The importance of mean and variance in predicting changes in temperature extremes — Source link

Sylvie Parey, Thi Thu Huong Hoang, Didier Dacunha-Castelle

Institutions: University of Paris-Sud

Published on: 16 Aug 2013 - Journal of Geophysical Research (John Wiley & Sons, Ltd)

Topics: Standard deviation and Generalized extreme value distribution

The importance of mean and variance in predicting changes in temperature extremes

Sylvie Parey, Thi Thu Huong Hoang, Didier Dacunha-Castelle

The importance of mean and variance in predicting changes in temperature extremes

Sylvie Parey and Thi Thu Huong Hoang, EDF/R&D, France

Didier Dacunha-Castelle, Laboratoire de Mathématiques, Université Paris 11, Orsay, France

Corresponding author: Sylvie Parey, EDF/R&D, Chatou, France (sylvie.parey@edf.fr)

Abstract The important role of the evolution of mean temperature in the changes of extremes has been recently documented in the literature, and variability is known to play a role in the occurrence of extremes too. This paper aims at further investigating the role of their evolutions in the observed changes of temperature extremes. Analyses are based on temperature time series for Eurasia and the United States and concern absolute minima in winter and absolute maxima in summer of daily minimum and maximum temperature. A test is designed to check whether the extremes of the residuals after accounting for a time-varying mean and standard deviation can be considered stationary. This hypothesis is generally true for all extremes, seasons and locations. Then, the comparison between the directly fitted parameters and the retrieved ones from those of the residuals compare favorably. Finally, a method is proposed to compute future return levels from the stationary return levels of the residuals and the projected mean and variance at the desired time horizon. Comparisons with return levels obtained through the extrapolation of significant linear trends identified in the parameters of the GEV distribution show that the proposed method gives relevant results. It allows taking mean and/or variance trends into account in the estimation of extremes even though no significant trends in the GEV parameters can be identified. Moreover, the role of trends in variance cannot be neglected. Lastly, first results based on two CMIP5 climate models show that the identified link between mean and variance trends and trends in extremes is correctly reproduced by the models and is maintained in the future.

1 Introduction

Global temperature has increased since the beginning of the last century and will most likely continue to do so in the next decades [*IPCC,* 2007]. This increasing trend may induce more frequent and more intense heat waves in the future [*Meehl and Tebaldi*, 2004; *Fischer and Schaer*, 2010; *Barriopedro et al.*, 2011]. *Coumou and Rahmstorf* [2012] recently showed that the unprecedented occurrence of record-breaking events in the last decade can be attributed to anthropogenic climate change. As temperature extremes may cause multiple severe social and economic impacts, their evolutions have been studied using different approaches. Some studies are based on the analysis of observed daily data, recently made available through homogenized or, at least, scrutinized series regarding homogeneity, like the European Climate Assessment and Dataset (ECA&D) project series or the *Caesar et al.* [2006] gridded dataset. Important decreases are found in the number of frost days, while coherent increases appear in extreme night time temperatures [*Alexander et al,* 2006; *Frich et al.*, 2002]. Generally, trends in extreme night time temperature are higher than trends in day time maximum temperature, and the warming is largest in the northern hemisphere during winter and spring. Moreover, *Kiktev et al.* [2003] showed that these evolutions are linked to anthropogenic greenhouse gas emissions. It is thus clear that the highest and lowest temperatures exhibit trends all over the world. One question thus concerns the link between these trends and that of the mean and/or of other moments of the distribution.

This question has been tackled by *Barbosa et al.* [2011], for daily mean temperature in Central Europe using quantile regression and clustering. They showed that for most of their studied stations, the slopes of the lowest and highest quantiles are not the same as those of the median, and thus that the trends are not the same for all parts of the distribution. Using a different approach, *Ballester et al.* [2010a] analyzed the link between trends in extreme and in mean temperature. Using climate simulation results from the European PRUDENCE project and the E-OBS gridded observation dataset [*Haylock et al.*, 2008] they showed that the increasing intensity of the most damaging summer heat waves over Central Europe is mostly linked to higher base summer temperatures.

Few papers analyze the most extreme events using statistical Extreme Value Theory (EVT). *Zwiers et al.* [2011] used Generalized Extreme Value (GEV) distributions and climate model simulations of the CMIP3 project database to detect anthropogenic influence. They found that the most detectable influence of external forcing is on annual maximum daily minimum temperature (TN) and the least detectable on annual maximum daily maximum temperature (TX). They also stated that the waiting time for the 1960's 20-year return level (expected to recur once every 20 years) has now increased for annual minimum TX and TN and decreased for annual maximum TN. *Brown et al.* [2008] went further in studying the link between the identified trends in extreme and in mean temperature. They used an EVT-model with time varying parameters to study the global changes in extreme daily temperatures since 1950 from the *Caesar et al.* [2006] gridded daily dataset. Applying the Marked Point Process technique, they found that only trends in the location parameter are significant and that both maximum and minimum TN present higher trends than their TX counterparts. They then compared the trends in the location parameter to the trends in mean, and found that the trends in extremes are consistent with the trends in mean.

Starting from these results, this paper aims at going further in researching the link between the evolutions of extremes and of the bulk of the distribution of temperature. It can obviously be expected that if the mean is changing, the induced shift of the tails of the distribution will lead to changes in extremes. *Katz and Brown* [1992] and *Fisher and Schär* [2009] highlighted the role of variability in the occurrence of extremes. Other moments of the distribution could be studied. For example, *Ballester et al.* [2010b] use standard deviation and skewness of the annual distribution of detrended temperature. Using climate model simulation results only, they stress the role of standard deviation change in the modification of frequency, intensity and duration of warm events, whereas skewness change is also important for cold extremes.

This study focuses on the estimation of temperature extremes in the climate change context. One commonly used methodology relies on the identification and estimation of trends in the parameters of the EVT distributions [*Coles,* 2001; *Parey et al.,* 2007; *Parey et al.,* 2010b]. However, such trends are identified on relatively short samples made of the highest (or lowest) observed values and may not be as robust as trends identified on the whole dataset. Therefore a systematic study of the link between trends in extremes and trends in mean and variance is helpful to determine whether extremes exhibit unique trends in addition to those induced by trends in mean and variance. If they do not, future extremes can be derived from the stationary extremes of the residuals, after accounting for a time-varying mean and standard deviation, and the changes in mean and variance of the whole dataset, as proposed in Parey et al. 2010b. The aim of this paper is then to check this link for a large number of time series of temperature from weather stations. It will therefore be organized as follows: section 2 is dedicated to the observational data and section 3 to methods descriptions. The link between the non-parametric trends in mean and variance and in extremes is investigated and discussed in section 4, as well as its use in the estimation of future return levels, before concluding with a discussion and perspectives in section 5.

2 Observational data

For Eurasia, weather station time series are taken from the ECA&D project database. The project gives indications of homogeneity through the results of different break identification techniques [*Klein Tank et al.,* 2002]. For this study the series which could be considered as homogenous (stated as "useful" in the database) over the period1950-2009 have first been selected for both TN and TX. Then, these series have been checked for missing data and those with more than 5% missing data have again been excluded. This selection left 106 series for TX and 120 for TN (many TX series, mostly in Russia, have missing values from 2007 onward whereas the corresponding TN series have missing values only in 2009).

For the United States, weather station TX and TN time series are obtained from the Global Historical Climatology Network – Daily Database (GHCN daily) [*Menne et al.,* 2011]. These time series have been quality checked through an automated quality insurance described in *Durre et al.* [2010]. The first step has been to select the highest quality time series, as stated by the quality indicators, with less than 5% of missing data. Then, only the series starting before 1966 and ending after 2008 are kept. Finally a new check-up for missing values has been conducted, together with a visualization of the evolution of annual mean values. ~~The TX time series for the station of Eureka (Arizona) and the TN time series for Ajo (California)~~ **One TX and one TN time series present** a stepwise-like evolution between 1970 and 1980 looking like a break and have been eliminated ~~(figure 1)~~, which leaves us with 86 series for TX and 85 for TN.

3 Statistical methods

3.1 Extreme value theory

EVT relies on the well known Extremal Types Theorem which states that, if the maximum of a large sample of observations, suitably normalized, converges in distribution to G when the sample size tends to infinity, then G belongs to the GEV family [*Coles,* 2001]. The assumptions behind the theorem are that the data in every block are stationary and weakly dependent with a regular tail distribution. Temperature maxima are expected to occur mostly in summer and temperature minima in winter. For each time series, the distribution of the 2, 3 or 5 highest or lowest values each year in the different months is computed. Then the months with more extremes than expected under the identical distribution assumption are selected. For maximal TN or TX, the months of June, July, August or July, August, September occur quite regularly as the favored ones, and thus the summer season is defined as a period of 100 days between the 14^{th} of June and the 21^{st} of September. The selection of 100 days is convenient but may appear somewhat arbitrary. However, it is a good compromise between length and weak remaining seasonality. In fact, tests with different selections in these months of June to September showed that the results are not sensitive to this choice (not shown). For minimal TN or TX, the minima rather occur during the month of January, followed by December or February, but no other months emerge. Thus the winter season is defined as the 90 days of the months of December, January and February (the 29 February is omitted during leap years, except if the temperature is lower than that of the 28 in which case it is considered as the temperature of the 28). Then the choice of block length is based on the classical bias / variance trade-off. Defining 2 blocks per season (blocks of 50 days in summer and 45 days in winter) have been chosen as a reasonable balance, leading, with series of around 50 to 60 years to more than 100 block maxima or minima.

Thus the GEV distribution will be fitted to the maxima of TN and TX in summer and the minima of TN and TX (maxima of the opposite series) in winter considering 2 blocks per season.

3.2 Trends

3.2.1 Non-parametric trends in mean and variance

Let X(t) be an observed temperature time series. For each day t, m(t) and s^2(t) (continuous time functions) represent the associated mean and variance, respectively. If Γ(t) is a (k,T) matrix, where T is the length of the time period**,** whose components are associated to different characteristics of the process at time t, then Γ(t) is called multidimensional trend [*Hoang et al.,* 2009]. For instance, Γ(t) consists here of the trends in mean and standard deviation, but skewness and kurtosis trends could also be considered. The goal is to estimate as objectively as possible Γ(t), in order to capture the structure in the data and in the same time, to smooth local extrema. As in *Hoang et al.* [2009] or in *Parey et al.* [2010a and b], the LOESS (Local regression, *Stone*, 1977) technique is used to do so. The choice of the smoothing parameter (and thus the window length) has to be adapted to the analyzed data to keep the trend identification as intrinsic as possible. This is made by using a modified partitioned cross-validation (MPCV) technique [*Hoang,* 2010]. Cross-validation has to be modified in order to eliminate as far as possible time dependence and take heteroscedasticity into account. The idea of MPCV is to partition the observations into *g* subgroups by taking every g^{th} observations, for example the first subgroup consists of observations 1, 1+*g*, 1+2*g*,..., the second subgroup consists of observations 2, 2+*g*, 2+2*g*,... The observations in each subgroup are then independent for high *g*. *Chu and Marron* [1991] define the optimal bandwidth for Partitioned Cross-Validation in the case of constant variance as $h_{PCV} = h_0 g^{1/5}$, with h_0 estimated as the minimiser of $PCV_g(h) = \frac{1}{g}\sum_{k=1}^{g} CV_{0,k}(h)$ ($CV_{0,k}$ is the ordinary Cross-Validation score for the k-th group). This approach has been modified to take heterocedasticity into account. Then, the optimal *g* corresponds to the minimum of a more complicated expression [*Hoang*, *2010*] and in practice, it is preferred to estimate h_{MPCV} (the optimal bandwidth of the Modified Partitioned Cross Validation) for different values of *g* and to retain the values of *g* for which h_{MPCV} is not too bad (that is not too close to zero and not higher than 0.7). For each *g* the trends m and s are estimated by loess with bandwidth $\hat{h}^g_{MPCV}$ to obtain an estimator of the expression to minimize. The value of *g* corresponding to the minimum value is retained, giving the corresponding optimal bandwidth h_{MPCV}. Up to now, this seems to be the best way to estimate the optimal bandwidth in this situation for which mathematical theory is not complete. For temperature, the dependence between the dates can be assumed as negligible if the dates are distant by more than 5 days. We used a cross validation method on data sampled every 10 days (g=10) to be conservative, and an optimal parameter is computed for each temperature time series.

3.2.2 Non-parametric trends in extremes

In the same way, if EVT can be applied and G(t) is the GEV distribution at time t, $\Theta(t)$ represents the parameters of G(t), that is location $\mu(t)$, scale $\sigma(t)$ and shape $\xi(t)$. The shape parameter ξ is the most difficult to estimate, and it could be tricky to differentiate possible evolutions from estimation errors. In their study, *Zhang et al.* [2004] did not consider any trend in this parameter, as they assume that it is not likely to show a trend in climate series. Tests on different periods of a long observation series have shown that this parameter does not significantly evolve with time [*Parey et al.,* 2007], and more sophisticated non-parametric studies lead to the same conclusion [*Hoang*, 2010]. Thus, in the following, the shape parameter ξ will be considered constant. Then, the trends in location and scale parameters are estimated in a non-parametric way using cubic splines (through penalized likelihood maximization, *Cox and O'Sullivan* [1996]) and the classical cross validation technique (in an iterative way) since the extremes are selected as independent values. Cubic splines are preferred here because they are convenient to deal with edge effects for the relatively short series of maxima. An iterative procedure is used to smooth both the location and scale parameters consistently. The estimation of constant parameters is obtained through likelihood maximization (see section 3.3).

3.3 Stationarity test

The question we wish to address is whether trends in extremes can mostly be characterized by trends in mean and variance. To analyse this, Y(t) is defined as the standardized residuals:

$$Y(t) = \frac{X(t) - m(t)}{s(t)} \qquad (1)$$

The hypothesis we want to test becomes: "the extremes of Y(t) in every block can be considered as a stationary sequence", which means that both the location μ and scale σ parameters are constant. A methodology to test this hypothesis has been proposed and detailed in *Hoang* [2010] and is summarized here. First, Y(t) is estimated as $\hat{Y}(t) = \frac{X(t) - \hat{m}(t)}{\hat{s}(t)}$ and the stationarity of its extremes is tested. The set of possible evolutions of the extreme parameters of Y(t) is very large. So the test cannot easily be formulated as a choice between two well defined alternatives. This is the reason why the use of a squared distance Δ between two functions of time, defined as:

$$\Delta(f, g) = \int_{t \in D} (f(t) - g(t))^2 dt \text{ , D being the time period,} \qquad (2)$$

is preferred. If any function of time *f* is estimated by *g*, Δ(f,g) is a measure of the quality of *g* as an estimate of *f*. Two different estimations of the parameters μ(t) and σ(t) can be made: they can be estimated non-parametrically as $\tilde{\mu}(t)$ and $\tilde{\sigma}(t)$ or as constant as $\hat{\mu}$, $\hat{\sigma}$. The stationarity hypothesis being true or not, $\tilde{\mu}(t)$ and $\tilde{\sigma}(t)$ converges to the 'real' values μ, σ when the sample size T tends to infinity, the rate of convergence depends on the supposed smoothness of the function. The situation is of course different for $\hat{\mu}, \hat{\sigma}$: if the stationarity hypothesis is true, they converge to μ, σ with a rate of the order of $\sqrt{T}$ and in this case $\Delta(\hat{\mu}, \tilde{\mu})$ is, for a large sample, very close to $\Delta(\mu, \tilde{\mu})$. On the contrary if the hypothesis is false, $\hat{\mu}$ converges to a constant which is of course different from the non constant function μ(t) and $\Delta(\hat{\mu}, \tilde{\mu})$ does not tend to zero and remains larger than some A>0. The intuitive reason is that we try to find μ in a set of functions "far away" from μ if the hypothesis is false. The same is true for Δ($\hat{\sigma}$, $\tilde{\sigma}$). A test could be based on an asymptotic result [*Hoang*, 2010]. We prefer the use of a numerical approach based on simulation. Our proposed solution is then to statistically evaluate (by simulation or bootstrapping) the distribution of Δ($\hat{\mu}$, $\tilde{\mu}$) if the hypothesis is true, that is the distribution of the distances between the non-parametric estimates and the best constant to estimate μ. To do this, we simulate a large number of samples of the stationary GEV (μ_Y,σ_Y,ξ_Y) distribution with the same size as the series of the maxima of Y(t). From each sample, we estimate the GEV parameters in two ways: first, by considering them as constant; second, by considering them as functions of time. Then we calculate the distances between these two estimates and obtain a distribution of the statistical error of estimation provided the hypothesis is true. If the distances obtained from the observations are found lower than the 90th percentile, then the hypothesis is considered satisfied: the distances cannot be distinguished from such arising due to statistical errors. The power of the test has been evaluated and is reasonable (see appendix).

4 Results for temperature time series

4.1 Stationarity test

Brown et al. [2008], among others, have shown that significant trends can be identified in the evolutions of temperature extremes, especially the location parameter. The investigated issue is whether these trends can mostly be characterized by trends in mean and variance. Therefore, the previously described test has been applied to different temperature time series for different variables (TN and TX), parameters (location and scale) and locations (Eurasia and the United States).

The results are shown in figure ~~2~~ **1**. Grey points indicate that the cross validation could not converge to an optimal smoothing parameter for the non-parametric estimation of the location and scale parameters, and thus, the test could not be performed. This mostly happens in winter in the United-States: around 20% of the stations (18.8% for minimal TN and 19.8% for minimal TX) experience this problem. The reason for this will have to be more carefully investigated in future work. For the other seasons and locations, this concerns less or around 10% of the stations. Among points where the test could be performed, the hypothesis is accepted for both location and scale parameters for around 80 to 90% of the stations (from 76.6% for maximum TN in summer in the United-States to 94.2% for minimum TN in winter in the United-States), and for at least one of the parameters for more than 94% of the stations (from 94.7% for maximum TX in summer in the United-States to 100% for minimum TX and minimum TN in winter in the United-States and minimum TX in winter in Eurasia). This means that the stationarity of the extremes of the standardized residuals can reasonably be assumed globally.

4.2 Impact on Return Level estimation

Previous results show that the trends in extremes closely follow that of mean and variance. The extreme distribution parameters of the observed temperature time series X(t) are linked to those of the standardized residuals Y(t) in the following way:

$$\begin{cases} \xi_X = \xi_Y \\ \sigma_X(t) = \sigma_Y(t) * s(t) \\ \mu_X(t) = m(t) + \mu_Y(t) * s(t) \end{cases} \qquad (3)$$

where μ, σ and ξ are respectively the location, scale and shape parameters of the GEV distribution, subscripts X and Y referring to the observed temperature time series and the residuals time series, and m(t) and s(t) are the trends in mean and standard deviation. We thus first compare**d** the non-parametric GEV parameters directly obtained from X(t), with their bootstrap confidence intervals, to the same parameters reconstructed from the constant Y(t) parameters and the non-parametric trends in mean and standard deviation of X(t) by using (3). ~~The plot obtained for the French station of Déols in figure 3 shows~~ **The obtained results** show that the reconstructed parameters **are reasonably comparable to the directly estimated ones (not shown)** ~~fall most of the time inside the 95% bootstrap confidence interval of the directly computed ones~~, which checks the validity of the tested hypothesis.

Then, the GEV parameters for a given future period can be derived from those of Y(t), which are constant, and future values of the mean and the standard deviation, to compute some future Return Level (RL), as proposed in *Parey et al.* [2010b].

As an example, 50-year RLs are computed for the year 2030 for TX in Eurasia:

1) through extrapolation of optimal linear trends (according to a likelihood ratio test with a ~~10% significance~~ **90% confidence** level) in location and scale parameters of the GEV for X(t)
2) through (3) with m(t) and s(t) being significant linear trends extrapolated to 2030 (future m and s are computed over 10 years around 2030). Trend significance is assessed with a Mann-Kendall test on seasonal means and variances with a ~~10% significance~~ **90% confidence** level.

In each case, confidence intervals are computed by bootstrapping, in order to take uncertainties in the identified trends into account. The obtained differences in RL do not exceed 3°C, and method 2 generally gives higher RLs. The confidence intervals of the two methods do not overlap for 16 out of the 106 TX time series (figure ~~4~~ **2**). The confidence intervals are said “not overlapping” if the RL computed with method 1 does not fall in the confidence interval of the RL computed with the method 2 and vice-versa. This avoids choosing a threshold to eliminate small overlapping. For 14 of them, no trends are found in the GEV parameters but a significant trend in mean, in variance or in both mean and variance is identified, and for the 2 others a significant trend is found for the location parameter of the GEV and in mean and variance. For these 16 TX time series, the second approach leads to a higher RL, except for Gurteen in Ireland (open red circle in figure ~~4~~ **2**). This can be explained by differences in the shape parameter obtained for the extremes of X(t) and those of Y(t) in this case. Theoretically, the shape parameters are identical (equation 3), but due to adjustment uncertainties, in practice, it may not be the case (the confidence intervals are large for this parameter). For the Gurteen TX time series $\xi_X = -0.13$ and $\xi_Y = -0.33$. If the RL is computed with $\xi_Y = \xi_X$ with method 2, then the two confidence intervals do overlap.

The role of a trend in variance can be illustrated by the TX time series of Dresden and Berlin in Germany. For these two time series, no significant trends are identified in the location and scale parameters of the GEV. If the non-parametric trends are drawn for these parameters, it can be seen that they show a small increasing trend, which is not found significant through the likelihood ratio test when looking for a linear trend (figure ~~5~~ **3**). The two time series differ regarding the mean and variance evolutions: whereas in Berlin a significant linear trend is found for both mean and variance, in Dresden, only the linear trend in mean is significant (figure ~~6~~ **4**). Then, the 50-year RL in Dresden computed with method 2 falls inside the confidence interval of the RL computed with method 1:

Method 1: RL=36.9°C [35.7;38.1]   Method 2: RL=37.8 [36.3;38.7]

whereas in Berlin, it does not:

Method 1: RL=38.2°C [37.2;39.3]   Method 2: RL=40.9°C [39.1;42.4]

The proposed method based on mean and variance trends allows taking changes in extremes into account, even though no significant trends in the GEV parameters are identified. Furthermore, the role of a variance change in the computed RL is not negligible and has to be taken into account.

4.3 First results with climate models

A preliminary study has been made with climate model results to check:

- whether the stationarity of the extremes of the residuals found with observations is reproduced
- whether this stationarity remains true in the future with continued increasing greenhouse gas emissions

The TN and TX **daily** time series for Eurasia and the United States for only two CMIP5 model simulations have been considered: IPSL-CM5B-LR and CNRM-CM5 (made available by the French teams of the Institut Pierre Simon Laplace and Météo-France/CERFACS), with the highest RCP8.5 emission scenario. For both models, the historical period is 1950-2005 and the considered future period extends from 2006 to 2100 for IPSL-CM5B-LR and from 2006 to 2060 for CNRM-CM5 (the downloaded results concern this period only, although the model simulations run to the end of the century). ~~Because the computation of the test is time consuming (500 simulations are done for each temperature time series), all grid points time series could not be considered for testing~~. **The interest here is on local extremes behavior, and thus grid point time series have to be considered. However, temperature shows important spatial correlations, and coherent regions can easily be identified. Therefore, it does not seem necessary to compute the test for all grid points, especially for the highest resolution models. Thus,** ~~O~~only the land grid points are considered, and among those, all are tested in the US and only one over two points in longitude for Eurasia for IPSL-CM5B-LR. For CNRM-CM5 one land point over two in the US and one over two in longitude in Eurasia are used for testing**, since this model grid has a higher resolution**. The results obtained for minimum TN in winter and maximum TX in summer show that for both periods and both models, our hypothesis is likely to be true (figures **5** ~~7~~ and **6** ~~8~~). This means that these models reliably reproduce the observed link between trends in extremes and trends in mean and variance, and maintain it in the future. This has the interesting consequence that future RLs can be computed with our proposed method by using climate model results, and thus, projections are possible at later time horizons, which is not reasonably possible when extrapolating observed linear trends. **This is however a very preliminary insight, a more complete study of the behavior of climate models regarding this link will have to be further investigated by considering more models and by better designing the testing methodology for an optimal set of grid points.**

5 Discussion and perspectives

In this paper, two sets of observed temperature time series, in Eurasia and in the United States, chosen to be as homogenous as possible over the period 1950-2009, have been used to extend studies on the role of mean and variance change in the evolutions of temperature extremes. **Only point-wise analyses are made first to avoid smoothing the extremes by spatial averages and secondly because return levels are required, in practice, for specific locations.**

~~This role may be well known, but here~~ **Although the role of mean and variance in the evolution of extremes has been previously documented, here** a test is proposed and applied to check the stationarity of the extremes of the residuals. The results show that, for **local daily** temperature, trends in mean and variance mostly explain the trends in extremes for both TN and TX, in winter and in summer, and in Eurasia and in the United States. This allows estimating future return levels from the stationary return levels of the residuals and the projected mean and variance at the desired future period. Trends in mean and variance are more robustly estimated than trends in the parameters of the extreme value distribution, as they rely on much larger samples. Then, in case significant trends in the parameters of the GEV distribution cannot be detected, this method allows computing the future return levels in taking mean and/or variance trends into account. Furthermore, some significant trends in variance are found and their impact on the estimated future return level is not negligible. One practical difficulty with the proposed method lies in the fitting of the shape parameters: although the shape parameters of the observed time series and of the residuals are theoretically the same, practically they may differ and induce differences in the return levels. If this happens, it is advised to consider the lowest of both values as the same shape parameter for both time series.

These results, and especially the identified trends in variance and their role in the evolution of extremes, although coherent with most of the previous

findings, seem to contradict some recent ones (Simolo et al., 2011; Rhines and Huybers, 2013). However, Rhines and Huybers 2013, following and commenting Hansen et al. 2012, analyze summer mean temperatures and discuss the role of changes in mean and variance in the recent occurrence of very hot summers. They conclude that variance does not change, but the variance they consider is rather interannual variability, whereas in the present paper, variance means daily variability. They indeed acknowledge that their analysis "pertains only to summer averages and that other analyses based on, for example, shorter-term heat waves or droughts, may yield different results." In Simolo et al. 2011, the study is made on spatial averages over three different sub-domains and deals with so called "soft extremes", that is high and low percentiles of the temperature distributions. Spatial averaging necessarily leads to a reduction in variance and a smoothing of extreme events. On the other hand, our study is devoted to more extreme events through the application of EVT. It is thus very difficult to compare the results.

~~Then~~ **Finally**, the reproduction by two climate models of the identified link between trends in mean and variance and trend in extremes for temperature has been verified. Moreover, the same models maintain the validity of the link in the future, until 2100, which allows the use of the proposed method to estimate future return levels based on model projected mean and variance at any desired future horizon. **The analysis of climate models behavior regarding this link needs however to be further investigated using more models and a more robust testing methodology. Physical mechanisms able to explain such a link need furthermore to be identified.**

These findings are important for practical applications, because most safety regulations are based on the estimation of rare events, defined as long period return levels. In the climate change context, at least for temperature, it is not yet possible to apply EVT as if the time series were stationary to make such estimations. The proposed method is a way of tackling this problem.

~~Only point-wise results are shown, and it could be interesting to further investigate field significances. However in practice, return levels are often required for specific locations.~~

6. Appendix: power of the test

A synthetic study is presented to check the ability of the test to assess stationarity of the GEV parameters. To do so, 1000 samples are drawn from a distribution with imposed trends in mean and standard deviation, but not in extremes:

X(t) = m(t) + s(t)ε, where m(t)=*a*t+*b* and s(t)=*c*t+*d* and ε is drawn from a GEV distribution with location 0, scale 1 and shape -0.15. Coefficients *a* to *d* has been chosen to be reasonable for temperature: $a=3.8*10^{-4}$; $b=23.8$; $c=4.4*10^{-5}$; $d=4.4$. For each sample, m(t) and s(t) are re-estimated through LOESS with a smoothing parameter of 0.17 to compute the residuals Y(t). Then non-parametric and constant GEV parameters for the extremes of Y(t) are computed in the previously described way, and the table of distances under stationarity is calculated, to test whether the GEV parameters are found constant, with a 10% significance level. The non-parametric (splines) estimates of the GEV parameters converge for 943 of the 1000 samples. Among these, the test accepts the stationarity of μ for 925 samples (98%), the stationarity of σ for 846 (≅90%) and the stationarity of both μ and σ for 837 samples (≅89%), which results in around 10% false rejection, coherent with the 10% significance level used.

Now, to compute the power of the test, we consider a sample for which stationarity is rejected. We then compute 500 distances between constant and non-parametric estimates of the GEV parameters of the extremes of Y(t) for a non stationary GEV and count the number of times the distance falls in the rejection region of the table computed with a stationary GEV. 84.4% of these distances fall in the rejection region, which gives a power of 84.4%.

Acknowledgments

The authors acknowledge the data providers in the ECA&D project (http://eca.knmi.nl), in the National Climatic Data Center in NOAA (www.ncdc.noaa.gov). We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

7 References

Alexander, L.V., X. Zhang, T.C. Peterson, J. Caesar, B. Gleason, A.M.G. Klein Tank, M. Haylock, D. Collins, B. Trewin, F. Rahimzadeh, A. Tagipour, K.R. Kumar, J. Revadekar, G. Griffiths, L. Vincent, D.B. Stephenson, J. Burn, E. Aguilar, M. Brunet, M. Taylor, M. New, P. Zhai, M. Rusticucci, and J.L. Vazquez-Aguirre (2006), Global observed changes in daily climate extremes of temperature and precipitation. Journal of Geophysical Research-Atmospheres, 111(D05109), doi:10.1029/2005JD006290

Ballester J., F. Giorgi and X. Rodo (2010b), Changes in European temperature extremes can be predicted from changes in PDF central statistics. Climatic change, vol 98, 1-2: 277-284, DOI: 10.1007/s10584-009-9758-0

Ballester J., X. Rodo and F. Giorgi (2010a), Future changes in Central Europe heat waves expected to mostly follow summer mean. Climate Dynamics, 35:1191–1205, DOI 10.1007/s00382-009-0641-5

Barriopedro D., Fischer E. M., Luterbacher J., Trigo R. M., Garcia-Herrera R. (2001), The hot summer of 2012: redrawing the temperature record map of Europe. Science 332, DOI: 10.1126/science.1201224

Barbosa S.M., M.G. Scotto and A.M. Alonso (2011), Summarising changes in air temperature over Central Europe by quantile regression and clustering. Natural Hazards and Earth System Sciences, 11, 3227-3233, doi:10.5194/nhess-11-3227-2011

Brown, S.J., J. Caesar and C.A.T. Ferro (2008), Global changes in extreme daily temperature since 1950. Journal of Geophysical Research-Atmospheres, 113(D05115), doi:10.1029/2006JD008091

Caesar, J., L. Alexander and R. Vose (2006), Large-scale changes in observed daily maximum and minimum temperatures: Creation and analysis of a new gridded data set. Journal of Geophysical Research-Atmospheres, 111(D05101), doi:10.1029/2005JD006280

Chu C. K. and Marron J. S. (1991), Comparison of two bandwidth selectors with dependent errors. Annals of Statistics, 19: 1906-1918

Coles S (2001), An introduction to statistical modelling of extreme values, springer series in statistics. Springer Verlag, London

Coumou D. and S. Rahmstorf S. (2012), A decade of weather extremes. Nature climate change, DOI:10.1038/NCLIMATE1452

Cox D. D. and O'Sullivan F. (1996), Penalized likelihood type estimators for generalized nonparametric regression. Journal of multivariate analysis, 516: 185-206

Durre I., Menne M. J., Gleason B. E., Houston T. G., Vose R. S. (2010), Comprehensive Automated Quality Assurance of Daily Surface Observations, Journal of Applied Meteorology and Climatology, 1615-1633

Fischer, E.M. and Schär C., (2009), Future changes in daily summer temperature variability: driving processes and role for temperature extremes. Climate Dynamics, 33(7-8), 917-935, DOI: 10.1007/s00382-008-0473-8

Fischer, E. M. and Schär C., (2010), Consistent geographical patterns of changes in high-impact European heatwaves, Nature Geoscience 3, 398 – 403, doi:10.1038/ngeo866

Frich, P., L.V. Alexander, P.M. 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 Research, 19(3), 193-212

Hansen J., Sato M., Ruedy R (2012), Perception of climate change. Proceedings of the National Academy of Science USA 109(37):E2415–E2423

Haylock, M. R., Hofstra, N. Klein Tank A. M. G., Klok, E. J., Jones P. D. and New M. (2008), A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, Journal of Geophysical Research, vol. 13, D20119, doi:10.1029/2008JD010201

Hoang T.T.H., S. Parey and D. Dacunha-Castelle (2009), Multidimensional trends: The example of temperature. The European Physical Journal Special Topics 174, 113-124, DOI: 10.1140/epjst/e2009-01094-6

Hoang T.T.H. (2010), Modélisation de séries chronologiques non stationnaires, non linéaires: application à la définition des tendances sur la moyenne, la variabilité et les extrêmes de la température de l'air en Europe. PhD thesis work (written in English), http://www.tel.archives-ouvertes.fr/tel-00531549/fr/

IPCC (2007), Summary for Policymakers. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [S. Solomon, D. Qin, M. Manning, Z. chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller (eds.)].

Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, pp. 1-18.

Katz R and B. Brown (1992), Extreme events in a changing climate: variability is more important than averages. Climatic Change, 21:289–302, DOI: 10.1007/BF00139728

Klein Tank AMG et al (2002), Daily dataset of 20th-century surface air temperature and precipitation series for the European ClimateAssessment. International Journal of Climatolology 22:1441–1453. Data and metadata available at http://eca.knmi.nl

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. Journal of Climate, 16(22), 3560-3571

Meehl, G. A. and C. Tebaldi (2004), More intense, more frequent, and longer lasting heat waves in the 21st century. Science 305, 994-997 (2004), DOI: 10.1126/science.1098704

Menne M.J., Durre I., Vose R. S., (2011), An overview of the Global Historical Climatology Network – Daily Database. Journal of Applied Meteorology and Climatology

Parey S., F. Malek, C. Laurent and D. Dacunha-Castelle (2007), Trends and climate evolutions: statistical approach for very high temperatures in France. Climatic change, 81, 331-352, DOI:10.1007/s10584-006-9116-4

Parey S., D. Dacunha-Castelle and T.T.H. Hoang (2010a), Mean and variance evolutions of the hot and cold temperatures in Europe. Climate Dynamics 34:345–369. doi:10.1007/s00382-009-0557

Parey S., T.T.H. Hoang and D. Dacunha-Castelle (2010b), Different ways to compute temperature return levels in the climate change context. Environmentrics 21(7–8):698–718. doi:10.1002/env

Rhines A. and Huybers P. (2013), Frequent summer temperature extremes reflect

changes in the mean, not the variance. Proceedings of the National Academy of Science USA 110(7) E546

Simolo C, Brunetti M, Maugeri M., Nanni T (2011) Evolution of extreme temperatures in a warming climate. Geophysical Research Letters 38(16):L16701.

Stone C. J. (1977), Consistent nonparametric regression. Annals of Statistics, 5: 595-620

Westra S. and S. A. Sisson (2011), Detection of non-stationarity in precipitation extremes using a max-stable process model, Journal of Hydrology, 406, 119-128.

Zhang X.F, F.W. Zwiers and G. Li (2004), Monte Carlo experiments on the detection of trends in extreme values. Journal of Climate, 17: 1945-1952

Zwiers F.W., X.F. Zhang and Y. Feng (2011), Anthropogenic influence on long return period daily temperature extremes at regional scales. Journal of Climate, 24(3), 881-892, DOI: 10.1175/2010JCLI3908.1

List of figures

~~**Figure 1**: Annual mean temperature evolution between 1950 and 2009 for TX in Eureka (top panel) and TN in Ajo (bottom panel)~~

Figure ~~2~~ 1: Results of the stationarity test of the GEV parameters (location μ and scale σ) of the residuals Y(t) for **a)** minimum TN in winter, **b)** maximum TN in Summer, **c)** minimum TX in winter and **d)** maximum TX in Summer in Eurasia (left panels) and in the United States (right panels). Non convergence means that the cross-validation could not converge to an optimal smoothing parameter and thus the non-parametric evolution of the GEV parameters could not be computed. Green means that stationarity is valid for μ and σ, blue for μ only, orange for σ only and red means that the hypothesis is rejected for both μ and σ

~~**Figure 3**: Non-parametric location μ (top panel) and scale σ (bottom panel) parameters for the extremes of X(t) directly obtained from the maxima of X(t) (black solid line) with their 95% bootstrap confidence interval (black dotted lines) and reconstructed from the stationary parameters of Y(t) and the non-parametric evolutions of the mean and standard deviation of X(t) (red line).~~

Figure 4 2: Comparison of the 50-year Return Levels for maximum TX in summer computed by extrapolation of the significant linear trends in the location μ and scale σ parameters of the fitted GEV ($RL_{\mu,\sigma}$) or by extrapolation of the significant linear trends in mean m and standard deviation s ($RL_{m,s}$). Red dots indicate that $RL_{m,s}$ falls outside the 95% confidence interval of $RL_{\mu,\sigma}$ and is higher (closed dots) or lower (open dots) and green points indicate that the 95% confidence intervals overlap.

Figure ~~5~~ 3: Non-parametric (green curve) and optimal parametric (red curve) trends in the location μ and scale σ parameters of the GEV distribution fitted on TX summer block maxima for the stations of Berlin and Dresden.

Figure ~~6~~ 4: Non-parametric (black curve) and linear (blue curve) trends in mean m and standard deviation s for TX in summer for the stations of Berlin and Dresden. The significance of the linear trend is indicated in the top left corner of each curve and is assessed by a Mann-Kendall test with a 10% significance level.

Figure ~~7~~ 5: Results of the stationarity test of the GEV parameters (location μ and scale σ) of the residuals Y(t) for minimum TN in winter for **a)** IPSL-CM5-LR and **b)** CNRM-CM5 model and maximum TX in Summer for **c)** IPSL-CM5-LR and **d)** CNRM-CM5 model in Eurasia (left panels) and in the United States (right panels) in the period 1950-2005. Non convergence means that the cross-validation could not converge to an optimal smoothing parameter and thus the non-parametric evolution of the GEV parameters could not be computed. Green means that stationarity is valid for μ and σ, blue for μ only, orange for σ only and red means that the hypothesis is rejected for both μ and σ.

Figure ~~8~~ 6: same as figure ~~7~~ **5** but for period 2006-2100 (2006-2060 for CNRM-CM5) with RCP8.5