# Generating Climate Extremes Indices (CLIMDEX) by Using Semi-Parametric Quantile Regression Models

Yang Li and Chi Yang* (Email: chi@bnu.edu.cn)
*College of Global Change and Earth System Science, Beijing Normal University, Beijing 100875, China*

## 1. Introduction

The Expert Team on Climate Change Detection and Indices (ETCCDI) (http://wcrp-climate.org/etccdi) had defined a set of 27 core indices representing the more extreme aspects of climate. A suite of in situ and gridded land-based global datasets of such indices has been produced and is maintained by the CLIMDEX project (http://www.climdex.org/indices.html). Quite a lot of the indices are based on the 10th or the 90th percentiles of daily temperature and precipitation data during the base period 1961−1990. To estimate the percentiles for each calendar day, a 5-day window centered on the day of interest is used throughout the base period so that 5×30 data are available. The percentiles are further used to calculate exceedance rate (ER) for both in-base and out-of-base periods.

The above procedure results in a jump in ER at the boundary between the in- and out-of-base periods (Zhang et al., 2005, referred to as ZHZK hereafter). An experiment with artificial data generated by AR(1) process

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

showed the jump clearly (Fig. 1 in ZHZK). ZHZK proposed an additional bootstrap resampling procedure to remove the inhomogeneity in the exceedance series. The whole steps had been taken as standard procedure by CLIMDEX for generating the datasets of indices.

Basically, ZHZK's method is empirical and lacks systematicness. The algorithm is designed on a daily basis. Data sample must be updated for each calendar day. Practitioners are difficult to implement the procedure by themselves except by using the developed software. Here we present a model-based approach to the generation of CLIMDEX indices. This approach is systematic, resistant to inhomogeneity (and thus no need to make up with extra steps), and easy to implement.

## 2. Quantile regression models

By building a quantile regression model (QRM) a certain quantile of an observed sample is regressed to some additive functions of covariates. Unlike the usual regression models in which it is the distribution parameters that are to be regressed, it is not necessary to assume a distribution for the response variable. To estimate percentiles for the CLIMDEX indices, the in-base period data will be used as a whole and no moving window is needed, so that the result is much less liable to be biased.

First look at the result for the artificial data generated by model (1). The QRM is simply

$$q_{0.9t} = \beta_0 + \beta_1 x_{t-1} \tag{2}$$

Once the coefficients are estimated, the 90th percentiles $\mathbf{q}_{0.9}$ for both in- and out-of-base periods can be 'predicted' by model (2). The QRM fitting has been implemented in many statistical softwares such as R, without any complicated scripting.

Fig. 1

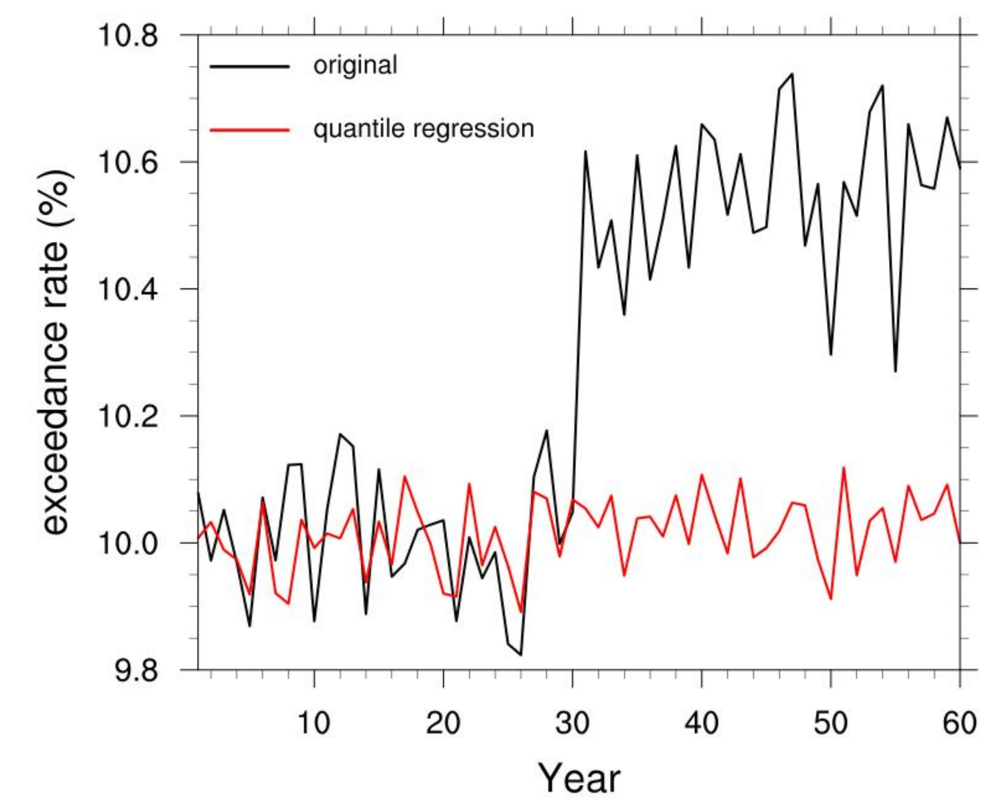


Fig.1 shows average ER of daily values greater than 90th quantile in 1000 simulations of model (1) by using the original moving window method and model (2), respectively. The estimated ER by model (2) is homogenous for the whole period and shows no jump at the boundary between the two half periods.

In applying QRM to real data such as daily temperature, The in-base period data should be used untouched and as a whole. There is no need to subtract the climatology from the data. Rather, the climatology will be modeled as a cyclic function of calendar days approximated by smoothing regression splines. A semi-parametric QRM for daily temperature taking climatology and autoregression into account looks like

$$q_{0.9t} = \beta_0 + f(d_t) + \beta_1 T_{t-1} + \beta_2 T_{t-2} + \beta_3 T_{t-3} \tag{3}$$

where $f(d_t)$ is a cyclic function of Julian date $d_t$ of the year for the $t$th datum of the whole series; $T_{t-1}$, $T_{t-2}$ and $T_{t-3}$ are lag 1-, 2- and 3-day temperatures.

Fig. 2

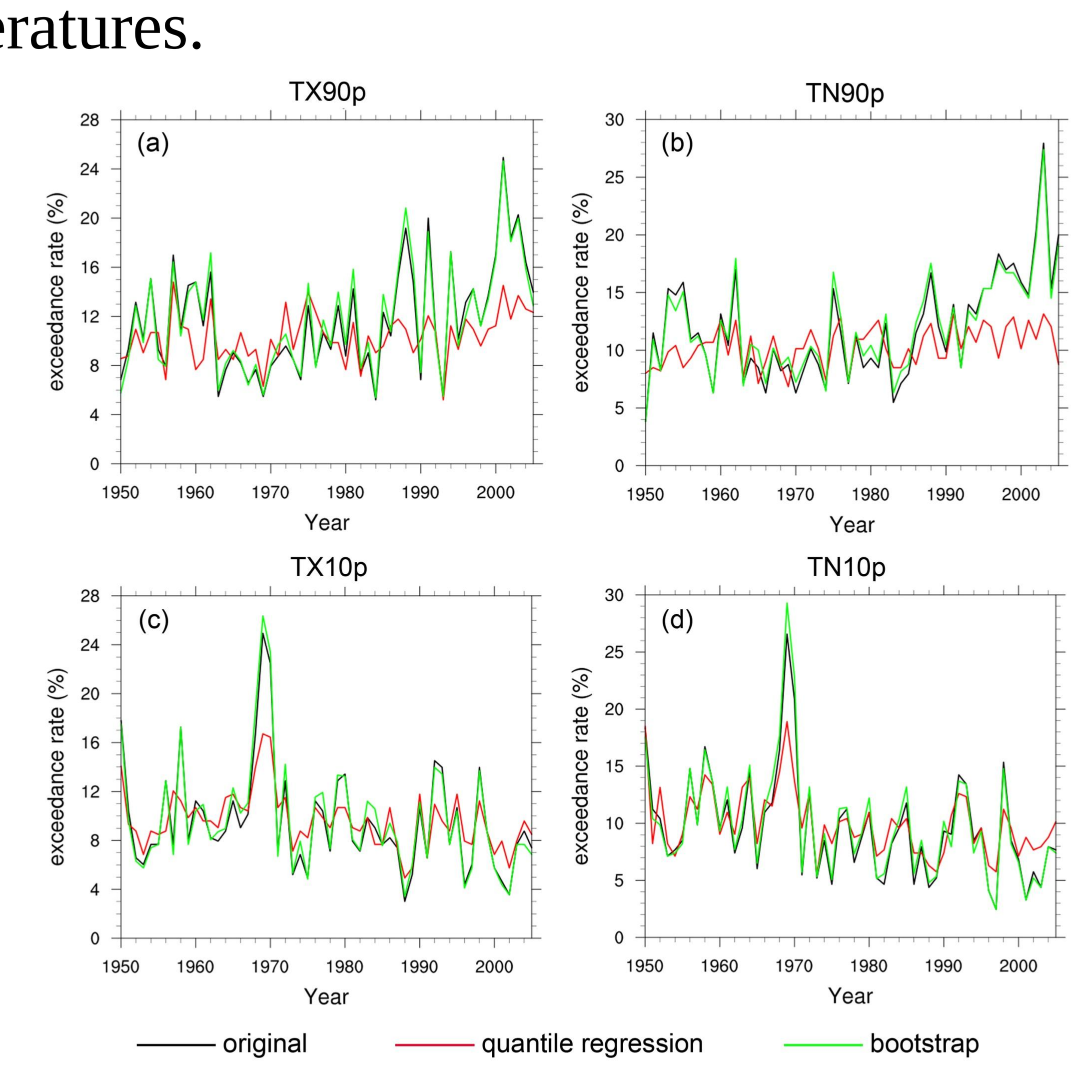


Fig.2 shows 4 indices for the grid point at 118.125E, 32.1N based on the BNU-ESM historical simulation, produced by the moving window, ZHZK and model (3), respectively. The model-predicted series (red line) has the least variation among the three. Its effect on trend analysis needs further investigation.

## 3. Summary

A semi-parametric QRM approach to generating percentile series for daily data is proposed. No jump appeared at the boundary between in- and out-of-base periods of derived ER series .

**Reference**
Zhang, X., et al. (2005): Avoiding Inhomogeneity in Percentile-Based Indices of Temperature Extremes. *J. Climate*, 18, 1641-1651.