5th International Workshop on Climate Informatics
September 24-25, 2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

CLUSTERING OF STATION OBSERVATIONS FOR EXTREME VALUE ANALYSIS

Andre R. Erler[1], W. Richard Peltier[1]

***Abstract*—The analysis of precipitation extremes with decadal return periods requires long data records; longer than is typically available from station observations. To address this problem, a method is proposed for clustering station records based on the similarity of their precipitation climatologies. It is demonstrated that the proposed aggregation method is superior to naive aggregation methods for Extreme Value Analysis and enables the detection of changes in the historical record that would otherwise not be distinguishable from noise.**

I. INTRODUCTION

The analysis of extreme events has attracted considerable attention in the climate change impacts community in recent years [1]. A statistical framework for the analysis is provided by the theory of Extreme Value Analysis (EVA), which typically involves fitting a particular distribution to a sample of extreme events. In order to constrain the parameters of the distribution, a large sample of events is needed; in particular, to constrain the extreme ends of the distribution. For example, to estimate the magnitude of a 100 year return period event, a record that is significantly longer than 100 years would be desirable. However, only a few meteorological stations have records of even 100 years. At the same time, it is often these rare events that are of greatest interest to stakeholders.

To overcome this problem, we propose to aggregate observations from similar stations, in order to increase the effective size of the record. A key assumption of EVA is that events are identically distributed, so that observations cannot be aggregated arbitrarily. Here we demonstrate that a good fit to the aggregated data can be achieved by clustering stations based on the similarity of their climatological seasonal cycle. In addition, this method can also be used to detect diverging trends in different regions that would otherwise be missed.

The data used for this proof-of-concept is precipitation data from Environment Canada meteorological stations for the province of British Columbia (BC) south of 55°N [2]. BC is characterized by complex topography and a strong precipitation gradient, ranging from very high amounts at the Coast Mountains to relatively dry conditions in the Interior Plateau. We compare two 40 year observation periods: 1920 - 1960 and 1970 - 2010.

II. STATION CLUSTERING

The `k-means` clustering algorithm was employed to group stations by the similarity of their climate regime. Each station was represented by a vector derived from the climatological seasonal cycle of monthly mean liquid and solid precipitation measured at the station (i.e. 24 values per station). To reduce the dimensionality of the station vectors, principal component analysis was applied prior to clustering, but this step is not essential. A total of nine clusters was chosen for this analysis; this number was determined by experimentation and is somewhat arbitrary. The division between stations at the coast and stations land inwards, which differ significantly in their winter precipitation amounts, emerges regardless of the number of clusters. However, with a larger number, the major clusters become more homogeneous, because “outliers” are moved into separate clusters. Note that precipitation extremes do not enter the clustering algorithm.

III. EXTREME VALUE ANALYSIS

The Extreme Value Analysis employed here is based on the Generalized Extreme Value (GEV) distribution. The GEV describes the distribution of the maxima of large blocks of independent and identically distributed samples. For this analysis the block size is one winter season (Dec. - Feb.). GEV distributions fitted to the precipitation maxima for each period are shown in Fig. 1, along with histograms of the underlying data. The three panels show different station groups: the entire province of BC south of 55°N (left), a cluster at the Pacific coast (middle), and a cluster in the Interior Plateau (right). The latter two were defined using the proposed clustering technique. Contrary to convention [3], only a single GEV distribution is fitted to the aggregated data from all

Corresponding author: A. R. Erler, University of Toronto, Toronto ON, Canada; aerler@atmosp.physics.utoronto.ca [1]Department of Physics, University of Toronto

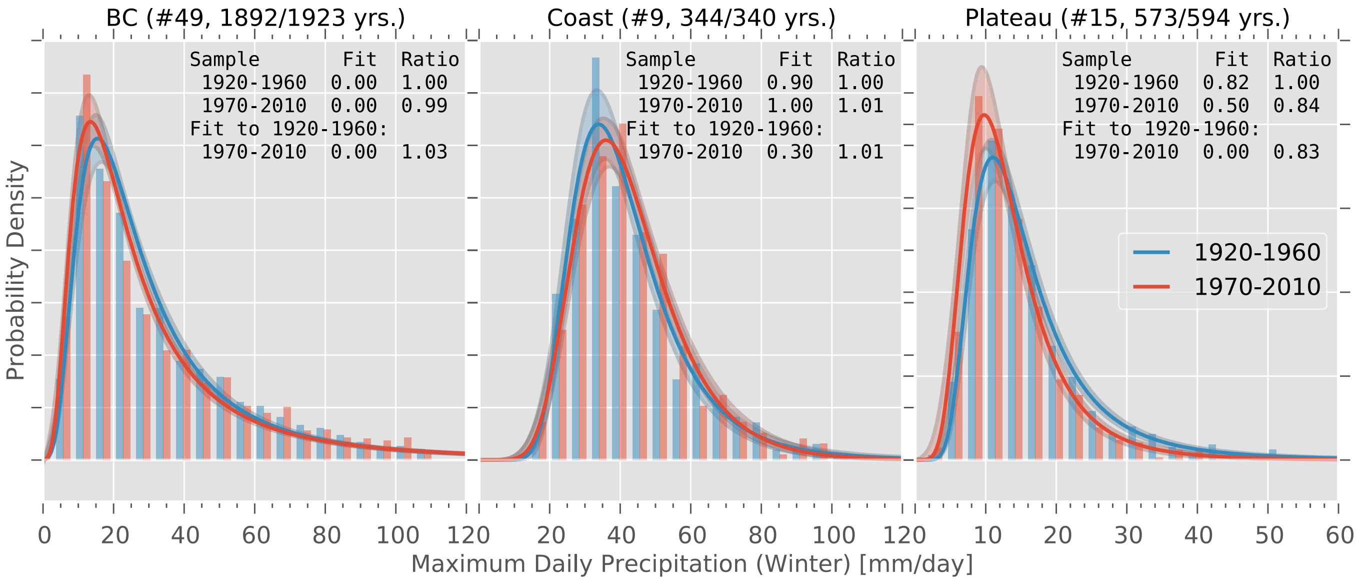

Fig. 1. Histograms and GEV fits to maxima of daily winter precipitation totals for two 40 year periods: 1920 - 1960 (blue) and 1970 - 2010 (red). Each panel shows data aggregated over the region indicated in the title (the number of stations and the number of valid years per period are also shown in the title). The colored bands are 95% confidence intervals (bootstrapping); see text for further explanation of annotation.

stations within one group, which significantly increases the number of data points constraining each fit.

The Kolmogorov-Smirnov (K-S) test is used to gauge the quality of the fit, as well as detect differences between samples from different observation periods. The K-S test measures the maximum difference between the cumulative distribution function of two distributions. p-values for the quality of fit (middle column), as well as the ratios of the sample means to the first sample (right column) are printed in the upper right corner of each panel.[1] K-S test results from the comparison of the second to the first period and the ratio of the two distribution means are shown in the last row. In either comparison, the null hypothesis associated with the p-value is that the two samples come from the same parent distribution.

The GEV fit to the data aggregated over the entire province of BC is shown in the left panel (Fig. 1): the quality of fit is clearly unacceptable ($p \approx 0$). It is not possible to fit a GEV distribution over data from all stations in BC, because the assumption of identical distribution is violated due to the strong precipitation gradient. However, it is possible to fit a single GEV distribution to aggregated data from station clusters that have been constructed using the proposed clustering algorithm, because their climatologies are sufficiently similar (also for precipitation extremes). GEV fits to aggregated data from a station cluster at the coast and in the Interior Plateau are also shown in Fig. 1 (middle and right panel): evidently a good fit to the data is achieved. Furthermore, a statistically significant change in winter precipitation extremes in the Interior Plateau cluster can be detected, while no change is evident at the coast. This divergent trend was not visible in the province-level aggregation, which illustrates another advantage of the clustering technique, namely that different climate zones are separated naturally.

Acknowledgments

ARE was partially supported by a SOSCIP Graduate Student Fellowship. The research of WRP is supported by NSERC Discovery Grant 9627.

References

[1] F. W. Zwiers, L. V. Alexander, G. C. Hegerl, T. R. Knutson, J. P. Kossin, P. Naveau, N. Nicholls, C. Schär, S. I. Seneviratne, and X. Zhang, "Climate extremes: challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events," in *Climate Science for Serving Society*, pp. 339–389, Springer, 2013.

[2] É. Mekis and L. A. Vincent, "An overview of the second generation adjusted daily precipitation dataset for trend analysis in Canada," *Atmosphere-Ocean*, vol. 49, no. 2, pp. 163–177, 2011.

[3] V. V. Kharin, F. Zwiers, X. Zhang, and M. Wehner, "Changes in temperature and precipitation extremes in the CMIP5 ensemble," *Climatic Change*, vol. 119, no. 2, pp. 345–357, 2013.

[1] The p-values for the quality of fit are biased upwards, because the hypothesis was constructed using the same data as used for testing.