Avoiding Inhomogeneity in Percentile-Based Indices of Temperature Extremes

XUEBIN ZHANG

Climate Research Branch, Meteorological Service of Canada, Downsview, Ontario, Canada

GABRIELE HEGERL

Nicholas School for the Environment and Earth Sciences, Duke University, Durham, North Carolina

FRANCIS W. ZWIERS

Canadian Centre for Climate Modelling and Analysis, Victoria, British Columbia, Canada

JESSE KENYON

Nicholas School for the Environment and Earth Sciences, Duke University, Durham, North Carolina

(Manuscript received 27 May 2004, in final form 10 November 2004)

ABSTRACT

Using a Monte Carlo simulation, it is demonstrated that percentile-based temperature indices computed for climate change detection and monitoring may contain artificial discontinuities at the beginning and end of the period that is used for calculating the percentiles (base period). This would make these exceedance frequency time series unsuitable for monitoring and detecting climate change. The problem occurs because the threshold calculated in the base period is affected by sampling error. On average, this error leads to overestimated exceedance rates outside the base period. A bootstrap resampling procedure is proposed to estimate exceedance frequencies during the base period. The procedure effectively removes the inhomogeneity.

1. Introduction

Successive reports of the Intergovernmental Panel on Climate Change (IPCC) have made increasingly strong statements on the human influence on the global climate. Since the greatest impacts of climate change may result from the changes in the extremes, rather than in the mean, analyzing climate extremes becomes very important. Monitoring, detecting, and attributing changes in climate extremes requires daily resolution data. However, the compilation, provision, and update of a globally complete and readily available daily dataset is a very difficult task. This comes about, in part, because not all national meteorological and hydrometeorological services are able to freely distribute the daily data that they collect. Consequently, indicators of climate extremes have been developed (e.g., Karl et al. 1999; Peterson et al. 2001) in the hope that they will come to be more widely obtainable than these daily data from which they are derived. These indicators have been used to analyze changes in climate extremes for various parts of the world (e.g., Jones et al. 1999; Frich et al. 2002; Easterling et al. 2003; Peterson et al. 2002; Klein Tank and Können 2003; Kiktev et al. 2003).

Several temperature indicators are calculated by counting the number of days in a year, or season, for which daily values exceed a time-of-year-dependent threshold. Such a threshold is usually defined as a percentile of daily observations in a fixed base period that fall within a few Julian days of the day of interest. For easy comparison of indices across stations with records of various lengths, and for easy update once new daily data are available, the thresholds are usually computed from a common base period, such as 1961–90, for all stations.

Folland et al. (1999) provisionally recommended a three-step procedure for the estimation of the thresholds: 1) remove the annual cycle by extracting the 30-yr

Corresponding author address: Dr. Xuebin Zhang, Climate Monitoring and Data Interpretation Division, Climate Research Branch, Meteorological Service of Canada, 4905 Dufferin Street, Downsview, Ontario M3H 5T4, Canada.
E-mail: Xuebin.Zhang@ec.gc.ca

mean values of each calendar day, 2) fit a probability distribution (such as the three-parameter gamma distribution) to the daily anomalies for each Julian day, and 3) compute the thresholds from the fitted probability distributions. Folland et al. (1999) also recommended that data from additional proximate calendar days be added to improve the stability of the probability distribution parameter estimates but that those days should be far enough apart such that data from different days are effectively independent. This method was implemented in Jones et al. (1999), who used five observations with 5-day intervals between them (referred to as the 5SD window hereafter). In many other applications (e.g., Frich et al. 2002; Klein Tank and Können 2003; Kiktev et al. 2003), thresholds have been estimated using data from five consecutive days centered on the day of interest (referred to as 5CD). In either case, the daily thresholds are, in effect, percentiles estimated from samples of no more than $5 \times 30 = 150$ days of data when a standard 30-yr base period is used.

Despite the importance of these indicators in the detection and monitoring of climate change, their statistical properties have not been well documented. For example, what differences would result in the index time series when 5CD and 5SD windows are used? Does the fact that the thresholds are "adapted" to (calculated from) the base period cause any systematic differences between the statistical properties of the index time series during the base period (in base) and before or after the base period (out of base)? Such differences need to be understood before the indices can be used with confidence for the purpose of climate change detection and monitoring.

The main objective of this paper is to examine, through Monte Carlo simulations, the characteristics of the index time series that are obtained when threshold functions are estimated with existing methods. We show that these threshold estimation methods produce substantial inhomogeneities in the index time series at the beginning and end of the base period in the sense that inhomogeneities become clearly apparent when a large number of station series are averaged (Fig. 1) as might be done in a climate change detection study. We propose an approach that corrects the problem. The remainder of this paper is organized as follows. We describe existing methods for calculating thresholds and index time series in section 2. The Monte Carlo experiment that is used to study the performance of these methods is also described in this section. Results are presented in section 3. An improved method for calculating the index time series is described and evaluated in section 4. Conclusions and discussion follow in section 5.

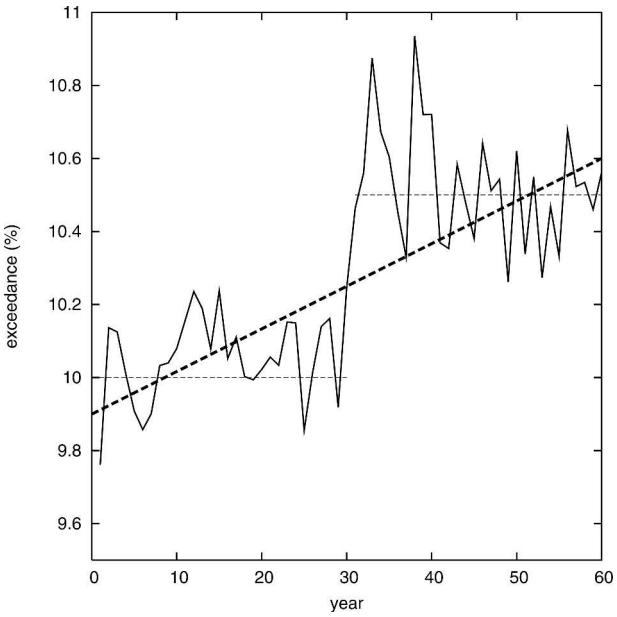

FIG. 1. Average of exceedance rate of daily values greater than the 90th percentile in 1000 simulations in which the lag 1-day autocorrelation has been set to 0.8. Thresholds are estimated using data from a 5-consecutive-day moving window and the empirical quantile as defined in the text. The first 30 yr are used as the base period. A jump (increase) in the exceedance rate is apparent at the boundary between the in-base and out-of-base periods, as indicated by 30-yr averages (thin dashed lines). Because of this jump, a highly significant trend (thick dashed line) can be identified if a linear trend is fitted to the exceedance time series, even though there is no trend in the simulated data.

2. Methods

a. Threshold function estimation

There are three aspects to consider in constructing an estimate of the threshold function. The first consideration is the choice of base period. To ensure that index time series can be easily extended into the future, the base period is usually chosen to be consistent with a recent World Meteorological Organisation (WMO) operational climatology base period (e.g., 1961–90 or 1971–2000). Most studies have used the 1961–90 base period because most indices of climate extremes were developed in the late 1990s (Karl et al. 1999) and because there is greater availability of data during this period than during other operational climatology base periods.

The second consideration is the type of subsampling that is used to select the data within the base period that will be used for threshold estimation. In this study, we use both the 5CD and 5SD windows. For example, to estimate the threshold for 13 January, the 5CD window selects data for all days in the base period dated 11–15 January. In contrast, all base period observations

dated 1, 7, 13, 19, and 25 January would be selected when the 5SD window is used. The latter approach uses only a small portion of available daily data between 1 and 25 January, and thus even though these observations are likely serially correlated, useful information has probably been discarded. For this reason, we also use all daily data available in the 1–25 January time window (25CD window) to estimate a threshold for 13 January.

The third consideration is the choice of method for estimating a threshold from a given dataset. One approach, as used by Frich et al. (2002) and others, is to use empirical quantiles that are obtained as follows. Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ be the n-sorted daily observations (i.e., order statistics) for a given day of the year that have been extracted from the base period with one of the data windows. In our case, $n = 5 \times 30 = 150$ for the 5CD and 5SD sampling methods, and $n = 25 \times 30 = 750$ for the 25CD sampling method. The empirical quantile corresponding to the pth percentile is computed by a linear interpolation of two values in the sorted data closest to the percentile. It is defined as

$$Q_p = (1 - f)y_{(j)} + fy_{(j+1)}, \tag{1}$$

with $j = \lfloor p^*(n + 1) \rfloor$ being the largest integer not greater than $p^*(n + 1)$, $f = p^*(n + 1) - j$, and $y_{(j)}$ is the jth largest value in the sample, for $1 \leq j < n$. The empirical quantile is set to the smallest or largest value in the sample when $j < 1$ or $j > n$, respectively. That is, quantile estimates corresponding to $p < 1/(n + 1)$ are set to the smallest value in the sample, and those corresponding to $p > n/(n + 1)$ are set to the largest value in the sample. Note that there are many different ways to estimate the empirical quantile corresponding to different ways of computing j (Hyndman and Fan 1996; Folland and Anderson 2002).

A second approach (e.g., Folland et al. 1999) is to fit a distribution to each sample and then to invert the fitted distribution to estimate the quantiles. As noted above, Folland et al. (1999) used a three-parameter Gamma distribution that can take a range of shapes. We will use the Gaussian distribution in this study because the data that we use in our Monte Carlo study have this distribution. Thus the choice of distribution does not add uncertainty in this study because it is known a priori. This is not the case in the real world. In general, uncertainty in the estimated distribution parameters and the choice of distribution will contribute to uncertainty in the estimated thresholds.

b. Exceedance indices

Once the threshold function is defined, the exceedance time series is estimated as described in Jones et al. (1999) and Frich et al. (2002). That is, the index for a given year, regardless of whether the year is inside or outside the base period, is the number of days in the year for which daily values have exceeded the estimated thresholds. As illustrated in Fig. 1, this seemingly correct approach may actually result in a discontinuity in the estimated exceedance time series at the boundaries between the in- and out-of-base periods. Consequently, trend analysis of these estimated time series may result in misleading conclusions.

The problem arises because the same base period observations are used to estimate the threshold function and in-base values of the index time series. Thus, as has also been noted in many other statistical applications in climatology (e.g., the "artificial predictability" issue discussed in Davis 1976), there is at least the potential for the in-base estimates of the exceedance series to be biased. Our threshold estimator (no matter how it is obtained) will be affected by sampling variability in the in-base sample. Thus the quantile estimate will never be identically equal to the true theoretical quantile, regardless of how the quantile is estimated. As a consequence, the mean out-of-base value of the exceedance time series will not be equal to the exceedance rate for the theoretical quantile. This means that while the in-base exceedance rate will be very close to 10% (if not exactly 10%—see below) by construction, the out-of-sample (out-of-base period) exceedance rate is unlikely to be exactly 10%.

c. Experimental design

Given that homogeneous time series are essential for monitoring and detecting climate change and that the thresholds are computed only from a portion of the data (usually a 30-yr base period), we designed a Monte Carlo simulation experiment to reveal whether inhomogeneities occur in the exceedance time series at the boundaries between the in-base and out-of-base periods. Daily values are usually serially correlated, which makes the effective sample size smaller than the actual sample size and hence influences the estimation of the thresholds and thus also the characteristics of the exceedance time series. Thus we use an auto regressive [AR(1)] process as described below to generate daily data values to also assess this effect.

Let X_t be a zero mean, unit variance AR(1) process

$$X_t = \alpha X_{t-1} + Z_t, \tag{2}$$

with lag 1-day autocorrelation α and white noise innovations Z_t with variance

$$\mathrm{Var}(Z_t) = 1 - \alpha^2. \tag{3}$$

We use $\alpha = 0.0$, 0.2, 0.4, 0.6, and 0.8 in order to study the impacts of different effective sample sizes. Note that values of α estimated from Canadian daily temperature data are typically between 0.6 and 0.8. For each α value, 60 yr of daily data are simulated using (2). The first and second 30-yr periods are assumed to be the in-base and out-of-base periods, respectively. Time series of annual exceedance rates are constructed as the number of days in the year for which daily values exceed the threshold estimated with (1). This procedure is repeated 1000 times. We then compare the statistical characteristics of the simulated exceedance time series in the two 30-yr periods.

To provide some insight regarding the sources of the discontinuity observed in Fig. 1, we also conducted a second set of Monte Carlo simulations as described below to examine the statistical properties and sampling errors of threshold and exceedance rate:

(a) We simulated 30 yr of autocorrelated daily data using (2).
(b) Daily data from each simulated year for days 1–5; for days 1, 7, 13, 19, and 25; and for days 1–25 were retained to estimate the 90th, 95th, and 99th percentiles ($\hat{Q}$) using the empirical quantile (1). Quantiles estimated in this way have the same properties as those obtained using the 5CD, 5SD, and 25CD windows. The probability $\hat{p}(X < \hat{Q})$ is obtained by inverting a standard Gaussian distribution. Note that $1 - \hat{p}$ is equivalent to the exceedance rate for the out-of-base period when that period is long.
(c) Steps a and b were repeated 5000 times.

The mean $\overline{\hat{Q}} = \Sigma\hat{Q}/5000$, standard deviation $\sigma_{\hat{Q}} = [\Sigma(\hat{Q} - \overline{\hat{Q}})^2/4999]^{1/2}$, and bias $\delta\hat{Q} = \overline{\hat{Q}} - Q$ of the quantile estimates were subsequently computed. The probability $p_{\overline{\hat{Q}}}$ corresponding to the average threshold $p_{\overline{\hat{Q}}} = p(X > \overline{\hat{Q}})$ was also computed by inverting the standard Gaussian distribution. The difference $\delta_{p_{\overline{\hat{Q}}}} = p_{\overline{\hat{Q}}} - p$ represents bias in the out-of-base threshold exceedance rate that is attributable to bias in the quantile estimate. The actual bias of exceedance rate $\delta_{\hat{p}}$ is $p - \Sigma(1 - \hat{p})/5000$. Results obtained from these two experiments are described in the following section.

3. Results

Figure 2 displays the relative bias in the exceedance rate estimated by using the 5CD window and empirical quantile. The bias is calculated as the difference between the average exceedance rate in 1000 simulations and the nominal rate expressed as the percentage of the nominal rate (the nominal rate is 10% when an estimate of the 90th percentile is used as the threshold).

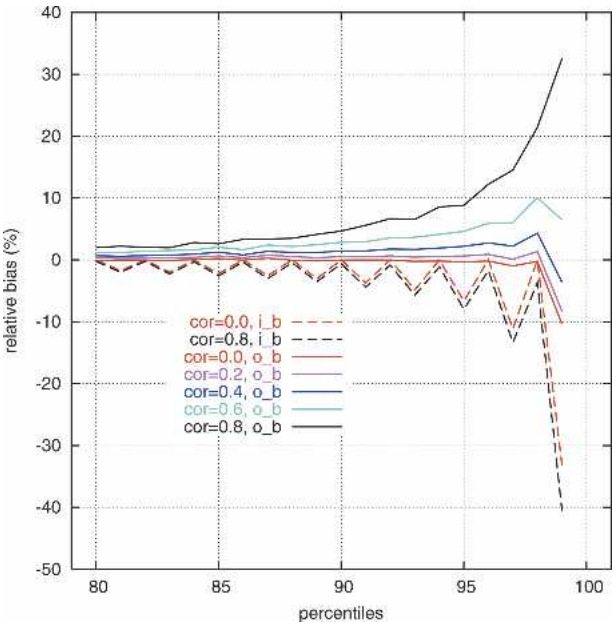

FIG. 2. Relative bias in the exceedance rate when thresholds are estimated by means of the empirical quantile with data from a 5-consecutive-day moving window, as a function of percentiles for in-base (i_b) and out-of-base (o_b) periods. Labels cor = 0.0, 0.2, . . . , 0.8 indicate that lag 1-day autocorrelation coefficients $\alpha = 0.0$, 0.2, . . . , 0.8 have been used, respectively.

The biases are shown for lag 1-day autocorrelation $\alpha = 0.0$, 0.2, 0.4, 0.6, and 0.8. Biases for the in- and out-of-base periods are very different.

In the in-base period, the exceedance rate bias is very small for some quantiles but is rather large with negative sign for other quantiles. The bias is not very sensitive to the value of α because the exceedance rate for the base period is adapted to the data. The estimated threshold always lies between the jth and $(j + 1)$ order statistic, where j is the integer portion of $p(n + 1)$, provided that the sample size is large enough. Thus the relative in-base bias will never be larger than $[(1/n) \times (100/1 - p)]$%.This holds regardless of whether we use the empirical quantile estimates described above, or another "plotting position" (i.e., another linear combination of the jth and $(j + 1)$ order statistics). The in-base bias varies systematically between zero and this bound as the percentile is varied.

To understand the cause of this variation, consider the estimated 90th and 91st percentiles. The number of exceedances for a sample of size 150 is 15 for the estimated 90th percentile, being equal to the nominal rate of 10% exactly. However, the number of exceedances for the estimated 91st percentile would be 13, giving an exceedance rate of 13/150 = 8.7%, which is smaller than the nominal rate of 9%. This bias is relatively

greater for higher percentiles. For example, there would be only one exceedance over the 99th percentile, giving an exceedance rate of 1/150 = 0.67%, which is much smaller than the nominal rate of 1%. Note that time series of exceedance rate for very high percentiles (e.g., 99th) also have other statistical properties that make the series undesirable for trend analysis and climate change detection. For example, the zero lower bound will be clearly apparent in these series, making it difficult to analyze trends with methods that assume a symmetric error distribution.

In this experiment, and other published studies, the estimated in-base exceedance rates are obtained by comparing a portion of the in-base sample data with the estimated thresholds. Bias in the exceedance rate for the in-base period will differ slightly from the above values because data from a moving window is used for threshold estimation, but additional experiments that we have conducted (not described above) indicate that the bias follows the pattern shown in Fig. 2 very closely. We use the term "rectification error" to denote this error. Because only the count number is involved, this bias is not sensitive to the use of different plotting positions, so long as the estimation of the threshold is based on interpolation between order statistics. However, the use of different plotting positions does affect the mean of the exceedance time series in the out-of-base period.

One possible approach for avoiding large in-base biases would be to carefully choose the combination of the window size and the quantile. However, this would be difficult to control in real applications where there are missing data within the base period and also perhaps an interest in multiple threshold levels. We note that the rectification error is closely related to sample size and can be reduced by using a larger sample, that is, by using a larger window such as the 25CD window. However, this may have the effect of reducing the amplitude and smoothing the annual cycle of thresholds, particularly in regions where the shape of the annual cycle is complex. The resulting thresholds may therefore have different expected exceedance rates for different calendar days as the annual cycle proceeds, making their interpretation more difficult and perhaps compromising the interpretation of the resulting index as an indicator of the frequency of moderate extremes. Another possible approach for reducing the rectification error without increasing the window size is to use a "fractional" exceedance rate where the integer number of observations above the threshold is "refined" by some fraction that depends linearly on the threshold and the two closest values above and below the threshold.

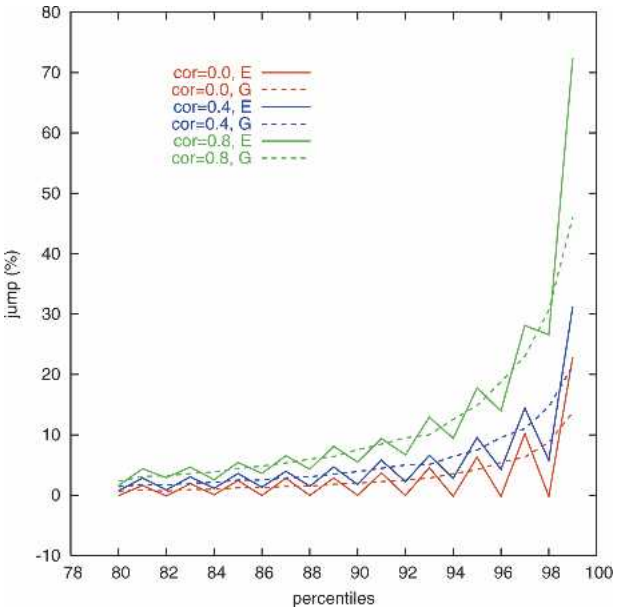

FIG. 3. Differences in exceedance rates between out-of-base and in-base periods expressed as a percentage of the nominal rate as a function of percentiles for different magnitudes of the lag 1-day autocorrelation (cor); E and G identify results that are obtained when thresholds are estimated with empirical quantiles or by fitting a Gaussian distribution to the data, respectively.

We repeated the above analyses, this time fitting a Gaussian distribution to the data from the in-base sample to estimate the quantile. Results indicate that quantiles tend to be underestimated, especially when the sample size is small and when autocorrelation is large (not shown), but the standard deviation is also smaller. As a result, exceedance rates for the out-of-base period are also overestimated. Figure 3 displays the differences in the exceedance rate between the out-of-base and in-base periods as a function of percentiles for different magnitudes of the lag 1-day autocorrelation. It is clear that the jump, a discontinuity in the mean value of the series at the boundary of in-base and out-base periods that is caused by a change of bias at the boundary, cannot be eliminated by estimating thresholds from a probability distribution that has been fitted to the in-base data.

The use of a 5SD window for quantile estimation results in the same amount of bias for the base period as the 5CD window because the bias is primarily the result of rectification error, which is an artifact of the size of the sample selected by the windows. When a 25CD window is used, the in-base error is greatly reduced due to the much larger sample size and hence reduced rectification effects. Note, however, that the possible effect of the attenuation of the annual cycle of thresholds, which may be large as discussed above, has not

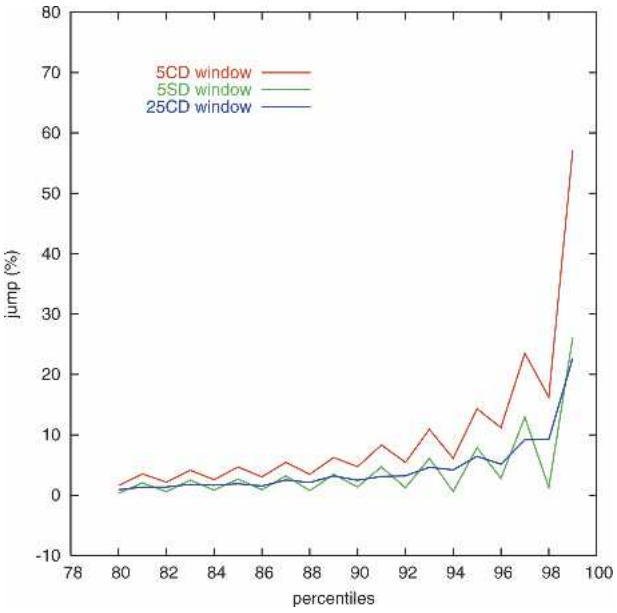

FIG. 4. Differences in exceedance rates between out-of-base and in-base periods expressed as the percentage of the nominal rate as a function of percentiles when 5CD, 5SD, and 25CD windows are used to estimate empirical quantiles. The lag 1-day autocorrelation, α, is set to 0.8.

been accounted for in these experiments. The exceedance rate biases for the out-of-base period are very similar for the 5SD and 25CD windows. As a result, a jump in the exceedance series is still apparent at the boundaries between the two periods (Fig. 4). The magnitude of the discontinuity is much smaller than for the 5CD window.

It appears that there will be a discontinuity in the exceedance series at the boundary of the in-base and out-of-base periods, no matter how the thresholds are obtained. This discontinuity may not be detectable in individual exceedance time series against the background of natural interannual variability, but it will become detectable when multiple exceedance time series are aggregated, as we will demonstrate in section 5.

We now briefly discuss the sources of the biases documented above. The biases displayed in Fig. 2 for the out-of-base period are generally positive, with larger relative biases corresponding to larger values of α and higher percentiles. These biases are affected by several factors. One is the bias of the quantile estimator, which affects the bias of the exceedance rate in a nonintuitive manner. For example, an unbiased quantile estimator will result in a biased exceedance rate estimator (the appendix). The second factor is sampling variability. Autocorrelation, when present, affects both of these factors by reducing the equivalent information in a sample of a given size.

Results from the *second* Monte Carlo simulation are summarized in Table 1. They show that the empirical quantile (1) is generally positively biased and that the bias tends to decrease with an increase in α. Table 1 also shows that the standard deviation of the quantile estimate increases when the percentile increases and when α increases. The latter result reflects the fact that when α increases, the same size of sample contains less information about the quantile, that is, the equivalent sample size is reduced. Finally, we see from Table 1 that $\overline{p}$ is always larger than $\overline{p}_{\hat{Q}}$. This suggests that the small negative bias in the out-of-base exceedance rate that is caused by overestimation of the quantile is more than overcome by a positive bias that results from sampling uncertainty in the quantile estimate.

TABLE 1. Biases ($\delta\hat{Q}$) and standard deviation ($\sigma_{\hat{Q}}$) of quantile estimates, percentage changes in probability corresponding to average quantile ($\delta p_{\hat{Q}}$), and percentage change in estimated exceedance rate ($\delta\hat{p}$) in 5000 simulations for 5CD, 5SD, and 25CD windows. The values for α = 0.0, 0.4, 0.8 are the lag-1 autocorrelation used in simulating the data. See text for details.

	Percentile (%)	90			95			99		
	α	0.0	0.4	0.8	0.0	0.4	0.8	0.0	0.4	0.8
5CD	$\delta\hat{Q}$	0.015	0.009	0.005	0.026	0.021	0.008	0.150	0.136	0.050
	$\sigma_{\hat{Q}}$	0.139	0.166	0.221	0.170	0.195	0.262	0.325	0.342	0.408
	$\delta p_{\hat{Q}}$	−2.6	−1.6	−0.8	−5.1	−4.2	−1.6	−33.6	−31.0	−12.5
	$\delta\hat{p}$	−0.5	1.4	4.6	−0.4	2.0	9.7	−9.8	−3.7	33.4
5SD	$\delta\hat{Q}$	0.013	0.012	0.009	0.027	0.026	0.021	0.155	0.161	0.141
	$\sigma_{\hat{Q}}$	0.139	0.141	0.153	0.172	0.174	0.184	0.319	0.319	0.332
	$\delta p_{\hat{Q}}$	−2.3	−2.1	−1.6	−5.5	−5.3	−4.4	−34.5	−35.6	−31.9
	$\delta\hat{p}$	−0.1	0.0	1.0	−0.7	−0.3	1.7	−11.4	−12.9	−6.3
25CD	$\delta\hat{Q}$	0.002	0.000	−0.003	0.005	0.002	−0.003	0.023	0.025	0.003
	$\sigma_{\hat{Q}}$	0.063	0.080	0.129	0.079	0.093	0.148	0.140	0.154	0.230
	$\delta p_{\hat{Q}}$	−0.4	0.0	0.5	−1.1	−0.5	0.7	−6.1	−6.4	−0.8
	$\delta\hat{p}$	0.0	0.6	2.4	0.0	1.0	4.4	−0.4	0.6	15

To understand this last point, let $\varepsilon_q > 0$, and let Q be a quantile in the right tail of the probability distribution. Then, because the probability density decreases monotonically in the right tail, we find that

$$[P(X > Q - \varepsilon_q) - P(X > Q)] > [P(X > Q) - P(X > Q + \varepsilon_q)].$$

This means that if the sampling uncertainty in the quantile estimate follows a symmetric distribution, then the out-of-base exceedance rate is positively biased even if the quantile estimate itself is unbiased. Note that quantile estimates for moderately large percentiles do, roughly, follow a symmetric distribution when the raw data are Gaussian and sample sizes are the same as those used in this study.

Autocorrelation, when present, appears to have two effects, both of which have a tendency to increase the overall bias in the out-of-base exceedance rate. First, autocorrelation appears to reduce the bias in the quantile estimate $\hat{Q}$, which has the effect of reducing or eliminating the corresponding negative bias in the exceedance rate. Second, autocorrelation increases the variability of $\hat{Q}$, which further increases the bias from that source as discussed above.

In summary, the overall bias in the out-of-base exceedance rate results from the bias of the quantile estimate and its variability. The former effect appears to be reduced when the daily data are serially correlated, but this apparent reduction is overwhelmed by the effects of increased sampling variability in the quantile estimate. As a result, the overall bias in the exceedance rate increases when the observations are positively serially correlated. The positive bias for the out-of-base period and the tendency for negative bias for the in-base period result in a jump in the exceedance rate at the boundaries between the in-base and out-of-base periods. Relative to the nominal rate, the jump becomes larger when higher percentiles are used to define extremes.

4. Removing the "jump"

We have shown that the seemingly simple exceedance time series is actually very difficult to estimate, and that there is a discontinuity in the expected threshold exceedance rate at the in-base and out-of-base boundaries. Several approaches may be considered to solve this problem. One approach would be to choose the base period entirely outside the period for which trends are calculated. In practice, this is difficult to implement since not all stations would have long enough data to cover such a base period. Alternatively, one could estimate the thresholds from whatever data are available for the station. However, this implies the use of different base periods for different stations, and it would be difficult to compare indices among the stations. Another method might be to use a more refined threshold estimate that has more consistent in-base and out-of-base exceedance rate properties. Our judgment, however, is that this would be a difficult task. Our experiments with different data windows, the empirical quantile estimate using various plotting positions, and a distribution function quantile estimator all suggest that this approach will not yield robustly and consistently improved results.

The fundamental difficulty is that in-base estimates of the threshold exceedance rate are not fully reliable estimates of the out-of-sample (out of base) exceedance rate. This is a familiar problem in climatology (e.g., Davis 1976; von Storch and Zwiers 1999) that can often be resolved by using a bootstrapping or cross-validation procedure. Thus instead of trying to adjust the threshold, we will attempt to estimate the in-base period exceedance rates in a manner that mimics exceedance rate estimation in the out-of-base period. In the latter case, the sample that is used to estimate the exceedance rate is independent of the sample used to estimate the threshold. By doing so, we accept that the mean exceedance rate will be different from the nominal rate. This is of secondary concern if a homogeneous index time series can be obtained for climate change monitoring and detection purposes.

Our procedure consists of the following steps:

(a) The 30-yr base period is divided into one "out of base" year, the year for which exceedance is to be estimated, and a "base period" consisting the remaining of 29 yr from which the thresholds would be estimated.
(b) A 30-yr block of data is constructed by using the 29 yr base period dataset and adding an additional year of data from the base period (i.e., one of the years in the base period is repeated). This constructed 30-yr block is used to estimate thresholds. Note that other resampling approaches for constructing a 30-yr block could also be used, perhaps equally as effectively. For example, one could select 30 yr from the 29 yr base period by means of simple random sampling with replacement (simple bootstrap). If there is concern about interannual serial correlation, then the block bootstrap (Wilks 1997) is also an alternative.
(c) The out-of-base year is then compared with these thresholds, and the exceedance rate for the out-of-base year is obtained.

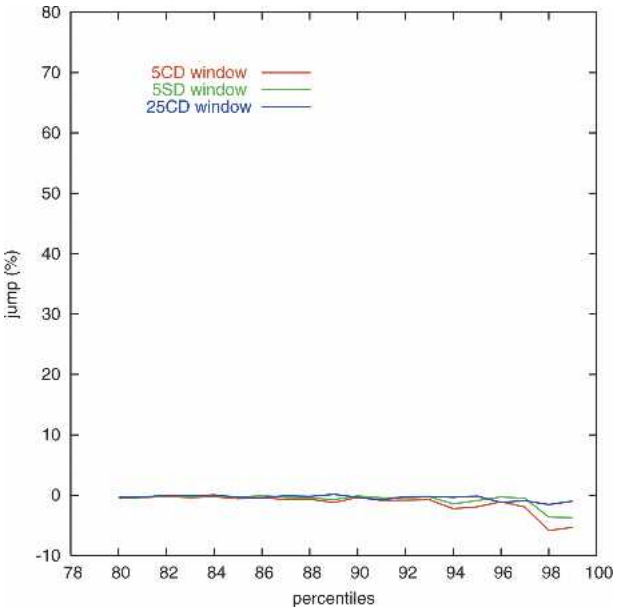

FIG. 5. Same as in Fig. 4, except the exceedance rates for the in-base period are estimated using a bootstrap resampling procedure as described in the text.

(d) Steps b and c are repeated an additional 28 times, by repeating each of the remaining 28 in-base years in turn to construct the 30-yr block.
(e) The final index for the out-of-base year is obtained by averaging the 29 estimates obtained from steps b, c, and d.

In this way, the year for which the exceedance rate is to be estimated is not used for estimating the thresholds. By repeating one of the 29 "in base" years, we insure that the rectification error in the threshold used to estimate the index in the withheld year is comparable to the rectification error experienced when calculating the out-of-base index values. This effectively makes the estimation of the exceedance rate for both the in-base and out-of-base periods comparable, greatly reducing the discontinuity.

Figure 5 shows the differences in the average exceedance rates obtained in 1000 Monte Carlo simulations between out-of-base and in-base periods when the lag 1-day autocorrelation α in (2) is set to 0.8 and when data from different windows are used. The thresholds used in this example were empirical quantiles. The jump in the exceedance series is almost entirely eliminated, with a small jump remaining evident only for the very largest quantiles when the 5-day data windows are used. The jump is essentially undetectable when the lag 1-day correlation is less than 0.8. Similar results are obtained when the quantiles are estimated by fitting a probability distribution to the data (not shown).

5. Conclusions and discussion

We have compared the performances of different methods of producing temporally homogeneous time series of exceedance rates. We used both an empirical probability distribution and also a fitted distribution to estimate thresholds from data selected with a 5CD (5 consecutive day) moving window, a 5SD (5 days spaced by 5 days) moving window, and a 25CD (25 consecutive day) moving window. Our performance evaluation was conducted with the aid of Monte Carlo simulation experiments. We found that the exceedance rate time series has discontinuities at the boundaries between the in- and out-of-base periods if the rate is estimated using existing methods. Our bootstrap resampling procedure overcomes this problem and produces much more homogeneous estimates of the exceedance rate across the two periods.

The 5CD moving window approach produces the largest bias in the estimated exceedance rate. The 5SD moving window that is used in Jones et al. (1999) offers some improvement for the out-of-base period. But the bias for the in-base period is the same as for the 5CD window since the same amount of data is used for the estimation of quantiles. The 25CD moving window approach yields the smallest bias for the base period. Note however that attenuation of the annual cycle of thresholds may become a problem when using large moving windows and that this effect may introduce large biases into the exceedance rate time series that might compromise its interpretation as an indicator of the frequency of moderately large extremes.

The difference in exceedance rates that results from using different methods for quantile estimation (empirical quantile or estimation from a fitted probability distribution) is small. Also, because we are primarily interested in monitoring change in exceedance rates over time, homogeneity of the exceedance rate time series is of substantially greater concern than modest biases in the exceedance rate, provided that those biases do not compromise the interpretation of the index time series as an indicator of the frequency of moderately large extremes. We therefore recommend the use of an empirical quantile for its simplicity along with the 5CD moving window to estimate thresholds. The exceedance series for the base period should be estimated using the bootstrap resampling procedure described above to avoid discontinuities at the in-base and out-of-base boundaries.

The inhomogeneity in the exceedance series estimated by existing methods could have profound impacts if the series are used for climate change monitoring and for trend computation in particular. For ex-

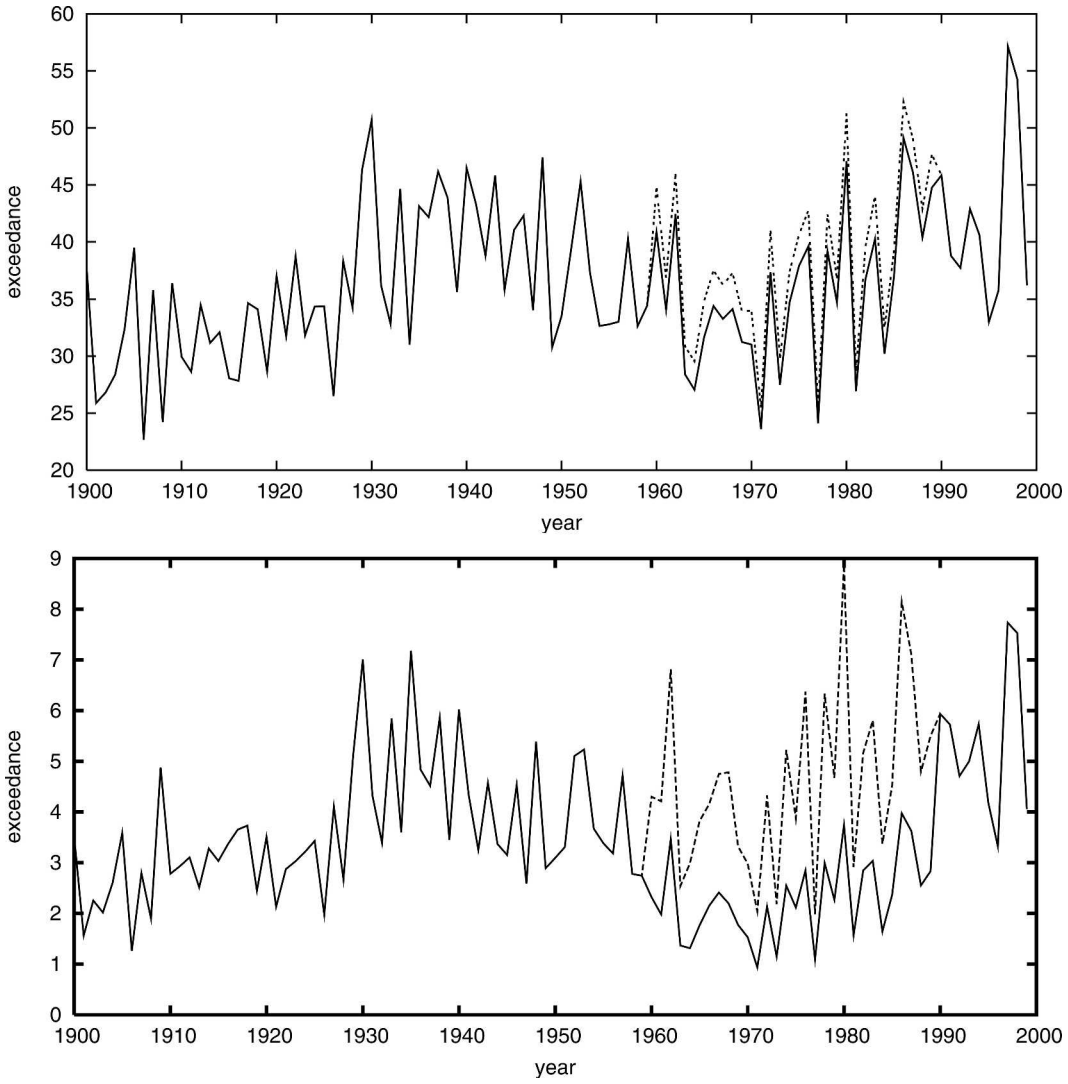

FIG. 6. Number of days on which daily mean temperature exceeded its (top) 90th and (bottom) 99th percentiles over Canada. Rates for the in-base period computed with the bootstrap resampling procedure described in the text are shown with the dashed curves. Note that the jump in the 90th percentile series is mainly due to the bias in the out-of-base estimates, while the biases in both the in-base and out-of-base periods contribute to the jump in the 99th percentile series.

ample, we show in Fig. 6 the average exceedance rates of daily temperature at 210 stations in Canada. The thresholds in this case are the 90th and 99th percentiles (empirical quantiles) of daily temperature that are obtained using the 5CD moving window. The daily temperature data that were used have been homogenized to remove step changes caused by changes in station location and/or measurement programs (Vincent et al. 2002). These data have been used previously for analyzing trends in daily and extreme temperatures (Bonsal et al. 2001). The exceedance rates are averaged across the 210 stations to obtain an extreme index for Canada. Clearly, the exceedance rates for 1961–90 are underestimated when the resampling procedure is not used. In this case, two artificial jumps are apparent in the time series of spatially averaged exceedance rates, one at the beginning of the base period and the other at the end of the period. Note that the jumps are greater when a higher percentile is used to define the threshold. The trend in the extreme indices would also be distorted if the existing method is used to estimate the in-base exceedance rate. The distortion would be greater if the base period is at the beginning or at the end of the time series, such as would be the case if estimating the trend in the index for the last three–four decades of the twentieth century.

More importantly, misleading conclusions could be reached if inhomogeneous indices series are used in climate change detection studies. The essence of climate change detection is to identify a weak climate change signal as simulated by coupled global climate models in observed data. If extreme indices for both observed and model-simulated data are computed similarly using existing methods, there would be artificial jumps in the series obtained from both observed and model-simulated data. This could easily become a part of the signal and lead to erroneous or overstated detection claims in a climate change detection study. Although such erroneous results might be preventable by also including base periods in data for climate variability, the presence of artificial jumps will still make results difficult to interpret. It is therefore important to use our resampling procedure to eliminate a small, but detectable and avoidable, inhomogeneity in the threshold exceedance indices.

Acknowledgments. GCH was supported by NSF Grants ATM-0002206 and ATM-0296007, by NOAA Grant NA16GP2683 and NOAA's Office of Global Programs, by DOE in conjunction with the Climate Change Data and Detection element, and by Duke University. We are very grateful to Richard Chandler for deriving the proof that appears in the appendix. We thank Nathan Gillett, Viatcheslav Kharin, Chris Ferro, Editor David Stephenson, and two anonymous reviewers for their comments that improved an earlier draft of this paper.

APPENDIX

An Unbiased Quantile Estimator Results in a Biased Estimate of the Exceedance Rate

Suppose $y_1, y_2, \ldots, y_t$ are identically distributed continuous random variables (not necessary independent) with probability density function f and corresponding distribution function F. Let q_α be the $(1 - \alpha)$th quantile of f so that

$$F(q_\alpha) = 1 - \alpha, \quad \text{and} \quad q_\alpha = F^{-1}(1 - \alpha). \quad \text{(A1)}$$

Let $\hat{q}_\alpha = \hat{q}_\alpha(y_1, y_2, \ldots y_t)$ be an estimator of q_α, and define $\hat{\alpha} = P(y > \hat{q}_\alpha) = 1 - F(\hat{q}_\alpha)$, where y is an additional random variable with the same distribution. Then, providing the effective sample size of the series $y_1, y_2, \ldots, y_t$ is large so that $|\hat{q}_\alpha - q_\alpha|$ is small, we have

$$\hat{\alpha} = 1 - F(\hat{q}_\alpha) = 1 - \left[F(q_\alpha) + (\hat{q}_\alpha - q_\alpha)F'(q_\alpha) + \frac{1}{2}(\hat{q}_\alpha - q_\alpha)^2 F''(q_\alpha) + \varepsilon \right]. \quad \text{(A2)}$$

For some $\hat{q}^*$ between $\hat{q}$ and q, and if f'' is continues in the neighborhood of α,

$$\varepsilon = (\hat{q}^* - q)^3/6. \quad \text{(A3)}$$

If f'' is small in the neighborhood of α, which is the case in the upper tail of most of probability distributions, we have

$$\hat{\alpha} \approx \alpha - (\hat{q}_\alpha - q_\alpha)f(q_\alpha) - \frac{1}{2}(\hat{q}_\alpha - q_\alpha)^2 f'(q_\alpha). \quad \text{(A4)}$$

It follows that if $\hat{q}_\alpha$ is unbiased, then the expected value of $\hat{\alpha}$ may be approximated by

$$E(\hat{\alpha}) \approx \alpha - \frac{1}{2}f'(q_\alpha)var(\hat{q}_\alpha). \quad \text{(A5)}$$

Thus, for smooth distributions other than the uniform, we expect $\hat{\alpha}$ to be biased upward in the upper tail (i.e., where $f' < 0$). Hence the exceedance rate will be biased when the quantile estimator is unbiased. According to (A5), the bias is approximately proportional to $var(\hat{q}_\alpha)$ and is thus likely to be larger for autocorrelated sequences.

REFERENCES

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.

Davis, R. E., 1976: Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.,* **6,** 249–266.

Easterling, D. R., L. V. Alexander, A. Mokssit, and V. Detemmerman, 2003: CCl/CLIVAR workshop to develop priority climate indices. *Bull. Amer. Meteor. Soc.,* **84,** 403–407.

Folland, C., and C. Anderson, 2002: Estimating changing extremes using empirical ranking methods. *J. Climate,* **15,** 2954–2960.

——, and Coauthors, 1999: Workshop on indices and indicators for climate extremes, Asheville, NC, USA, 3–6 June 1997, Breakout Group C: Temperature indices for climate extremes. *Climatic Change,* **42,** 31–43.

Frich, P., L. V. Alexander, P. Della-Marta, B. Gleason, M. Haylock, A. M. G. Klein Tank, and T. Peterson, 2002: Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Res.,* **19,** 193–212.

Hyndman, R. J., and Y. Fan, 1996: Sample quantiles in statistical packages. *Amer. Stat.,* **50,** 361–365.

Jones, P. D., E. B. Horton, C. K. Folland, M. Hulme, D. E. Parker, and T. A. Basnett, 1999: The use of indices to identify changes in climatic extremes. *Climatic Change,* **42,** 131–149.

Karl, T. R., N. Nicholls, and A. Ghazi, 1999: CLIVAR/GCOS/WMO workshop on indices and indicators for climate extremes: Workshop summary. *Climatic Change,* **42,** 3–7.

Kiktev, D., D. M. H. Sexton, L. Alexander, and C. K. Folland,

2003: Comparison of modeled and observed trends in indices of daily climate extremes. *J. Climate,* **16,** 3560–3571.

Klein Tank, A. M. G., and G. P. Können, 2003: Trends in indices of daily temperature and precipitation extremes in Europe, 1946–99. *J. Climate,* **16,** 3665–3680.

Peterson, T. C., C. Folland, G. Gruza, W. Hogg, A. Mokssit, and N. Plummer, 2001: Report on the activities of the Working Group on Climate Change Detection and Related Rapporteurs 1998–2001. World Meteorological Organization Rep. WCDMP-47, WMO-TD 1071, Geneva, Switzerland, 143 pp.

——, and Coauthors, 2002: Recent changes in climate extremes in the Caribbean region. *J. Geophys. Res.,* **107,** 4601, doi:10.1029/2002JD002251.

Vincent, L. A., X. Zhang, B. R. Bonsal, and W. D. Hogg, 2002: Homogenization of daily temperatures over Canada. *J. Climate,* **15,** 1322–1344.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

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