9.6 STATISTICAL LIMITATIONS FOR DIAGNOSING CHANGES IN EXTREMES FROM CLIMATE MODEL SIMULATIONS

Christoph Frei*
Atmospheric and Climate Science ETH, Zürich, Switzerland

1. INTRODUCTION

Interest into the possibility that global climate change could be associated with changes in extreme weather has stimulated diagnostic analyses of climate model experiments for extreme events (e.g. Meehl et al. 2000; McGuffie et al. 2000). The purpose of such analyses is to quantify the simulated frequency/intensity of droughts, heavy precipitation, windstorms etc. and to compare the results from a sensitivity experiment (e.g. a model integration for future climate) with those from a reference experiment (e.g. an integration for current climate).

Classical analyses of extremes in climate models have focussed on moderately rare events with a typical return period of 100 days or shorter (1% quantile of daily values; e.g. Hennessy et al. 1997, Durman et al. 2001). Moderate event thresholds warrant for sufficiently large samples and hence for statistical robustness of the results. Although this approach provides insights into changes of the (model simulated) frequency distribution, it is difficult to extrapolate the results towards damage-relevant extremes with a return period much larger than those considered. Recently, several model diagnoses of extremes have been conducted, adopting methods of extreme value statistics, which permit, in principle, to estimate changes for events with a much larger return period (20 years and larger; see e.g. Kharin and Zwiers 2000, Jones and Reid 2001). Although the results are of more immediate relevance for impacts and damages, the uncertainty due to limited sample size can be substantial.

Diagnosis for a change between two model simulations requires an assessment of statistical significance. This is commonly accomplished by means of a statistical test, comparing the magnitude of the change (the signal) against its statistical uncertainty (the noise). The reliability with which a given change between two model integrations can be detected as statistically significant will then depend on the signal-to-noise ratio. Large statistical uncertainty, such as that expected from diagnostics on very rare extremes, limits the detectability of a change. Quantitative knowledge of these limits is an important prerequisite for designing the diagnostic procedures and for interpreting results adequately.

In this study the statistical limits for detecting a change in extremes from two climate model integrations is theoretically quantified. The probability of detection is defined as the chance for identifying, from one pair of model simulations, a prescribed change. This quantity is equivalent to the power of the statistical test and it depends on the magnitude of the change, the length of the model integrations and the rarity of events under consideration. Here we derive quantitative estimates of the detection probability for the classical method of extreme value statistics, recently applied as climate model diagnostic.

2. EXTREME VALUE STATISTICS FOR TWO SAMPLES

Extreme value statistics is an asymptotic theory on extreme values from a large sample, very much like the 'law of large numbers' for the sample mean. Founded by Fisher and Tippett (1928) and developed for application by Gumbel (1958), the theory distinguishes specific parametric families of distribution functions, which describe sample extremes in the limit of large sample size. The theory of extreme value statistics and its applications in climatology are reviewed e.g. in Palutikov et al. (1999) and Coles (2001).

In this study we consider the classical block maximum method which deals with extreme values

* *Corresponding author address:* Christoph Frei, Institute for Atmospheric and Climate Science, ETH, Winterthurerstr. 190, CH-8057 Zürich, Switzerland;

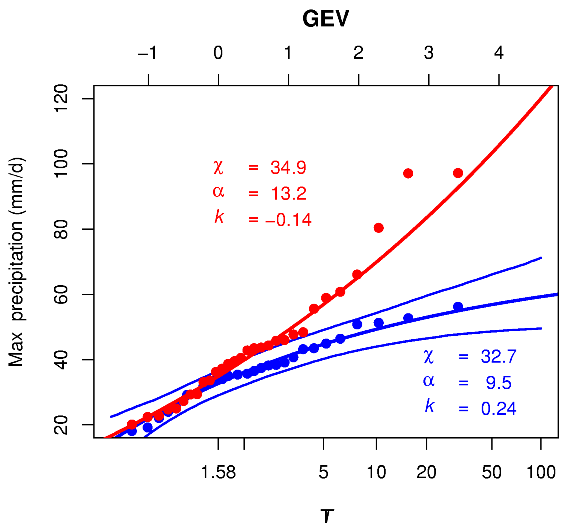

Figure 1: Example Gumbel Diagram showing two samples of maximum summer precipitation observed in Zurich Switzerland, (blue: 1931-1950, red: 1971-2000) together with the corresponding fits of a generalized extreme value distribution. Values of location (χ), scale (α) and shape (*k*) parameters are given in the inset. Blue lines flanking the fit are 95% confidence lines of the fit.

(maxima or minima) taken from a large sample, typically the seasonal or annual extremes of daily values. In this case the limiting distribution is the Generalized Extreme Value distribution (GEV). Fitting a GEV to the sample of extremes provides estimates of extreme values as a function of return period. The GEV is a three-parameter distribution family with location, scale and shape parameter. It comprises the Gumbel distribution as a special case. Results of a GEV analysis are commonly displayed in a Gumbel diagram, showing the cumulative distribution with a transformed return period axis, such that the Gumbel distribution appears as a straight line. An example Gumbel diagram is displayed in Fig. 1 showing observed maxima in daily summer rainfall together with the corresponding GEV fit. Several methods are available for the estimation of distribution parameters. In this study, estimation is based on the L-moments method, which in real applications was demonstrated to be particularly robust (Hosking 1990, 1992).

Application of extreme value statistics for diagnosing a change from a control integration to a scenario integration of a climate model involves estimation of distributions for each of the two integrations. The situation is illustrated in Fig. 1 with samples and distribution functions evaluated for distinct 30-year periods of the observational record (red and blue samples). In this example the fitted distributions imply an increase of extreme values from the first (blue) to the second (red) sample for return periods larger than about 5 years. There are several possibilities of a statistical test for assessing the significance of this change. In this study we adopt the Null-Hypothesis that the extreme value at return period T for the scenario $X_S(T)$ is equal to that of the control sample $X_C(T)$. This hypothesis is accepted when $X_S(T)$ is comprised in the 95% confidence interval for $X_C(T)$ and rejected otherwise. (Note that this testing is asymmetric between the control and scenario samples. A symmetric approach would consider overlap of 78% confidence intervals from both samples (e.g. Kharin and Zwiers 2000). The asymmetric test was chosen here because of easier interpretation of results later.) Indeed, for the example of Fig. 1, estimates for the second sample (red) are outside the confidence interval for the first sample (blue) and the difference is statistically significant even for large return periods. (Note that this is an extreme example chosen to illustrate the method.)

3. DETECTION PROBABILITY

The detection probability is defined as the chance with which a given change of an extreme value $X(T)$ (e.g. the 20-year return period rainfall) can be identified as statistically significant. A schematic illustration of the detection probability is depicted in Fig. 2. Let X_c and X_s denote the true extreme values and the bell-shaped curves depict the uncertainty of estimates for these

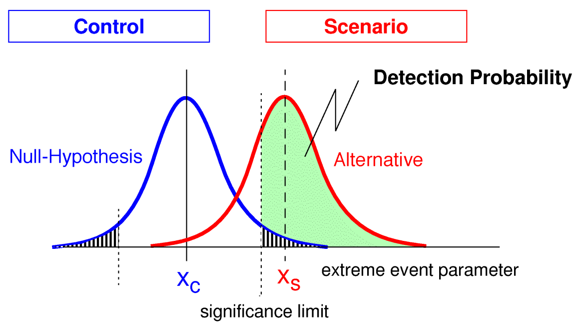

Figure 2: Schematic of definition of detection probability. X_c and X_s denote true (unknown) extreme values under control and scenario conditions. Bell-shapes depict uncertainty of estimations of X_c and X_s. Significance limit denotes the minimum distance between one pair of estimates for X_c and X_s such that Null-Hypothesis (X_c=X_s) is rejected.

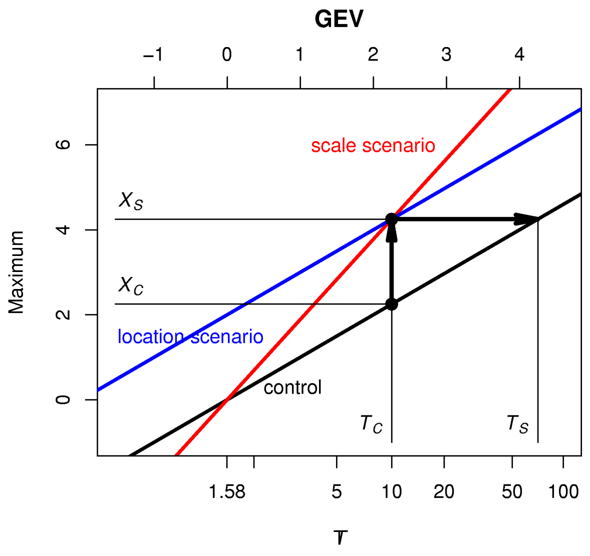

Figure 3: Schematic Gumbel diagram showing distribution functions for control and for scenario conditions that have been used in the calculation of detection probability. Both, the the scale scenario (red, change in scale parameter) and the location scenario (blue, change in location parameter) correspond to the same absolute change (X_S–X_C) at return period T_C.

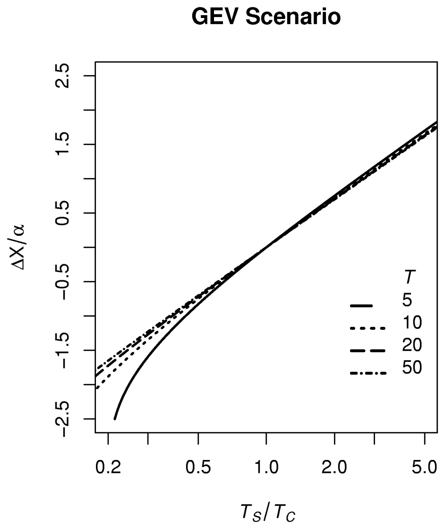

Figure 4: Relationship between relative change in return period T_S/T_C (x-axis, log-transformed) and absolute change in extreme value $\Delta X=X_S–X_C$ (y-axis) for scenarios in GEV distributions. See also Fig. 3 for definition of symbols.

values. One pair of estimates is taken to be statistically significant if their difference exceeds the significance limit. The chance of an arbitrary pair of estimates to be statistically significant is equal to the fraction of the green area under the red bell shape. The detection probability is equal to the power of the statistical test. (A more detailed description of the concept of detection probability is given in Frei and Schär (2001)).

It is evident from Fig. 2 that the detection probability increases with the magnitude of the prescribed change X_S–X_C, (determining the separation of the two bell shapes), and with the length of the simulations and hence the available sample size (determining the width of the bell shapes). On the other hand, the detection probability for a similar change in extreme values decreases with increasing return period T, as a consequence of the widening of confidence bounds towards more rare extremes (see e.g. Fig. 1).

In this study the detection probability is calculated for specified changes in extreme values. The specification requires the definition of distribution functions (GEVs) under control and under scenario conditions. For simplicity the predefined distribution functions will be taken from the family of Gumbel distributions. (Note that the subsequent estimation does not restrict to the Gumbel distribution. This is relevant in estimating the detection probability for the full distributional flexibility of extreme value statistics.) The procedure for defining distribution functions for control and scenario conditions is displayed schematically in Fig. 3. Without loss of generality it can be assumed that the control distribution function has a position parameter of 0.0 and a scale parameter of 1.0. (These parameters define origin and units.) A predefined change of the extreme value (X_S–X_C) at return period T_C can be accomplished by a change in the location parameter or the scale parameter or a combination of both. The simple cases correspond respectively to a vertical shift or a rotation of the distribution functions in the Gumbel diagram. The two basic examples will be termed *location* and *scale scenarios* respectively.

The detection probability will depend on the change of extreme values $\Delta X=X_S–X_C$. However it will be more illustrative to express the imposed change in terms of changes in the return period. Here we define the change in return period as the ratio T_S/T_C where T_C is the return period under consideration (i.e. under CTRL) and T_S is the return period which the changed extreme value X_S at T_C would have under control conditions. (See also Fig. 3). The relationship between ΔX and T_S/T_C is approximately exponential (Fig. 4) although

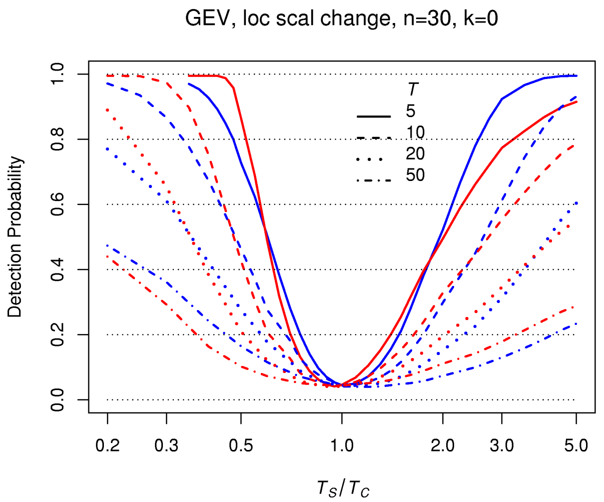

Figure 5: Detection probability as a function of the magnitude of the relative change in return period (T_S/T_C). Probabilities for an increase (decrease) in extreme value are to the rigth (left) of T_S/T_C =1.0. Blue lines are for the *location*, red lines for the *scale scenario* respectively (see also Fig. 3). Results are given for the detection at various return periods *T* in years (see line styles in legend).

there are slight deviations from this, especially for negative ΔX and smaller return periods *T*.

The calculation of the detection probability involves the following sequential steps:

(a) Specification of distribution functions for control and for scenario, consistent with the predefined change in extreme value ΔX.

(b) Simulation of parametric bootstrap samples of predefined size *n* from the control and scenario distributions.

(c) Fitting of GEV distributions to each sample and estimation of extreme values X_C and X_S.

(d) Calculation of detection probability by counting pairs of estimates exceeding the significance limit.

4. RESULTS

Results of the theoretical estimation of the detection probability are displayed in Fig. 5 as a function of the relative change in return periods (T_S/T_C). These estimates are valid for a sample size of *n*=30 both in the control and scenario samples. The situation is representative, for instance, for a diagnostic of annual temperature extremes in two climate model time-slices of 30 years length each. Results are shown for different rarities (return periods *T* ,line types) and for the location and scale scenarios respectively (blue and red). A significance level of 5% has been used throughout.

The limitations for the statistical detection of a change in extremes are evident: For events with a return period of 5 years, a change by a factor of 2 in frequency is detected with a probability of about 50%. However for very rare events with a return period of 50 years, the detection probability drops to less than 20% for the same relative change T_S/T_C. In fact, even a change by a factor of 5 (increasing or decreasing) would only be identified as statistically significant with a probability of less than 50%.

The results for the *location* scenario are roughly symmetric between increases and decreases for the same relative change T_S/T_C. Strictly, similar values would have been obtained for similar positive and negative changes of extreme values ΔX. The slight asymmetry for the location scenario in Fig. 5 is an effect of the curvature in the relationship between absolute and relative changes (see Fig. 4). Changing from the *location* to the *scale scenario* the asymmetry between increases and decreases becomes more pronounced. For a similar increase in T_S/T_C the *scale scenario* has a lower (higher) detection probability than for the *location scenario*, if the probability is above (below) 50%. The situation is reversed for decreases. This behavior can be understood from the variation of confidence bands between the two scenarios. Whilst for a *location scenario* the confidence interval is similar to that of the control, it is larger (smaller) than for control with an increase (decrease) of the scale parameter in a *scale scenario*. Fig. 2 can be used to make clear, that the detection probability of a *scale scenario* should change from that of a *location scenario* as depicted in Fig. 5.

Despite the fact that there is some sensitivity of the detection probability upon the type of scenario, the quantitative results are actually not overtly different (not more than 20%). This suggests that the detection probabilities depicted in Fig. 5 can be considered as rough estimates for a change in extreme values, essentially independent from the details of how this change is related to changes in scale and location parameters (or both) of the distribution function.

With the recent advance of ensemble modeling also for high-resolution global and regional climate modeling, there is some scope for sample sizes larger than the standard of *n*=30 depicted in Fig. 5. Fig. 6 displays the sensitivity of the detection probability upon sample size

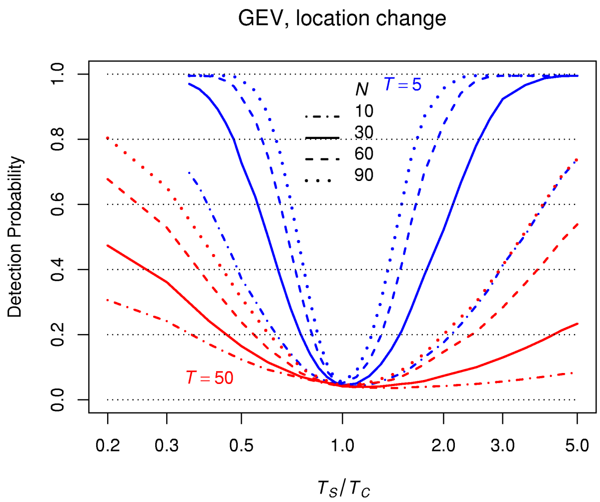

Figure 6: Comparison of detection probability for different settings of sample size (n=10,30,60,90; see legend for line styles). Results are shown for the location scenario at return periods of *T*=5 years (blue) and *T*=50 years (red).

(*n*=10, 30, 60, 90). For simplicity only results for the location scenario are displayed. For the return period of 5 years (blue lines) there is indeed a remarkable increase of detection probability with larger sample size. For example, with *n*=90 (e.g. 3 ensembles of a 30 year time-slice) the detection probability for a relative change in return period (T_S/T_C) by a factor of 2.0 or 0.5 raises to above 90%. In contrast, the probability with a sample of only 10 years is significantly reduced. Again, the detection probability for a change in the 50-year extreme value (red lines) is sensitive to sample size. However the improvement with *n*=90 still leaves considerable limitations. For example changes by a factor of 3 or lower can still be detected with only less than 50% probability.

5. CONCLUSION

A theoretical estimation has been undertaken of the probability with which changes in extreme events can be statistically detected in a diagnostic analysis of climate model simulations. The estimation was carried out for the standard procedure of extreme value statistics adopted to a control and a scenario sample and using generalized extreme value distributions. The results pinpoint to the potential limitations of such a diagnostic especially for very rare events. For example, using annual extremes from two time-slice simulations of 30 years length each, a change of the 5-year extreme value to the 10-year (or to the 2.5 year) extreme value can be detected with a probability of about 50%. However this value drops to below 20% for change of the 50 year event to the 100 year (or to the 25) year event. For these rare events even changes in the order of a factor of 5 are detected with a probability lower than 50%.

These results do not question the theoretical fundaments and principal suitability of extreme value statistics for diagnosing extremes. However, they pinpoint to the limited evidence on changes that can be drawn from its applications with a small sample size such as that commonly available from global and regional climate models.

The results have a range of implications for the interpretation and design of model diagnosis on changes in extremes: Firstly, the absence of evidence for a change (i.e. a change that turns out to be statistically non-significant) should not be interpreted as the absence of a change. Even quite substantial changes can fail to be detected purely on statistical grounds. Secondly, diagnostics dealing with moderately rare, rather than extreme events, might be an important complement to classical extreme value statistics. Although the results may not simply be extrapolated to the very far tail of the distribution function, such diagnostics may be helpful in identifying statistically sound changes in the frequency distributions which themselves can guide further model diagnostics aiming at a physical understanding of possible changes of extremes. Finally, every possible measure should be taken to optimize the sample size for a model diagnostic on extremes. The use of ensemble model simulations and the consideration of long simulations are particularly important.

ACKNOWLEDGMENT

This research was supported by the 5th Framework Program of the European Union (project STARDEX, contract No. EVK2-CT-2001-00115) and by the Swiss Ministry for Education and Science (BBW).

REFERENCES

Coles, S., 2001: *An introduction to statistical modeling of extreme values.* Spinger, London.

Durman, C.F., J.M. Gregory, D.C. Hassell, R.G. Jones and J.M. Murphy, 2001: A comparison of extreme European daily precipitation simulated by a global and a regional climate model for present and future climates. *Q. J. Roy. Meteorol. Soc.,* **127,** 1005-1015.

Frei, C. and C. Schär, 2001: Detection probability of trends in rare events: Theory and application to heavy precipitation in the Alpine region. *J. Climate,* **14,** 1564-1584.

Gumbel, E.J., 1958: *Statistics of extremes.* Columbia University Press, New York.

Hennessy, K.J., J.M. Gregory and J.F.B. Mitchell, 1997: Changes in daily precipitation under enhanced greenhouse conditions. *Climate Dyn.,* **13,** 667-680.

Hosking, J.R.M., 1990: L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. R. Statist. Soc. B,* **52,** 105-124.

Hosking, J.R.M., 1992: Moments or L-moments? An example comparing two measures of distributional shape. *The American Statistician,* **46,** 186-189.

Jones, P.D. and P.A. Reid, 2001: Assessing future changes in extreme precipitation over Britain using regional climate model integrations. *Int. J. Climatol.,* **21,** 1337-1356.

Kharin, V.V. and F.W. Zwiers, 2000: Changes in the extremes in an ensemble of transient climate simulations with a coupled Atmophere-Ocean GCM. *J. Climate,* **13,** 3760-3788.

McGuffie, K., A. Henderson-Sellers, N. Holbrook, Z. Kothavala, O. Balachova and J. Hoekstra, 1999: Assessing simulation sof daily temperature and precipitation variability with global climate models for present and enhanced greenhouse climates. *Int. J. Climatol.,* **19,** 1-26.

Meehl, G.A., F. Zwiers, J. Evans, T. Knutson, L. Mearns and P. Whetton, 2000: Trends in extreme weather and climate events: Issues related to modeling extremes in projections of future climate change.*Bull. Amer. Meteorol. Soc.,* **81,** 427-436.

Palutikof, J.P., B.B. Brabson, D.H. Lister and S.T. Adcock, 1999: A review of methods to calculate extreme wind speeds. *Meteorol. Apps.,* **6,** 119-132.