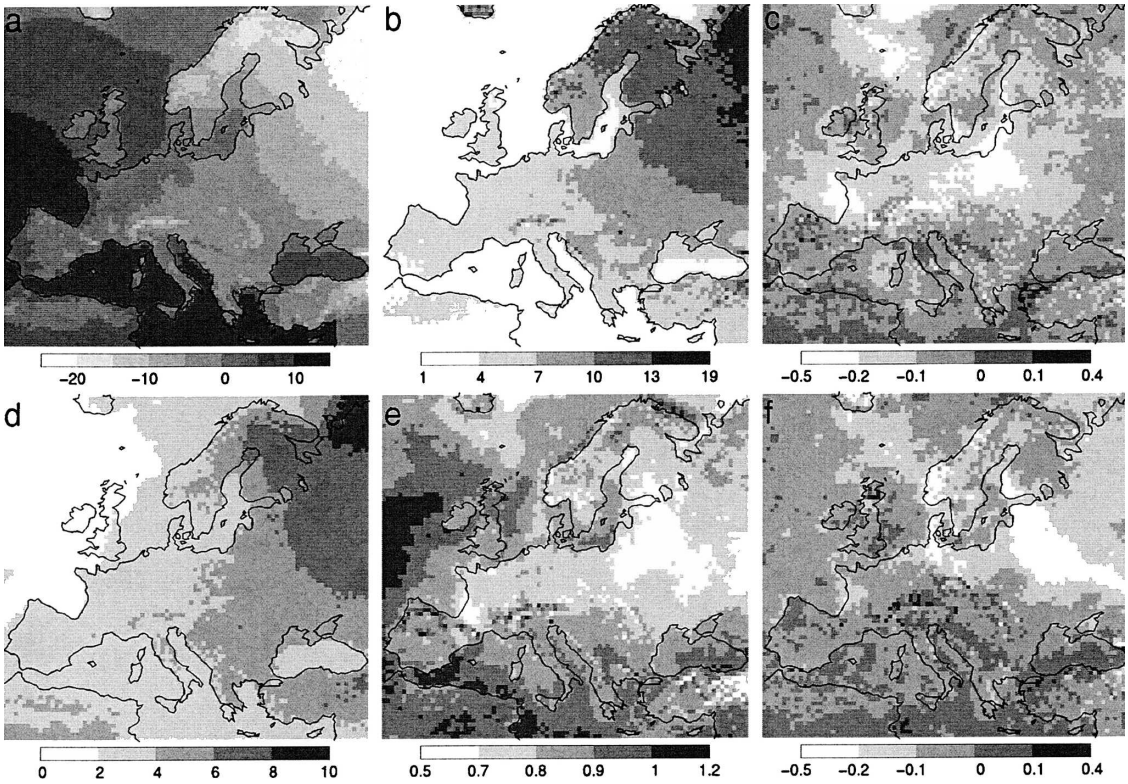

FIG. 4. (a) Median (°C), (b) interquartile range (°C), and (c) skewness measure for DJF daily minimum temperatures in the control; (d) differences in the medians and (e) ratios of the interquartile ranges between the scenario and control; (f) skewness measure in the scenario.

a. Temperature

The summary statistics (2) are displayed in Fig. 4. In the control, cooler median temperatures are found over northeastern Europe, regions that also exhibit greater variability, and there is widespread negative skewness. The uniform increase in scenario median temperatures is greater in cooler regions, the general decrease in variability is greatest in the continental interior, and skewness moves closer to zero except, most noticeably, for a band in the east that may correspond to snow retreat (Kjellström 2004).

That the probability distribution of scenario temperatures is not everywhere merely a location shift of the control distribution is evident in Fig. 5, where the greater increase in cold quantiles, particularly in the continental interior, is clear. Figure 6, showing the location-adjusted quantile differences, indicates that the distributional changes over much of the seas and western Europe can be described by a shift in location. Figure 7 indicates that additional changes in scale describe some of the remaining changes in northern and eastern Europe, but that some regions exhibit a more complex distributional change.

b. Precipitation

Location shifts may be inappropriate descriptions of changes in distributions of nonnegative variables such as precipitation. A positive location shift would exclude values near zero; a negative shift would admit negative values. Although common measures of location such as the median may still change, such effects are better described by shifts in scale and shape. Distinguishing between wet and dry days, the former denoting days on which the precipitation strictly exceeds an amount u, may also be important.

Let F and G denote the distribution functions for *excess* wet-day precipitation above $u = 1$ mm in the control and scenario integrations. Let also $\hat{x}_p$ and $\hat{y}_p$ be the estimators for the p quantiles of F and G. The scale- and shape-change hypothesis is formulated as

$$H_{SS} \quad F(\sigma_X z^{\alpha_X}) = G(\sigma_Y z^{\alpha_Y})$$

for all $z > 0$ and unknown, positive constants σ_X, σ_Y, α_X, and α_Y. This hypothesis corresponds to the nonlinear relationship

$$y_p = \sigma_Y (x_p/\sigma_X)^{\alpha_Y/\alpha_X}.$$

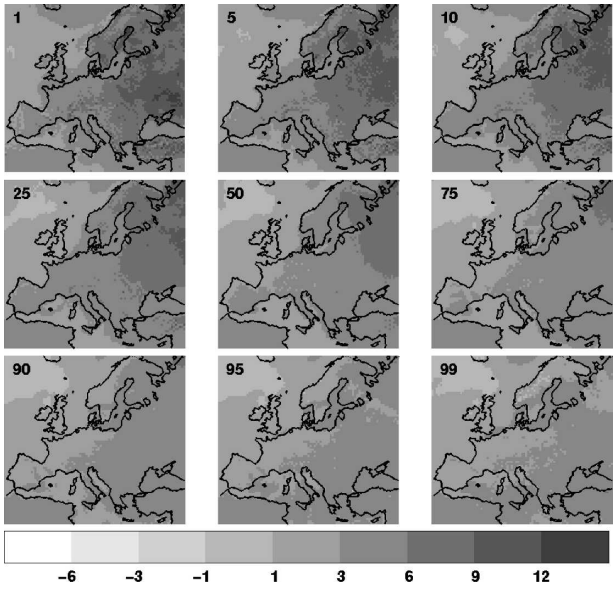

FIG. 5. Differences (°C) in nine quantiles of DJF daily minimum temperatures between the scenario and control.

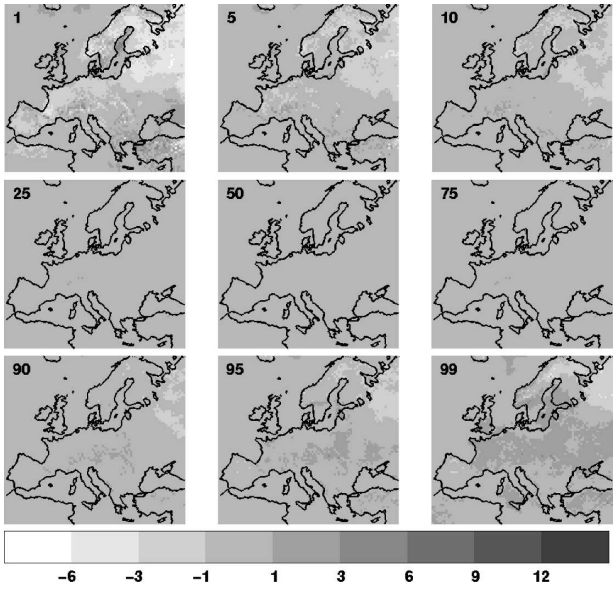

FIG. 7. Same as in Fig. 5, after adjusting for location and scale.

Taking logarithms yields

$$y_p^* = \log\sigma_Y + \alpha_Y(x_p^* - \log\sigma_X)/\alpha_X,$$

where $x_p^* = \log x_p$ and $y_p^* = \log y_p$. This has the same form as the final equality (3) and shows that a scale-shape change of wet-day precipitation excess is equivalent to a location-scale change of log-transformed excess. Adjusted quantile differences can therefore be defined as in section 2c. For example, if hypothesis H_{SS} holds, then the estimators

$$\hat{y}_p^* - \left\{ m_Y^* + s_Y^*\left(\frac{\hat{x}_p^* - m_X^*}{s_X^*}\right)\right\}$$

for the location- and scale-adjusted transformed quantile differences are expected to be zero, where $m_X^* = \hat{x}_{0.5}^*$, $s_X^* = \hat{x}_{0.75}^* - \hat{x}_{0.25}^*$, and m_Y^* and s_Y^* are defined similarly. Inverting the transformation yields *scale- and shape-adjusted* quantile ratios:

$$\frac{\hat{y}_p}{m_Y(\hat{x}_p/m_X)^{s_Y^*/s_X^*}}.$$

Maps of these scale- and shape-adjusted quantities are useful for diagnosing scale and shape changes. Setting $s_X^* = s_Y^*$ yields quantities appropriate for diagnosing a pure scale change.

These methods are illustrated by application to winter daily total precipitation from the control and scenario integrations. Summary statistics are reproduced in Fig. 8. In the control, greater median precipitation amounts are found windward of steep altitude gradients, regions that also exhibit greater variability, and there is widespread positive skewness. Scenario median precipitation decreases only in the Mediterranean and over the Scandinavian mountains, a pattern that is replicated for scale, but the spatially complex changes in skewness are difficult to summarize. The change in proportion of wet days is shown in Fig. 9, revealing a decrease in the Mediterranean and in the far north, and an

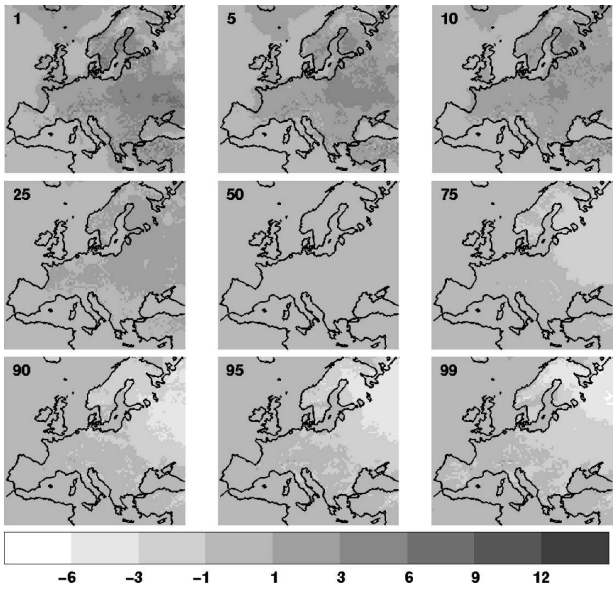

FIG. 6. Same as in Fig. 5, but after adjusting for location.

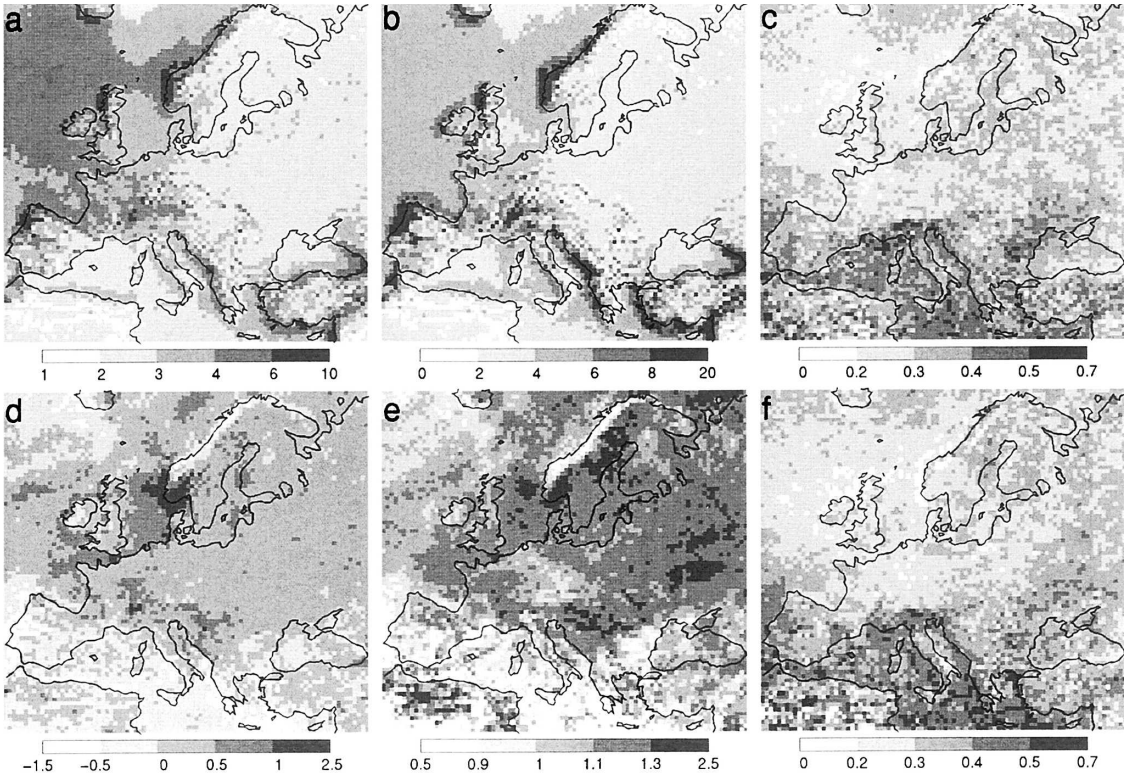

FIG. 8. Same as in Fig. 4, but for DJF daily total precipitation; (a) and (b) interquartile range are in mm.

increase in the intervening latitudes that is strongest around the North and Baltic Seas.

The changes in six quantiles are investigated in Fig. 10. (Low quantiles are of less interest so they are excluded.) All quantiles decrease in the Mediterranean and over the Scandinavian mountains. Largest increases are found elsewhere in Scandinavia and in eastern Europe. The scale-adjusted quantile ratios in Fig. 11 indicate that changes in scale of the distributions can explain many of the differences over northern Europe, the Alps, and around the Adriatic. Accounting for additional changes in shape (not shown) explained little of the remaining differences.

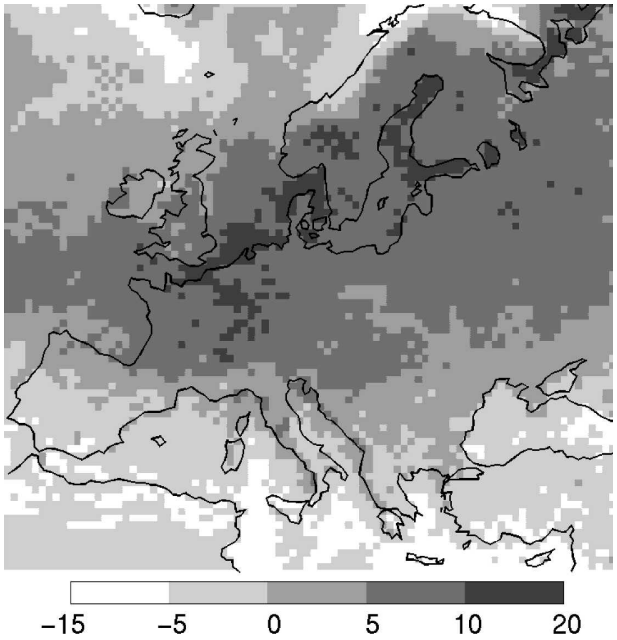

FIG. 9. Difference in the proportions (%) of wet days between the scenario and control.

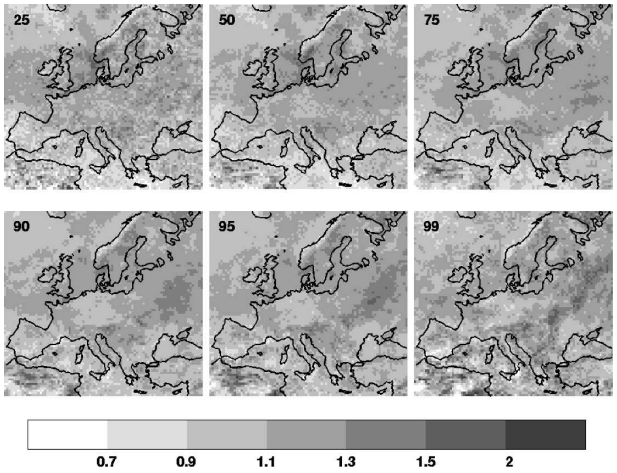

FIG. 10. Ratios of six quantiles of DJF daily total precipitation between the scenario and control.

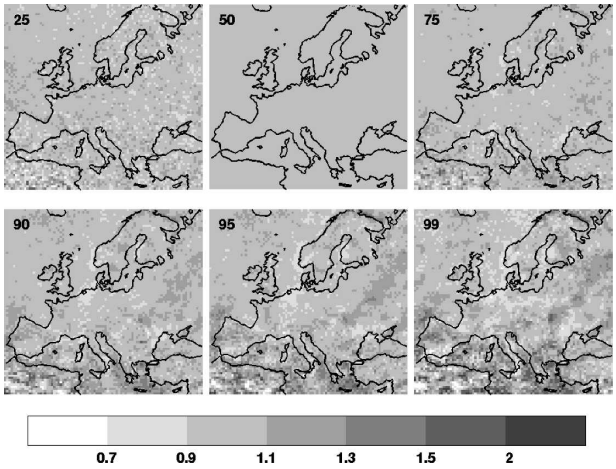

FIG. 11. Same as in Fig. 10, but after adjusting for scale.

4. Discussion

The methods presented here are simple and flexible tools for comparing entire distributions of meteorological variables. Such investigations are able to highlight differences, such as changes in the tails of a distribution, that have important, practical consequences and that could be missed by examining only means and variances. The application considered in section 3 demonstrates the ability of the methods to highlight the main features in large, gridded datasets, a situation in which traditional, graphical comparisons would be impracticable. The analysis can also be used to motivate a closer examination of sites that exhibit complex distributional changes.

Acknowledgments. We thank Simon Brown, Ole Christensen, and Erik Kjellström for encouraging our investigation of the methods presented here. DBS wishes to thank IPCC for the opportunity to present early ideas on this work at the IPCC workshop on extremes held from 11 to 13 June 2002 in Beijing. The work was supported by the European Union Programme Energy, Environment and Sustainable Development under Contract EVK2-CT-2001-00132 (PRUDENCE).

APPENDIX A

Bootstrap Confidence Intervals

This appendix first describes a method for obtaining pointwise confidence limits with coverage probability (4). If

$$P(l \le \hat{d}_p - d_p \le u) = 1 - \alpha,$$

then $L_p = \hat{d}_p - u$ and $U_p = \hat{d}_p - l$ are valid confidence limits. Requiring

$$P(\hat{d}_p - d_p < l) = P(\hat{d}_p - d_p > u) = \alpha/2 \quad \text{(A1)}$$

ensures equal-tailed intervals, a useful property that highlights any asymmetry in the uncertainty associated with $\hat{d}_p$, but does not necessarily yield the shortest interval. Bootstrap estimates of l and u can be obtained in the following way. Let $\{X_1^*, \ldots, X_m^*\}$ and $\{Y_1^*, \ldots Y_n^*\}$ be samples formed by resampling the original sequences. Discussion of how to perform this resampling is postponed to appendix B. When the original samples are representative of the populations from which they were drawn, the distribution of $\hat{d}_p - d_p$ is well approximated by that of $\hat{d}_p^* - \hat{d}_p$, where $\hat{d}_p^* = \hat{y}_p^* - \hat{x}_p^*$, $\hat{x}_p^* = X^*_{(\lfloor pm+0.5 \rfloor)}$ and $\hat{y}_p^* = Y^*_{(\lfloor pn+0.5 \rfloor)}$. The distribution of $\hat{d}_p^* - \hat{d}_p$ can be approximated numerically by creating a large number, B, of resamples: if $\hat{d}_{p1}^*, \ldots, \hat{d}_{pB}^*$ are the values of $\hat{d}_p^*$ for the B resamples, and $\hat{d}_{p(1)}^* \le \ldots \le \hat{d}_{p(B)}^*$ are the order statistics, then estimates of l and u satisfying equalities (A1) are

$$l^* = \hat{d}_{p(b_1)}^* - \hat{d}_p \quad \text{and} \quad u^* = \hat{d}_{p(b_2)}^* - \hat{d}_p,$$

where $b_1 = \lfloor (\alpha/2)\, B + 0.5 \rfloor$ and $b_2 = \lfloor (1 - \alpha/2)\, B + 0.5 \rfloor$. Taking $B = 1000$ typically yields a sufficiently close approximation. The resulting confidence limits,

$$L_p = \hat{d}_p - u^* \quad \text{and} \quad U_p = \hat{d}_p - l^*,$$

define what is known as a "basic" bootstrap confidence interval.

However, there are technical reasons and evidence from simulation studies (Falk and Kaufmann 1991) that another type of bootstrap interval should be preferred for constructing confidence intervals for quantiles. Suppose that there exists a function $h(\cdot)$ such that the distribution of $\hat{c}_p = h(\hat{d}_p)$ is symmetric about $c_p = h(d_p)$. Estimates of l and u that satisfy $P(l \le \hat{c}_p - c_p \le u) = 1 - \alpha$ can then be found as before: $l^* = \hat{c}^*_{p(b_1)} - \hat{c}_p$ and $u^* = \hat{c}^*_{p(b_2)} - \hat{c}_p$. By symmetry, however, it is also the case that $P(l \le c_p - \hat{c}_p \le u) = 1 - \alpha$, so valid confidence limits for c_p are $L_p = l^* + \hat{c}_p = \hat{c}^*_{p(b_1)}$ and $U_p = u^* + \hat{c}_p = \hat{c}^*_{p(b_2)}$. Applying the inverse transformation yields "percentile" bootstrap confidence limits for d_p:

$$L_p = \hat{d}^*_{p(b_1)} \quad \text{and} \quad U_p = \hat{d}^*_{p(b_2)}.$$

Note that the symmetrizing function $h(\cdot)$ is not used, and so need not be known, to compute these limits.

Bootstrapping confidence intervals for quantiles is mathematically sound (e.g., DiCiccio and Romano 1988) but is more difficult than for many other quantities, such as means, in the sense that larger sample sizes are required to obtain the same level of accuracy. Sev-

eral modifications of the bootstrap, such as smoothing (Hall et al. 1989), have been proposed to improve matters, but these are more complicated to implement and are not considered here.

Simultaneous intervals (5) can be estimated using a method described by Davison and Hinkley (1997, their section 4.2.4). From the B bootstrap samples obtained previously, compute $\hat{d}^*_{p1}, \ldots, \hat{d}^*_{pB}$ for each p of interest. Equal-tailed, simultaneous confidence intervals have limits $L'_p = \hat{d}^*_{p(k)}$ and $U'_p = \hat{d}^*_{p(B+1-k)}$ for some $1 \leq k \leq B/2$. For any k, the bootstrap estimate of the coverage probability (5) is

$$\frac{1}{B}\sum_{b=1}^{B} I(\hat{d}^*_{pb} \leq \hat{d}^*_{p(k)} \quad \text{or} \quad \hat{d}^*_{pb} \geq \hat{d}^*_{p(B+1-k)} \quad \text{for at least one } p),$$

where $I(A) = 1$ when A is true, and 0 when A is false. It is sufficient, therefore, to choose k such that this estimate is as close as possible to $1 - \alpha$. This value can be found with an appropriate search routine.

APPENDIX B

Bootstrap Resampling

When the original samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ are independent of one another, and each comprises independent and identically distributed variables, bootstrap resampling is straightforward: new samples $\{X^*_1, \ldots, X^*_m\}$ and $\{Y^*_1, \ldots, Y^*_n\}$ are formed by resampling uniformly and with replacement from the appropriate, original sample.

If there is dependence between or within the original samples, then this must be reproduced in the resamples for the bootstrap approximation to be accurate. Dependence between samples can be preserved by resampling Y_i whenever X_i is chosen, for $1 \leq i \leq \min(m, n)$. This would be appropriate if X_i and Y_i were coincident (paired) measurements at two sites, for example. Several approaches have been developed to account for serial dependence within samples, two of which (prewhitening and moving-blocks resampling) are discussed by Wilks (1997). In our application, each sample is a time series of daily gridpoint values for consecutive winters. If dependence between winters is weak, then resampling data blocked into winters is an acceptable solution. To be precise, each winter comprises $r = 90$ values and the X_i form blocks $Z_j = \{X_{(j-1)r+1}, \ldots, X_{jr}\}$ for winters $j = 1, \ldots, m/r$. We resample uniformly and with replacement from $\{Z_1, \ldots, Z_{m/r}\}$ to obtain $\{X^*_1, \ldots, X^*_m\}$, and similarly for the Y_i.

Another potential complication is the presence of time trends in the data. In this case, the stationarity assumption that each X_i has distribution F and each Y_i has distribution G is unreasonable. The methods described in this article are not designed for such data, so any trends should be removed prior to the analysis. In general, however, bootstrap techniques are easily modified to incorporate trends. For example, if $\{\hat{X}_1, \ldots, \hat{X}_m\}$ is an estimate of an additive trend in the X_i, then the detrended data, $\tilde{X}_i = X_i - \hat{X}_i$, should be resampled before adding back the estimated trend component to obtain $X^*_i = \tilde{X}^*_i + \hat{X}_i$.

APPENDIX C

Power Study

The power of the hypothesis test proposed in section 2 for H_0 and based on simultaneous, percentile bootstrap confidence intervals for nine quantiles (0.01, 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95, and 0.99) with 1000 bootstrap samples is compared with the power of the Kolmogorov–Smirnov test, the two-sample t test, and the F test.

Monte Carlo estimates of power are obtained from 1000 simulated datasets. Each dataset comprises two, independent samples of 300 independent normal random variables. Serial dependence is suppressed so that the t and F tests are applicable without adjustment (see, e.g., chapter 6 of von Storch and Zwiers 2001). The sample size is the effective size (von Storch and Zwiers 2001, p. 115) of samples of length 2700 generated by a first-order autoregressive process with correlation 0.8 at first lag. The power of the tests to detect changes in location is determined by setting the variances of the two samples equal to 1 and allowing the difference (δ) between the means to range from 0 to 0.5. The power of the tests to detect changes in scale is determined by setting the means of the two samples equal to 0 and allowing the ratio (ρ) of the standard deviations to range from 1 to 1.5.

The results are plotted in Fig. C1. The t and F tests are designed and are most powerful for detecting changes in, respectively, the means and variances of normal distributions. The quantile and Kolmogorov–Smirnov tests, on the other hand, are sensitive to any distributional changes and do not make any such distributional assumptions. The results show that the quantile test is conservative and less powerful than the Kolmogorov–Smirnov test for detecting location changes but is generally more powerful for detecting scale changes. Note that these results also give the powers of detecting changes in scale (exp δ) and shape (ρ)

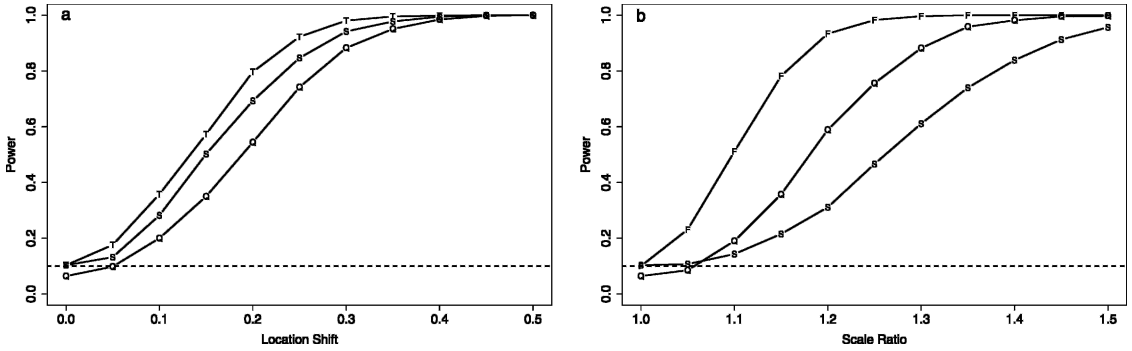

FIG. C1. (a) Powers of the *t* (T), Kolmogorov–Smirnov (S), and quantile (Q) tests for detecting changes in location; (b) powers of the *F* (F), Kolmogorov–Smirnov (S), and quantile (Q) tests for detecting changes in scale. The nominal significance level is marked (dashed line).

when the data have lognormal distributions and the tests are applied to the log-transformed data.

REFERENCES

Antoniadou, T., P. Besse, A. L. Fougères, C. Le Gall, and D. B. Stephenson, 2001: L'Oscillation Atlantique Nord (NAO) et son influence sur le climat Européen. *Rev. Stat. Appl.,* **49** (3), 39–60.

Beniston, M., and D. B. Stephenson, 2004: Extreme climatic events and their evolution under changing climatic conditions. *Global Planet. Change,* **44,** 1–9.

Bonsal, B. R., X. Zhang, L. A. Vincent, and W. D. Hogg, 2001: Characteristics of daily and extreme temperatures over Canada. *J. Climate,* **14,** 1959–1976.

Christensen, J. H., T. Carter, and F. Giorgi, 2002: PRUDENCE employs new methods to assess European climate change. *Eos, Trans. Amer. Geophys. Union,* **82,** 147.

Christensen, O. B., J. H. Christensen, B. Machenhauer, and M. Botzet, 1998: Very high-resolution regional climate simulations over Scandinavia—Present climate. *J. Climate,* **11,** 3204–3229.

Conover, W. J., M. E. Johnson, and M. M. Johnson, 1981: A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics,* **23,** 351–361.

Davison, A. C., and D. V. Hinkley, 1997: *Bootstrap Methods and Their Application.* Cambridge University Press, 592 pp.

DiCiccio, T. J., and J. P. Romano, 1988: A review of bootstrap confidence intervals. *J. Roy. Stat. Soc.,* **50B,** 338–354.

Dunn, P. K., 2001: Bootstrap confidence intervals for predicted rainfall quantiles. *Int. J. Climatol.,* **21,** 89–94.

Falk, M., and E. Kaufmann, 1991: Coverage probabilities of bootstrap-confidence intervals for quantiles. *Ann. Stat.,* **19,** 485–495.

Hall, P., T. J. DiCiccio, and J. P. Romano, 1989: On smoothing the bootstrap. *Ann. Stat.,* **17,** 692–704.

Katz, R. W., and B. G. Brown, 1992: Extreme events in a changing climate: Variability is more important than averages. *Climatic Change,* **21,** 289–302.

Kjellström, E., 2004: Recent and future signatures of climate change in Europe. *Ambio,* **33,** 19–24.

Lanzante, J. R., 1996: Resistant, robust and nonparametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.,* **16,** 1197–1226.

Lehmann, E. H., 1975: *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, 457 pp.

McGregor, G. R., C. A. T. Ferro, and D. B. Stephenson, 2005: Projected changes in extreme weather and climate events in Europe. *Extreme Weather Events and Public Health Responses,* W. Kirch, B. Menne, and R. Bertollini, Eds., Springer, 13–23.

Mearns, L. O., R. W. Katz, and S. H. Schneider, 1984: Extreme high-temperature events: Changes in their probabilities with changes in mean temperature. *J. Climate Appl. Meteor.,* **23,** 1601–1613.

Nakićenović, N., and R. Swart, Eds., 2000: *IPCC Special Report on Emission Scenarios.* Cambridge University Press, 599 pp.

Parrish, R. S., 1990: Comparison of quantile estimators in normal sampling. *Biometrics,* **46,** 247–257.

Sun, Y., S. Sun, and Y. Diao, 2001: Smooth quantile processes from right censored data and construction of simultaneous confidence bands. *Commun. Stat. Theory Methods,* **30,** 707–727.

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

Watson, R. T., and Core Writing Team, Eds., 2001: *Climate Change 2001: Synthesis Report.* Cambridge University Press, 398 pp.

Wilcox, R. R., 1997: *Introduction to Robust Estimation and Hypothesis Testing.* Academic Press, 296 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

——, 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate,* **10,** 65–82.

Zhang, Z., and Q. Yu, 2002: A minimum distance estimation approach to the two-sample location-scale problem. *Lifetime Data Anal.,* **8,** 289–305.