A semi-parametric regression approach to climatological quantile estimation for generating percentile-based temperature extremes indices

Chi Yang1* and Jing Xu2
1College of Global Change and Earth System Science, Beijing Normal University, Beijing, China
2State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

Abstract

A semi-parametric regression approach to quantile estimation for daily temperature data is proposed, in which both the biases and inhomogeneity are negligible, and is applied to the calculation of the six percentile-based Expert Team on Climate Change Detection and Indices (ETCCDI) temperature extremes indices. Comparisons of the results with those from the CLIMDEX datasets show that the three warmth indices in the latter are probably biased such that their linear trends under the RCP4.5 scenario seem to be overestimated. In order to avoid drawing misleading conclusions, it is necessary to re-examine currently adopted algorithms and available datasets, and to develop new methods for generating the percentile-based ETCCDI indices.

Keywords: climate extremes; daily temperature data; ETCCDI indices; CLIMDEX project; semi-parametric quantile regression; CMIP5

1. Introduction

Changes in climate extremes have been a hot spot in climate change research, because of their potential associations with increased natural disasters. To facilitate the monitoring and analysis of climate extremes, the joint CCI/WCRP/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI) (http://wcrp-climate.org/etccdi) had defined a set of 27 core indices describing the more extreme aspects of climate based on daily temperature and precipitation data (Karl et al., 1999). Another joint project CLIMDEX (http://www.climdex.org) has finished producing public domain datasets of such indices for both global land-based observations and output from global climate models (GCM) participating in the Coupled Model Intercomparison Project Phase 3 (CMIP3) and Phase 5 (CMIP5). Among the 27 ETCCDI indices, 6 of them are based on percentiles of daily temperature data during the base period 1961–1990 (Table 1). They represent the tail behavior of climatological distributions. The CLIMDEX project used the empirical method as its fundamental algorithm to estimate sample quantiles (Zhang et al., 2005, referred to as ZHZK hereafter). Let \( \{y_1, \ldots, y_n\} \) be the order statistics of a sample \( \{y_1, \ldots, y_n\} \), the \( p \)th percentile estimate can be formally expressed as:

\[
\hat{Q}_p = \left[1 - \gamma(j)\right] y(j) + \gamma(j) y(j+1)
\]

(1)

There are many different ways to determine \( j \) and \( \gamma(j) \) associated with \( p \) (Hyndman and Fan, 1996; Folland and Anderson, 2002), resulting in slightly different estimates. To estimate the data percentiles for each calendar day, a 5-consecutive-day window centered on the day of interest is used throughout the base period so that \( 5 \times 30 = 150 \) sample data are available for that day. The percentile-based ETCCDI temperature extremes indices are then calculated accordingly for both the in-base and out-of-base periods.

ZHVK studied the effect of empirical quantile estimator on percentile-based indices of temperature extremes, and pointed out that the above algorithm resulted in a jump (increase) in annual exceedance rate (ER) at the boundary between the in-base and out-of-base periods. They designed a Monte Carlo experiment with 60-year daily series generated by an order-1 autoregressive process [AR(1)]

\[
X_t = \alpha X_{t-1} + Z_t, \quad Z_t \sim N\left(0, \sqrt{1 - \alpha^2}\right)
\]

(2)

with different values of autocorrelation coefficient \( \alpha \). The first and second 30-year periods are assumed to be the in-base and out-of-base periods, respectively. ZHZK generated ensembles of 1000 simulations using model (2) with \( \alpha = 0.0, 0.2, 0.4, 0.6 \) and 0.8, respectively, to study the impacts of different effective sample sizes on the usually positive biases for the out-of-base period. The black curve in Figure 1 is the ensemble-mean annual ERs for the 90th percentile (0.9 in quantile) with \( \alpha = 0.8 \), showing clearly the jump at the boundary of two periods such that a fake linear trend is presented (dashed black line). They concluded that the biases for the out-of-base period are mainly affected by the...
quantile estimator, sampling variability and autocorrelation in data. In fact, the first two sources of biases are all closely related to the limited sample size. For the 90th percentile, the expected exceedance is only 10% of, or 15, sample data. The higher the percentile, the less the exceedance is, resulting in more significant biases. ZHZK finally proposed an additional bootstrap re-sampling procedure to remove the inhomogeneity in the annual ER series. The improved algorithm had been adopted by the CLIMDEX project for generating the first four indices of temperature extremes in Table 1 [it seems that for unknown reasons, the last two indices were calculated still based on the empirical quantile estimator (1) rather than the improved algorithm]. The CLIMDEX datasets and software have been widely used in climate extremes analysis and GCM performance assessment (e.g. Bürger et al., 2012, 2013; Li et al., 2013; Sillmann et al., 2013a, 2013b; Dong et al., 2015; Zhou et al., 2016).

However, ZHZK’s method did not really correct the biases. They admitted that ‘the mean exceedance rate will be different from the nominal rate’, but they thought that ‘this is of secondary concern if a homogeneous index time series can be obtained for climate change monitoring and detection purposes’. Nevertheless, as shown later, even though the annual ER series seems to be homogeneous, its linear trend may still be distorted by intrinsic biases. Of course, if we put aside the use of a base period required by ETCCDI and simply use all the available data to estimate quantiles (e.g. Yan et al., 2002), homogeneity can be guaranteed. However, if the spans and periods of two data series are quite different, their quantiles may lose comparability if estimated individually without using a common base period. In addition, the autocorrelation issue usually tends to be overlooked. Although ZHZK pointed out that the existence of autocorrelation may increase the overall bias in the out-of-base ER, their method does not actually deal with it either. Therefore, it is still necessary to improve the base-period approach by reducing biases while taking autocorrelation into account.

In this article, we present a model-based approach to climatological quantile estimation, aiming at producing bias- and inhomogeneity-negligible ER series throughout in-base and out-of-base periods for daily temperature data. All the data from the base period are used as sample data to make the sample size large enough. Quantiles to be estimated are regressed to nonlinear functions representing the seasonality, and to previous days’ observations accounting for autocorrelation. This approach is systematic and easy to implement, providing practitioners with a framework for solving their own quantile estimation problems in climatic change research. In the following, Section 2 briefly describes the quantile regression model (QRM) and examines the model performance by Monte Carlo experiments. Section 3 presents an example of applying the semi-parametric QRM to CMIP5 GCM output. The new results are compared with their counterparts in the CLIMDEX dataset. Conclusions and remarks are given in Section 4.

2. Semi-parametric regression model for quantile estimation

2.1. Quantile regression

Quantile regression is a methodology for estimating quantile functions of response conditional on covariates (Koenker, 2005). Unlike the usual linear regression that leads to least squares problems, quantile regression leads to linear programming problems that can be solved by simplex method. The implementation of quantile regression has been included in

Table 1. Percentile-based ETCCDI temperature extremes indices

| ID     | Name                | Definition                                                                 | Unit          |
|--------|---------------------|-----------------------------------------------------------------------------|---------------|
| TN10p  | Cool nights        | Percentage of days when daily minimum temperature < 10th percentile        | %             |
| TX10p  | Cool days          | Percentage of days when daily maximum temperature < 10th percentile        | %             |
| TN90p  | Warm nights        | Percentage of days when daily minimum temperature > 90th percentile        | %             |
| TX90p  | Warm days          | Percentage of days when daily maximum temperature > 90th percentile        | %             |
| WSDI   | Warm spell duration | Annual count of days with at least 6 consecutive days when daily maximum temperature > 90th percentile | Days          |
| CSDI   | Cold spell duration | Annual count of days with at least 6 consecutive days when daily minimum temperature < 10th percentile | Days          |

Detailed definitions are given at http://www.climdex.org/indices.html.

Figure 1. Ensemble-mean annual ERs for the 90th percentile of daily data generated by the AR(1) model with \( \alpha = 0.8 \) (curves), estimated by empirical estimator and QRM, respectively, and their linear trends (dashed lines).
numerous statistical softwares, such as R: A Language and Environment for Statistical Computing (R Core Team, 2016). The QRM for estimating the 90th percentile of AR(1)-simulated data is simply

\[ q_{90}(t) = \beta_0 + \beta_1 x(t - 1) \]  

(3)

where \( q_{90}(t) \) is the 90th percentile for the \( t \)th day, \( x(t - 1) \) is the previous day’s observation, \( \beta_0 \) and \( \beta_1 \) are coefficients. In Figure 1, the red curve is the counterpart of the black curve estimated by model (3), showing actually no linear trend throughout the 60 years (dashed red line). The means of the series for both the in-base and out-of-base periods are all 10.0, almost exactly the ‘nominal’ ER. Compared with the empirical results, biases in the QRM estimates are negligible so that there is no discernible jump at the boundary between the two periods.

Following ZHZK, a Monte Carlo experiment is carried out to generate ensembles of 1000 simulations using model (2) with \( \alpha = 0.0, 0.2, 0.4, 0.6, 0.8 \), respectively. Sample percentiles are then estimated by using model (3) instead of the empirical estimator (1). Figure 2(a) shows the ensemble-mean relative bias in ER as a function of percentile for each combination of period and \( \alpha \) value individually. It can be seen that for in-base period relative biases are so close to 0 that their curves are almost overlapped (solid curves); only for out-of-base period they deviate a little bit from 0, and increase with percentile in general (dashed curves). Figure 2(b) shows the ensemble-mean difference in ER between out-of-base and in-base periods (‘jump’) relative to its nominal percentile as a function of percentile for each \( \alpha \) value individually. As the ERs for in-base period are almost bias-free (Figure 2(a)), these jump curves have very similar shapes with those of their counterparts (dashed curves) in Figure 2(a). Figures 2(a) and (b) are comparable with ZHZK’s Figure 2 and Figures 3–5, respectively. The percentage magnitudes in Figure 2 are all less than 1%, showing that the QRM results are not only much better than the empirical estimates, but also better than ZHZK’s improved estimates.

2.2. Semi-parametric model for quantile estimation

The daily temperature at a site during a base period can be regarded as a periodically stationary process. There

![Figure 2](image-url)

**Figure 2.** (a) Ensemble-mean relative bias in ER (difference of mean ER with the nominal ER relative to the nominal ER) as a function of percentile for each combination of period and \( \alpha \) value individually. (b) Ensemble-mean jump (difference of mean ER between out-of-base and in-base periods relative to the nominal ER) as a function of percentile for each \( \alpha \) value individually.

![Figure 3](image-url)

**Figure 3.** (a) Ensemble-mean annual ERs for the 90th percentile of the artificial daily temperature data (curves), estimated by ZHZK’s method and semi-parametric QRM, respectively, and their linear trends (dashed lines); (b) Ensemble-mean monthly ERs for the 90th percentile of the same data as in (a), averaged over each calendar month, for the in-base (solid curves) and out-of-base (dashed curves) periods, estimated by ZHZK’s method and semi-parametric QRM, respectively.
are two main sources accounting for its variability: radiative forcing and weather processes, resulting in seasonality and autocorrelation in data. Seasonality can be represented as a nonlinear function through non-parametric regression method (Green and Silverman, 1994). The lagged effect for daily temperature is usually significant for up to 3 days, which are accounted for by parametric autoregressive terms. Put the non-parametric and the parameter terms together, the semi-parametric QRM for daily temperature can be expressed as:

\[ q_r(t) = \beta_0 + s(d_r) + \beta_1 x(t - 1) + \beta_2 x(t - 2) + \beta_3 x(t - 3) \]  

where \( q_r(t) \) is the \( r \)th percentile for the \( r \)th day of the base period; \( s(d_r) \) is the non-parametric term for seasonality as a nonlinear function of calendar date \( d_r \in \{1, 2, \ldots, 365\} \) for the \( r \)th day; the other terms are similar to those in model (3).

The non-parametric term \( s(d_r) \) is implemented through smoothing splines. To attain the optimized smoothness, model (4) is fitted alternatively through the vector generalized additive model (VGAM) approach (Yee, 2015), by which the response is assumed to follow the asymmetric Laplace distribution (ALD) (Kotz et al., 2001). The ALD has the following density function:

\[ f(y; \xi, \sigma, \kappa) = \begin{cases} \frac{\sqrt{2}}{\sigma} \frac{\kappa}{1+\kappa} \exp \left( -\sqrt{2} \frac{\kappa}{\sigma} |y - \xi| \right), & \text{for } y \leq \xi, \\ \frac{\sqrt{2}}{\sigma} \frac{\kappa}{1+\kappa} \exp \left( -\sqrt{2} \frac{\kappa}{\sigma} |y - \xi| \right), & \text{for } y > \xi, \end{cases} \]  

where \(-\infty < \xi < \infty\) and \(\sigma, \kappa > 0\). An important property of the ALD is that \(P(Y \leq \xi) = \tau\) where \(\tau = \kappa^2/(1 + \kappa^2)\) so that \(\kappa = \sqrt{\tau/(1 - \tau)}\). Therefore, if we build a simplified VGAM such as

\[ \xi(t) = \beta_0 + s(d_r) + \beta_1 x(t - 1) + \beta_2 x(t - 2) + \beta_3 x(t - 3), \]  

where \(\sigma(t)\) is assumed to be constant, then \(\xi(t)\) is the \(r\)th percentile for the \(r\)th day. Model (6) can be estimated by maximizing its penalized likelihood

\[ l_p = l - \frac{\lambda}{2} \int_1^{365} \left[ s''(t) \right]^2 dt \]  

where \(l\) is the likelihood of model (6), \(\lambda\) is a positive parameter controlling the smoothness of splines. Thanks to the R package VGAM (Yee, 2016), model (6) can be easily implemented as an alternative approach to model (4). In VGAM, B-splines are used as the basis functions of a smooth term. To make sure that the seasonal variability for each calendar day is fully considered, the effective degree of freedom of the smooth term is set to be the highest value of 365 (suitable for the no-leap-year calendar).

### 2.3. Monte Carlo experiment with artificial temperature data

The AR(1)-simulated data were used by ZHZK to examine the empirical and their improved quantile estimation methods, and are used again in 2.1 to assess the QRM method for comparison. However, such data are not representative of daily temperature. Here we use an AR(3) process

\[ X_t = 0.8X_{t-1} - 0.2X_{t-2} + 0.1X_{t-3} + Z_t, \]  

\[ Z_t \sim N(0, 1) \]  

superimposed on a trigonometric function

\[ s(t) = 10 + 20 \sin \left[ 2\pi (t - 100) / 365 \right] \]  

to mimic daily temperature observations typical in mid-latitude land areas. A total of 5000 simulations of 60-year daily series are generated. The first 30 years are assigned to be the base period. Semi-parametric QRM and ZHZK’s method are applied to these simulations, respectively, to estimate annual and monthly ERs for the 90th percentile. ZHZK’s method is implemented using the R package climdex.psic (Bronaugh, 2015), which is also the software producing the CLIMDEX datasets.

Figure 3(a) shows the ensemble-mean annual ER series and their linear trends. It can be seen that both methods still have tiny jumps at the boundary between in-base and out-of-base periods, but resulting in opposite linear trends. The jumps (defined as the same as in Figure 2(b)) for the red and the black curves are 1.8 and 0.3%, respectively, both of which are negligible. However, the mean ERs for the base period in red and black curves are 10.00 and 9.92, respectively, showing that the latter is biased. The bias in ZHZK’s result is more obvious through the ensemble-mean monthly ERs further averaged over each calendar month (Figure 3(b)). In Figure 3(b), although the mean monthly ERs for in-base and out-of-base periods by ZHZK’s method are very close to each other (black curves), they seem to have a common half-year-periodic deviation from 10 such that they range from 9.3 to 10.6 within the year. This will significantly distort the monthly ER estimation in particular. On the contrary, although by QRM the mean monthly ERs for the out-of-base period are a little bit higher than those for in-base period (red curves), they are all very stable throughout the year; particularly, those for the in-base period are very close to 10, showing that even on a monthly basis the QRM performs just as well. As trends in climate extremes need to be analyzed on either a seasonal or an annual basis, to make the result credible, it is far from enough to avoid inhomogeneity by simply removing the jump in annual ER series. Rather, jumps and biases in ERs on both the monthly and annual bases should be reduced as much as possible.
3. Example results

We now apply the semi-parametric QRM to real GCM output to calculate the six temperature extremes indices listed in Table 1, and compare the results with their CLIMDEX counterparts. Ensemble member r1i1p1 under Historical (1950–2005) and RCP4.5 (2006–2100) greenhouse gas emission scenarios is chosen from simulations performed by Beijing Normal University Earth System Model (BNU-ESM) for CMIP5 (Ji et al., 2014) as an example for the comparison. BNU-ESM consists of fully coupled atmospheric, land, oceanic and sea ice components along with carbon cycles, with 128 × 64 horizontal dimensions and 26 vertical levels. Model (6) is first fit to the daily maximum and minimum temperature data during the base period 1961–1990 at each grid point individually, for the 10th and 90th percentiles, respectively; the fitted models are then used to predict these percentiles throughout the Historical and RCP4.5 scenarios. The six indices for each grid point in BNU-ESM are calculated accordingly.

Figure 4 shows comparisons of the six indices for BNU-ESM, globally averaged over the 8192 grid points, between the two sources. Linear trends under the RCP4.5 scenario are also given. It can be seen

Figure 4. Comparisons of the six temperature extremes indices for BNU-ESM output under Historical (1950–2005) and RCP4.5 (2006–2100) emission scenarios, globally averaged over the 8192 grid points, between the CLIMDEX datasets and the semi-parametric QRM-based results (curves); dashed lines are linear trends under the RCP4.5 scenario.
that changes in the first four indices from CLIMDEX datasets seem to be exaggerated. While the cold indices TX10p and TN10p go from about 16% in 1950 down to under 2% after 2050, the warmth indices TX90p and TN90p go from about 10% up to above 50% in the same time. In the QRM results, however, changes in cold and warmth indices throughout the whole period are only about 8 and 15%, respectively. Under the future RCP4.5 scenario, linear trends in TX90p and TN90p from CLIMDEX datasets are much larger than those from the QRM results. A possible reason for the seemingly exaggeration of future trends especially in the warmth indices from CLIMDEX datasets is that, ZHZK’s method tends to underestimate the annual ERs for the base period (Figure 3(a)); although it can reduce the jump, it cannot correct biases. If the data are periodically stationary, such as those simulated by the superposition of Equations (8) and (9), the trend estimation would still be reasonable because of the reduction of jumps. However, the CMIP5 simulations for various emission scenarios are absolutely non-stationary: their long-term trends in temperature under future scenarios are generally increasing. Once the annual ERs for the base period are underestimated, those for the future period may probably be overestimated, resulting in exaggerated trends in the warmth indices. As for cold spell duration indicator (CSDI) and warm spell duration indicator (WSDI) from the CLIMDEX datasets, as they were calculated still based on the empirical quantile estimator, the non-stationarity in data in addition to the jump effect can also result in changes similar to those in the first four indices as well. Therefore, trends in the three warmth indices under the future scenario may probably be overestimated using the CLIMDEX datasets.

4. Conclusions

In this article, a semi-parametric regression approach to quantile estimation for daily temperature data is proposed, and is applied to the calculation of the six ETCCDI temperature extremes indices using the CMIP5 BNU-ESM output as an example. Through Monte Carlo experiments, performances of empirical quantile estimator, ZHZK’s method and QRMs presented here are carefully examined and compared. The results show that although ZHZK’s method can effectively reduce jumps in the annual ER series, biases still exist, and are particularly obvious in the monthly ERs. The QRM approach, however, can keep both jumps and biases in ERs as small as negligible, on either a monthly or an annual basis, so that the derived linear trends are much more credible.

Further comparisons of the six indices for the BNU-ESM output between the CLIMDEX datasets and the QRM results are also carried out. The three warmth indices from the CLIMDEX datasets are probably biased such that their linear trends under the future RCP4.5 emission scenario seem to be overestimated, whereas those from the QRM results are much more reasonable. Owing to the fact that the ETCCDI indices have almost been standards for the description of changes in climate extremes, and that the CLIMDEX datasets and software has been widely used as convenience to climate change research community, in order to avoid drawing misleading conclusions for policy making, it is necessary to re-examine currently adopted algorithms and available datasets, and to develop new methods for generating the percentile-based ETCCDI indices.

Acknowledgement

This work was supported by the Fundamental Research Funds for the Central Universities of China (grant 2012LZD12).

References

Bronaugh D. 2015. climdex.pcic: PCIC Implementation of Climdex Routines. R package version 1.1-6. https://CRAN.R-project.org/package=climdex.pcic (accessed 22 July 2016).

Bürger G, Murdock TQ, Werner AT, Sobie SR, Cannon AJ. 2012. Downscaling extremes – an intercomparison of multiple statistical methods for present climate. Journal of Climate 25: 4366–4388.

Bürger G, Sobie SR, Cannon AJ, Werner AT, Murdock TQ. 2013. Downscaling extremes: an intercomparison of multiple methods for future climate. Journal of Climate 26: 3429–3449.

Dong SY, Xu Y, Zhou BT, Shi Y. 2015. Assessment of indices of temperature extremes simulated by multiple CMIP5 models over China. Advances in Atmospheric Sciences 32(8): 1077–1091.

Folland C, Anderson C. 2002. Estimating changing extremes using empirical ranking methods. Journal of Climate 15: 2954–2960.

Green PJ, Silverman BW. 1994. Nonparametric regression and generalized linear models. Chapman & Hall: London.

Hyndman RJ, Fan Y. 1996. Sample quantiles in statistical packages. The American Statistician 50: 361–367.

Ji D, Wang L, Feng J, Wu Q, Cheng H, Zhang Q, Yang J, Dong W, Dai Y, Gong D, Zhang RH, Wang X, Liu J, Moore JC, Chen D, Zhou M. 2014. Description and basic evaluation of Beijing Normal University Earth System Model (BNU-ESM) version 1. Geoscientific Model Development 7: 2039–2064.

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

Koenker RW. 2005. Quantile Regression. Cambridge University Press: Cambridge.

Kotz S, Kozubowski TJ, Podgorski K. 2001. The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance. Birkhauser: Boston, MA.

Li J, Zhang Q, Chen YD, Singh VP. 2013. GCMs-based spatiotemporal evolution of climate extremes during the 21st century in China. Journal of Geophysical Research. Atmospheres 118: 11,017–11,035.

R Core Team. 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria. https://www.R-project.org/ (accessed 22 July 2016).

Sillmann J, Kharin VV, Zwiers FW, Bronaugh D. 2013a. Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. Journal of Geophysical Research. Atmospheres 118: 1716–1733.

Sillmann J, Kharin VV, Bronaugh D. 2013b. Climate extremes indices in the CMIP5 multimodel ensemble: Part 2. Future climate projections. Journal of Geophysical Research. Atmospheres 118: 2473–2493.

Sun Z, Jones PD, Davies TD, Moberg A, Bergström H, Camuffo D, Cochoo C, Mauger M, Demarée GR, Verhoeve T, Thoen E, Barren dos M, Rodríguez R, Martín-Vide J, Yang C. 2002. Trends of extreme
temperatures in Europe and China based on daily observations. *Climatic Change* 53: 355–392.

Yee TW. 2015. *Vector Generalized Linear and Additive Models: With an Implementation in R*. Springer: New York, NY.

Yee TW. 2016. VGAM: Vector Generalized Linear and Additive Models. R package version 1.0-2. http://CRAN.R-project.org/package=VGAM (accessed 22 July 2016).

Zhang X, Hegerl GC, Zwiers FW, Kenyon J. 2005. Avoiding inhomogeneity in percentile-based indices of temperature extremes. *Journal of Climate* 18: 1641–1651.

Zhou B, Xu Y, Wu J, Dong S, Shi Y. 2016. Changes in temperature and precipitation extreme indices over China: analysis of a high-resolution grid dataset. *International Journal of Climatology*, 36: 1051–1066.