ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-899

Mining Extremes: Severe Rainfall and Climate Change

Debasish Das[1,2,*] and **Evan Kodra**[2] and **Zoran Obradovic**[1] and **Auroop R. Ganguly**[2]

Abstract. Theoretical developments for the analysis and modeling of extreme value data have tended to focus on limiting cases and assumptions of independence. However, massive datasets from models and sensors, space-time dimensionality, complex dependence structures, long-memory, long-range and low frequency processes all motivate the need for sophisticated methods for correlated and finite data that follow complex processes. The importance of extremes has been rapidly growing in areas ranging from climate change and critical infrastructures to insurance and financial markets. Here we briefly discuss the state-of-the-art and key gaps, through the case of rainfall extremes under climate change. Preliminary analysis suggests new directions and points to research areas that deserve further attention.

1 INTRODUCTION

Extreme events are growing in importance across disciplines like finance, insurance, hydrology [1] and climate [2-3]. Rare events mining in artificial intelligence (AI), which includes classification of imbalanced datasets through synthetic over-sampling [4], are typically not concerned with extremely high or low values. In the latter case, Gaussian assumptions do not hold, the extremes may not even be present in the data, and the generation processes may be continuous. Extreme value theory (EVT) is among the few statistical methods doing true extrapolation; parametric relations are developed to infer about tails of the distribution (e.g., a 100-year, or a one in a thousand, event) with values that are adequately large but not necessarily at the extreme tails [5]. The selection of adequately large values may be based either on the block maxima over a time window (e.g., annual) or as a peak over threshold, which in turn may be fixed or variable (e.g., a percentile).

Despite decades of development, EVT remains an area with open challenges, many of which may be resolved through statistics, data mining and AI. The growing importance of extremes, for example in the context of climate change and severe rainfall, motivates urgent solutions. The open challenges [6] include the selection and justification of EVT approaches, exploring parameter uncertainties, modeling space-time dependence as well as the use of covariates to reduce uncertainty, relating to space-time outliers or change, and blending multiple information sources. Climate change is selected as an exemplar both because of the societal importance [7] and to validate the methods with massive data from sensors and models.

2 PROBLEM DESCRIPTION

Rainfall extremes are typically characterized by their intensity, duration and frequency (IDF) for applications from water resources management, flood hazards, and dam design [8]. Recent research has explored changes in the IDF curves under climate change [9].

The *n-year return level,* (RL_n), defined as the level that is reached or exceeded once every *n-years* on the average (alternatively, the probability of exceedance on any given year is *1/n*). The three [5, 8] ways to describe extreme values are the Generalized Extreme Value (GEV) distribution fitted to block maxima (BM) or blocks of time windows like an annual maxima time series, the Poisson arrival of extremes followed by the Generalized Pareto distribution (GPD) fitted to the excesses above a threshold, leading to the Peak-over-Threshold (PoT) as well as the Point Process (PP) approach. From a pragmatic standpoint, the approaches generate estimates of the return levels along with associated uncertainties per time series, but require either the selection of a block size or a threshold. The distributions (GEV or GPD) arise from limiting cases for large sample sizes as well as when the maxima or excess data are independent and identically distributed. Thus, the tradeoffs during the choice of a block size or a threshold may be expressed as a bias versus variance issue: larger block sizes or higher thresholds may imply lower bias but larger variance while smaller block sizes and lower thresholds may imply larger variance. For most practical applications in climate and rainfall, the typical choice of the annual maxima for BM-GEV minimizes correlation but wastes data, while the use of PoT-GPD typically results in correlated excesses but can use more data. Thus, research in rainfall extremes has typically used the GEV for annual maxima (e.g., [9-10]) as well as the GPD for excesses above user-selected percentile-based thresholds after temporal aggregation (e.g., [11] used weekly extremes). One data mining challenge is whether the applicability of EVT may be automated to an extent where they can scale to massive data, for example, simulated data from the current generation of global climate models, which in turn is rapidly approaching the petabyte scale. However, this scalability needs to be achieved without compromising accuracy or precision. Our preliminary results explore the tradeoffs between data size and correlation for BM-GEV and PoT-GPD respectively as well as the computational issues in parameter and uncertainty estimation.

3 PRELIMINARY RESULTS

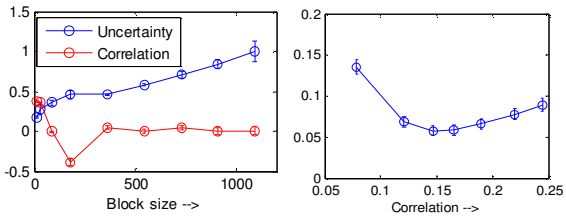
First, we evaluate the effects of sample size and temporal correlation - present among the samples of an observed time-series - on the precision of the estimated return levels with the GEV and the GPD. Let us designate RL_T as the true (n-year) return level and $\widehat{RL}_{BM}$ and $\widehat{RL}_{PoT}$ as estimated return levels from BM and PoT approaches, respectively. Let us assume for simplicity, without loss of generality, that these are unbiased Gaussian estimators:

$$\widehat{RL}_{BM} \sim \mathcal{N}(RL_T, \lambda_{BM}^{-1}), \quad (1a)$$

$$\widehat{RL}_{PoT} \sim \mathcal{N}(RL_T, \lambda_{PoT}^{-1}). \quad (1b)$$

[1]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia; [2]Department of Civil and Environmental Engg., Northeastern University, Boston, MA, USA; * Corresponding author: d.das@neu.edu

We used daily precipitation time-series observed over 200 different locations across India [10] between 1951-2003 to explore the comparative precision (inverse variance) of our estimators, λ_{BM} and λ_{PoT}, as they vary functionally with sample size L, and temporal correlation, ρ, among chosen samples respectively (L is primarily expected to affect λ_{BM} and ρ is expected to influence λ_{BM}). For BM, we varied L by changing the block-size and computedλ_{BM} and sample correlation, which is plotted in Figure 1a; for PoT, we varied ρ by varying the threshold from 80 to 99 percentile (sample correlation decreases with increasing threshold) and computedλ_{PoT}, which is plotted in Figure 1b. In both cases, average over 200 locations is plotted. For BM, uncertainty is less for smaller block size, but correlation fluctuates. This suggests the need for balancing the dual concerns. Further tests are needed to determine if the uncertainty versus correlation plot shown (Fig. 1) for the PoT may generalize.



(a) (b)

Figure 1: (a) BM with GEV- plot of parameter uncertainty and correlation vs block size (days), and (b) PoT with GPD – parameter uncertainty vs correlation.

Second, we show the increase in computation time for MLE-based parameter estimation of the PoT-GPD as a function of the number of time series. Figure 2 shows a linear dependence and therefore leaves scope for improvement. The time for parameter and uncertainty estimation, including the use of the bootstrap [10], typically relies on the MLE hence this is critical to address.

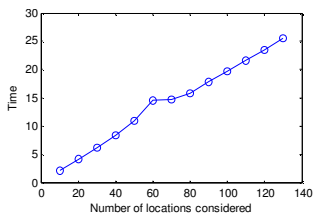


Figure 2: Computation time (sec) for parameter estimation vs. number of locations considered

4 FUTURE WORK

Applications to massive data as well as precise and accurate predictive insights on extremes, for example in the context of heavy rainfall events under climate change, require automated declustering to reduce temporal correlations in extremes [12], downscaling of extremes [13], as well as quantifying tail dependence [14]. Model parameter estimation, whether via maximum likelihood (ML), L-moments estimation, or the bootstrap for either of the two, may impact accuracy [15] and computation.

A key concern in future research is to relate to the statistical insights from the data and the physical or process understanding of the domain (hydro-climate in our case) to each other. In addition, a relation needs to be drawn to the expected sources of uncertainty [16] for understanding the accuracy as well as for enhanced predictions. The complexity grows when multisource and multi-resolution data [17], some of which are sparse, need to be fused. Covariates such as temperature or humidity may hold information content for enhancing predictions of rainfall extremes [18] at multiple space-time scales. The data-mining community is well positioned to make a difference in the theory and algorithms of extremes as well as their applications to climate extremes and generalizations to multiple domains.

REFERENCES

[1] Reiss, R-D., Thomas M: Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields, 3rd edition, 2007, Springer, 511 pp.

[2] Min, S.-K., Zhang, X., Zwiers, F.W., Hegerl, G.C., Human contribution to more-intense precipitation extremes. Nature, 470, 2011, 378-381.

[3] Lozano, A.C., Li, H., Niculescu-Mizil, A., Liu, Y., Perlich, C., Hosking, J., Abe, N., Spatio-temporal causal modeling for climate change attribution, Proc. 15th ACM SIGKDD, KDD 2009, 587-596.

[4] Chawla, N.V., Boyer, K.W., Hall, L.O., Kegelmeyer, W.P., SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, 16, 2002, 321-357.

[5] Coles, S. G., An introduction to statistical modeling of extreme values, 2001, Springer-Verlag, 208 pp.

[6] Fuentes, M., Reich, B., and Lee, G., Spatial-temporal mesoscale modelling of rainfall intensity using gage and radar data, Annals of Applied Statistics, 2, 2012, 1148–1169.

[7] Field, C.B., Et Al., IPCC, Summary for Policymakers. In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Special Report of the Intergovernmental Panel on Climate Change, 2012, Cambridge University Press, pp. 1-19.

[8] Katz, R. W., Parlange, M. B., Naveau, P: Statistics of extremes in hydrology. Advances in Water Resources, 25, 1287–1304.

[9] Kao, S. - C., Ganguly A. R: Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios, Journal of Geophysical Research, 116(D16), 2011, 14 pp.

[10] Ghosh, S., Das, D., Kao, S.-C., Ganguly A.R., Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes, Nature Climate Change 2, 2012, 86–91.

[11] Khan, S., Kuhn, G., Ganguly, A. R., Erickson III, D. J., and Ostrouchov, G: Spatio-temporal variability of daily and weekly precipitation extremes in South America, Water Resources Research, vol. 43, W11424, 2007, 25 pp.

[12] Ferro, C.A.T., and Segers, J., Inference for clusters of extreme values, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2), 2003, 545-556.

[13] Mannshardt-Shamseldin, E.C., Smith, R.L., Sain, S.R., Mearns, L.O., and Cooley, D., Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data, Annals of Applied Statistics, 4(1), 2010, 484-502.

[14] Kuhn, G., Khan, S., Ganguly, A.R., and Branstetter, M.L., Geospatial-temporal dependence among weekly precipitation data with applications to observations and climate model simulations in S. America, Advances in Water Resources, 30(12), 2007, 2401-2423.

[15] Martins E.S., Stedinger J.R., Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, Water Resources Research, 36, 2000, 737-744.

[16] Wehner, M., Sources of uncertainty in the extreme value statistics of climate data. Extremes, 13(2), 2010, 205-217.

[17] Smith R.L., Tebaldi, C., Nychka D., Mearns L.O., Bayesian modeling of uncertainty in ensembles of climate models, Journal of the American Statistical Association, 104 , 2009, 97-116.

[18] O'Gorman, P. A., and Schneider, T: The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. Proc. Natl. Acad. Sci. USA, 106(35), 14773-14777, 2009.