AN EXTREME ANALYSIS FOR THE 2010 PRECIPITATION EVENT AT THE SOUTH OF SASKATCHEWAN PRAIRIE

K.P. CHUN*
H.S. WHEATER

ABSTRACT

After a prolonged drought period in the early 2000s, the Canadian prairie experienced a remarkably wet year in 2010. Five stations near the edge of the Saskatchewan boreal forest recorded historically high cumulative precipitation (from April to September). The exceptional wet year causes the public concerns on flood controls and land use management in the region. Using the Canadian National Climate Data Achieve, characteristics of six-month cumulative precipitation sums over Saskatchewan prairie are investigated by the Generalised Extreme Value (GEV) Theory. Based on the unconstrained GEV distribution, the 2010 event is outside the estimated 95% confidence intervals for the five Canadian prairie stations. On the contrary, the exceptional high 2010 cumulative perception sums for the five stations are still bounded by the estimated confidence bounds if the GEV distribution is constrained to the Gumbel distribution (i.e. setting the shape factor of the GEV distribution to be zero). These results demonstrate that the classical extreme analysis is useful for planning unprecedented extreme events in the Canadian Prairie, if the GEV distribution is constrained to the Gumbel distribution with the estimated uncertainty bounds based on the order statistics.

KEYWORDS: Generalised Extreme Value (GEV) theory, Saskatchewan, prairie, confidence intervals, Gumbel distribution.

1. INTRODUCTION

Extreme precipitation events are the major cause of fluvial floods and natural hazard such as landslides which have major socioeconomic implications (Wheater, 2002). Ecological systems and agriculture practices are vulnerable to excessive high precipitation (Katz *et al*., 2005; Semenov, 2008). Moreover, using peak runoff for drainage system design is a common engineering practice (e.g. City of Saskatoon, 2008, Hong Kong Drainage Services Department, 2000). Characterising extremes is important for water resources management.

Since Gumbel (1957) proposed to use the theory of extreme values to project unobserved extremes, daily annual extreme precipitation is widely analysed based on the Generalised Extreme Value (GEV) theory which consists of three asymptotic probability distributions (e.g. Beirlant *et al*., 2004). However, many studies (e.g. Fowler and Kilsby, 2003) proposed that multi-day annual extreme sums may be more important for design and planning practices in flood control than annual single day extreme precipitation. For rainfall-runoff processes, cumulative precipitation extremes resulted from prolonged wet days is important because they affect antecedent conditions and flood response in catchments (c.f. Jakeman and Hornberger, 1993 and Wagener *et al*., 2004). Moreover, accumulative precipitation above normal over a long period (such as months) can lead to flooding caused by groundwater in low-lying area (Hughes *et al*., 2011). Therefore, with the considerations of flood and water resources, studying unusual consecutive precipitation is motivated.

In 2010, although there is no record breaking in single monthly extreme sums at the Canadian prairie, the consecutive wet summer months (from April to September) is very atypical for the region,

especially after prolonged drought years in the early 2000s. The aim of this paper is to explore whether the classic extreme value theory can be used to model the precipitation extremes in the

Central Canada in view of the 2010 precipitation event. The six month consecutive precipitation sums (from April to September) are used for this investigation. The reason for studying precipitation between April and September is that the period is generally the rainfall season of the Canadian prairie and it usually has the annual maximum of six month running sums. In the next session, details of the data are explained. Then, the results and the implications of using the GEV distribution and a simpler Gumbel distribution to characterise the prairie six month precipitation sums are discussed. In the concluding session, the importance of using suitable distribution to characterise the extremes are summarised and possible further research are presented.

2. DATA AND METHODOLOGY

Historical time series for this extreme analysis are from the National Climate Data and Information Archive (NCDIA). The data before 2007 are extracted from the NCDIA Canadian Daily Climate Data (CDCD) which contains data for 7815 stations. The data after 2007 is downloaded from the NCDIA Climate Data Online. Apart from the station data, the Adjusted and Homogenised Canadian Climate Data (AHCCD) (Mekis and Vincent, 2011) are also used as a control long record for assessing the general quality of the extracted NCDIA data.

Over Saskatchewan, there are 44 AHCCD stations but only 11 stations have enough data for studying the 2010 events. Nevertheless, the 11 AHCCD stations (Table 1 and Figure 1) spread fairly randomly across Saskatchewan. Five locations which AHCCD data does not have the 2010 data are also included this study because they have relative long record and located near the edge of the boreal forest where is considered to be the boundary of the climate zone shift.

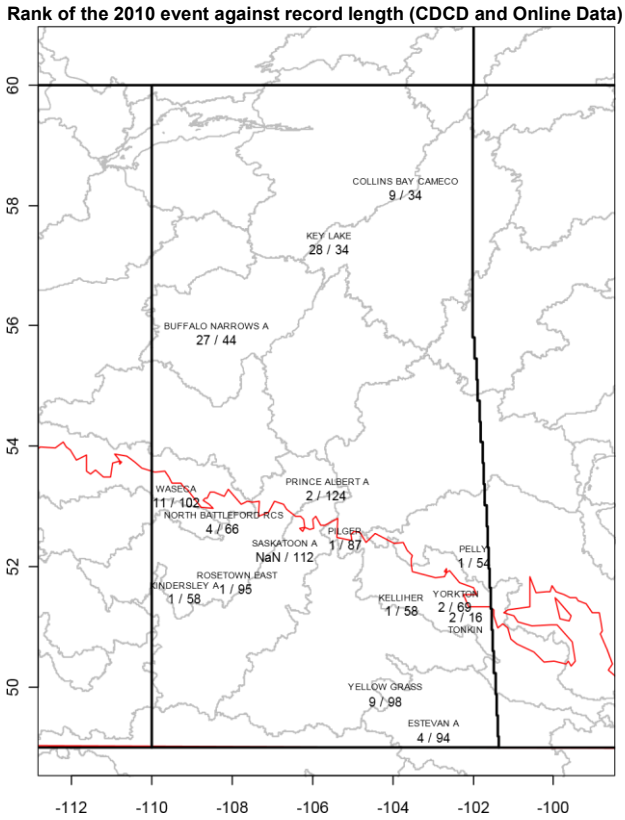


Figure 1. Rank of the 2010 event against record length (CDCD and online data)

Table 1. Details of the station data

CSN is climate station number; Y in the AHCCD column indicates that the station is an Adjusted and Homogenised Canadian Climate Data stations (Mekis and Vincent, 2011)

Name	District ID	CSN	Longitude	Latitude	Elevation	AHCCD
BUFFALO NARROWS	406	980	-108.5	55.9	421	
BUFFALO NARROWS	406	981	-108.5	55.8	423	
BUFFALO NARROWS A	406	982	-108.4	55.8	440	Y
BUFFALO NARROWS (AUT)	406	983	-108.4	55.8	440	
COLLINS BAY	406	1629	-103.7	58.2	492	
COLLINS BAY	406	1630	-103.7	58.2	492	
COLLINS BAY CAMECO	406	1632	-103.7	58.2	490	Y
ESTEVAN	401	2390	-103.1	49.2	566	
ESTEVAN A	401	2400	-103.0	49.2	581	Y
KELLIHER	401	3660	-103.8	51.3	676	Y
KEY LAKE	406	3755	-105.6	57.3	509	Y
KEY LAKE A	406	3759	-105.6	57.3	510	
KINDERSLEY A	404	3900	-109.2	51.5	694	Y
KINDERSLEY CDA EPF	404	3904	-109.2	51.5	681	
KINDERSLEY KY	404	3920	-109.2	51.5	683	
PELLY	408	6000	-101.9	52.1	509	Y
PELLY 2	408	6001	-101.9	51.7	499	
PILGER	405	6120	-105.2	52.4	552	Y
TONKIN	401	9082	-102.2	51.2	527	Y
WASECA	404	8520	-109.4	53.1	638	Y
YELLOW GRASS	401	9040	-104.2	49.8	580	Y
SASKATOON A	405	7120	-106.7	52.2	504	Y
SASKATOON RCS	405	7165	-106.7	52.2	504	
SASKATOON SRC	405	7180	-106.6	52.2	497	
YORKTON A	401	9080	-102.5	51.3	498	
YORKTON	401	9085	-102.5	51.3	498	
YORKTON CDA EPF	401	9090	-102.4	51.1	504	
NORTH BATTLEFORD A	404	5600	-108.3	52.8	548	
NORTH BATTLEFORD	404	5605	-108.3	52.8	548	Y
ROSETOWN	404	6879	-108.0	51.6	586	
ROSETOWN CDA EPF	404	6880	-107.9	51.5	591	
ROSETOWN EAST	404	6884	-107.9	51.6	586	
PRINCE ALBERT	405	6230	-105.8	53.2	437	
PRINCE ALBERT A	405	6240	-105.7	53.2	428	Y

2.1 Extreme value theory

As multiple-day extremes may have an important role for design and planning practices in flood control (e.g. Fowler and Kilsby, 2003), six month cumulative sums (from April to September) are studied here. The reason of studying these sums is that the classic extreme value theory may be applicable. Generally, the cumulative sums from April to September are the peak cumulative sum of a year (i.e. a block), so that, according to the extreme value theory, the distribution of the cumulative sums should converge weakly to the Generalised Extreme Value (GEV) distributions G(z).

$$1 \qquad \frac{z-\mu}{\qquad}\Big)\Big]^{-\frac{}{\xi}}$$

where μ, σ and are the position, scale and shape parameters respectively. There are two constraints of the GEV distribution: and σ > 0.

An iterative maximum likelihood approach (Coles, 2001) is used for the extreme value distribution parameter estimation. The main advantage of using the maximum likelihood approach is that the variance matrix of the parameters is also estimated during the parameter estimation. As a general result of maximum likelihood estimation, the estimated parameters are asymptomatically normal distributed. Although it is shown that L-moment or other method may provide more robust parameter estimation for the short extreme series (e.g. Hosking and Wallis, 1997), the record lengths in this study are relatively long (>30 years), the parameter estimation is expected to be not sensitive to parameter estimation methods

In addition to estimating the extreme distribution for the cumulative sums, the confidence bounds of cumulative sum for different non-exceedance probability are also studied as a statistical inference tool. For the extreme distributions, the normal approximation is one of the most widely used confidence bounds estimations. In the normal approximation (Coles, 2001), the quantile estimator (∅ is inferred as normal distributed

$$\emptyset \sim \quad \emptyset ,$$

where the estimator variance is expressed as

$$_{\emptyset}\emptyset \qquad\qquad \emptyset$$

with

$$\nabla\emptyset = \left[\frac{\partial\emptyset}{\partial\theta}, \ldots, \frac{\partial\emptyset}{\partial\theta}\right]$$

given that the d-dimensional parameters ( ), the parameter covariances ( ) and the quantile function (∅).

However, in a simulation experiment, Chun (2011) showed that the confidence interval derived by normal approximation is not adequate for the high order extremes and the order statistics (e.g. Castillo *et al*., 2005) give more consistent confidence bound estimation. Both the normal approximation and the order statistics are used to estimate confidence bounds of the six month sum series as a performance comparison.

3. RESULTS

Figure 1 and Table 1 show 16 places in Saskatchewan where their precipitation is investigated. In some of 16 locations, there are observations from more than one station because of various reasons such as change of instruments, additional measurements, shifts of stations etc. For example, the investigated observations of Saskatoon are from three stations. As a comparison, Figure 2 shows the AHCCD time series (a black line) against the data from the three Saskatoon stations (points). Generally, all the data from different stations and the AHCCD series are consistent with each other, and the 2010 precipitation sum is exceptional high. Moreover, Figure 2 shows that the Saskatoon Airport (4057120) station has the longest historical precipitation record but the 2010 event for the station is missed. Providentially, two other Saskatoon stations (4057180 and 4057165) near Station 4057120 have the 2010 record. Moreover, Figure 3 shows that the two stations (4057180 and 4057120) are significantly correlated (p-value is near to 0 based on F-statistic). Therefore, the

Saskatoon time series used for extreme analysis is a combination of the Station 4057120 data and the 2010 precipitation data from Station 4057180.

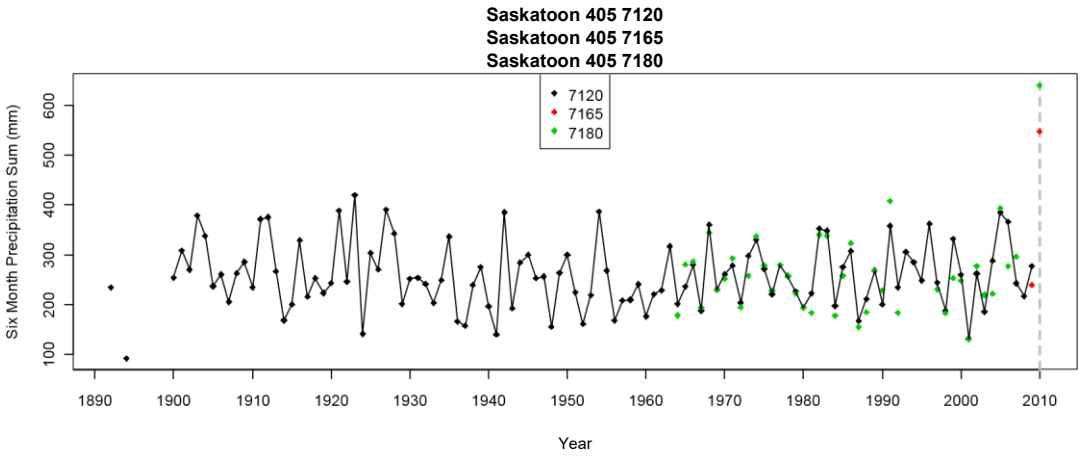


Figure 2. Historical Saskatoon Station precipitation

The earliest record of Saskatoon is back to the 1890s. The black line and points is the Saskatoon Airport (4057120) data. Generally, Stations 4057120 and 4057180 are corresponding well. It may be questionable to use the 2010 event of Station 40571810 to be that of the Saskatoon Airport (4057120) but this is the best available as the data for the Saskatoon Airport is missing in 2010

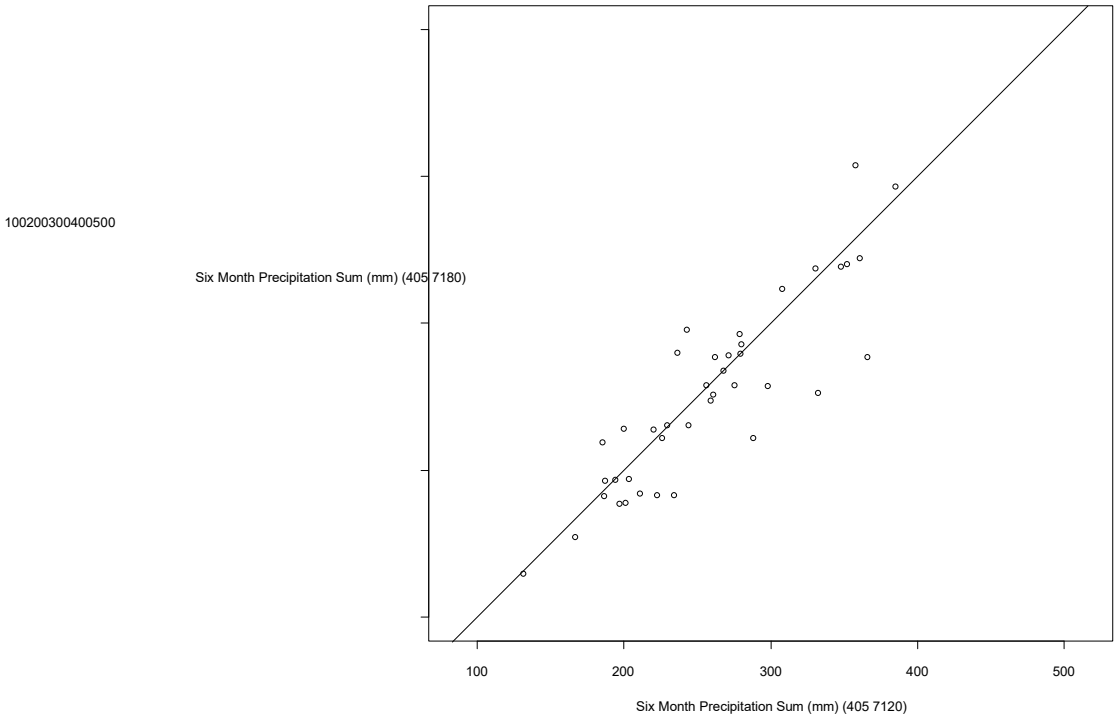
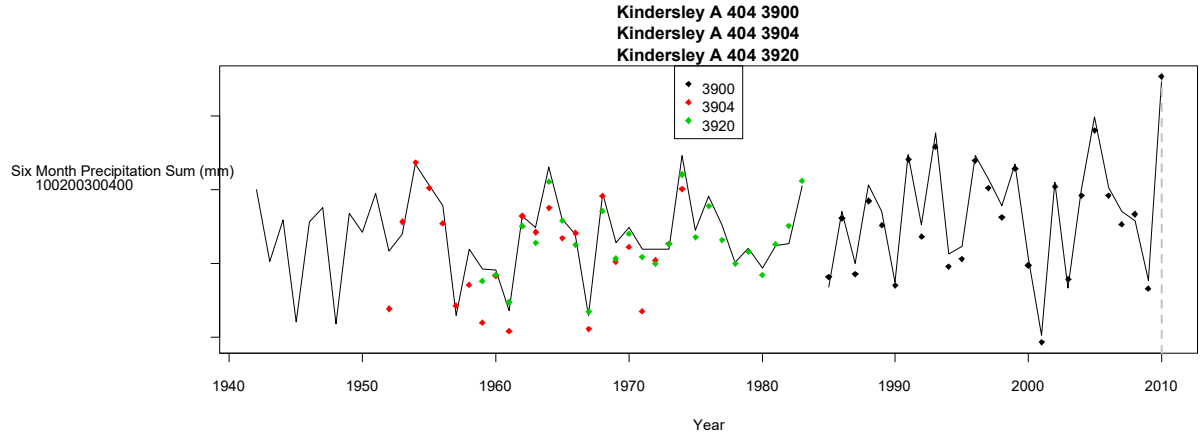
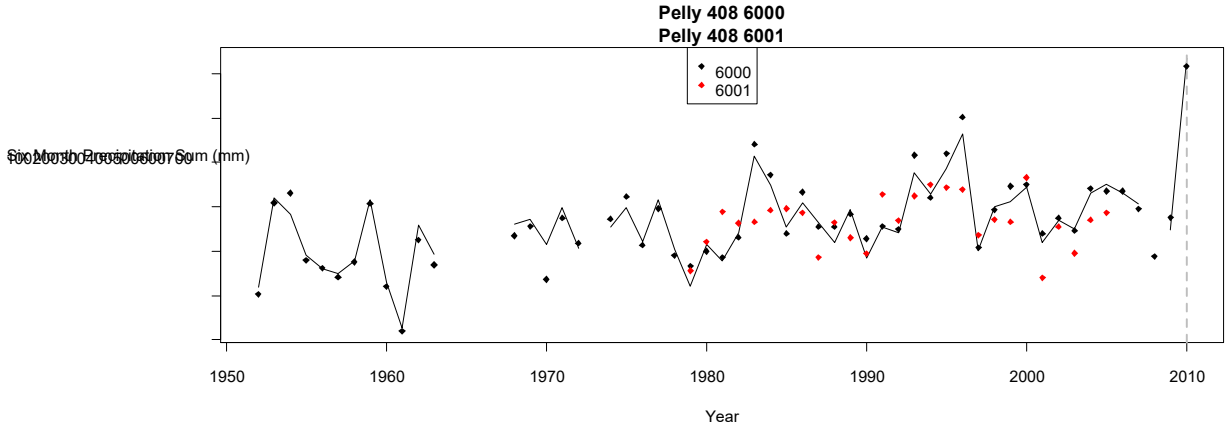


Figure 3. The Station 4057120 data is plot against the station 4057180 data. The black line is a 1:1 line as a correlation comparison reference

Using the pooled series of 16 locations, Figure 1 shows the ranks of the 2010 event against their record lengths. Similar to Saskatoon, four stations near Saskatoon along the boundary of the boreal forest have the highest rank for the 2010 event. Time series of these four locations (Kindersley, Pelly, Pilger and Rosetown) are shown in Figure 4, and they all indicate that the 2010 precipitation are unprecedented high for the area. As the 2010 six month precipitation sum is identified to be exceptional high for some locations near the boreal forest, the next question to be addressed in this paper is whether classic extreme value theory can be used for characterising for these locations.





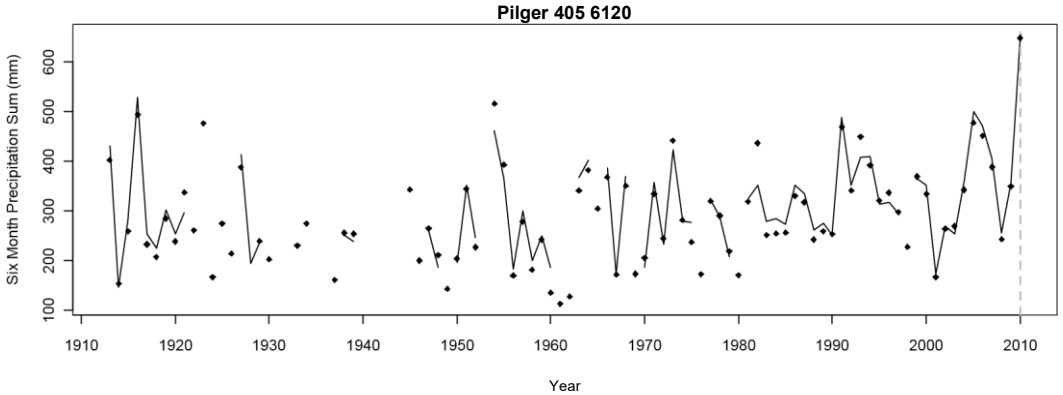


Figure 4. Historical South Saskatchewan Station precipitation
Different colour points indicate the data from different climate stations. The black lines are the AHCC data. Rosetown East does not have the corresponded AHCC data

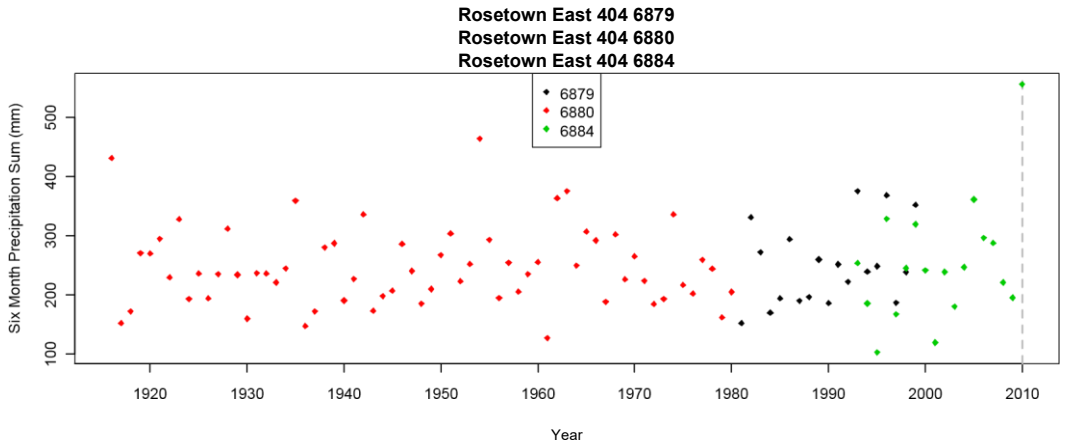


Figure 4 (continued). Historical South Saskatchewan Station precipitation
Different colour points indicate the data from different climate stations. The black lines are the AHCC data. Rosetown East does not have the corresponded AHCC data

The left panel of Figure 5 shows the histogram of the Saskatoon six-month precipitation sum and the GEV probability density estimated by the maximum likelihood. The GEV density estimation appears to match the histogram. In the right panel of Figure 5, the empirical autocorrelation plot show that the Saskatoon six-month precipitation sums do not violate the independent assumption of the GEV theory as the empirical autocorrelations are generally not over two dash lines which indicate whether the autocorrelations are significantly different from zero.

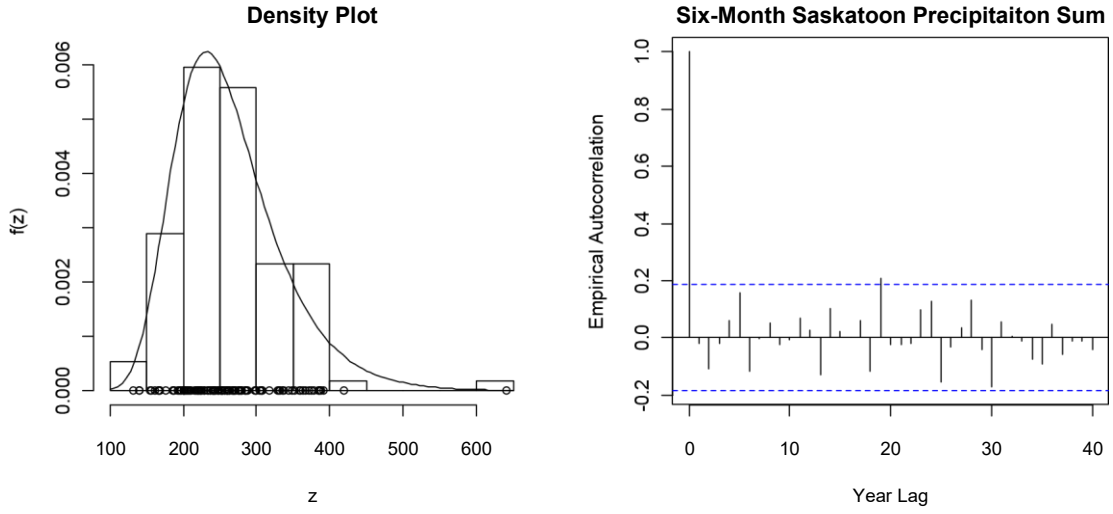


Figure 5. Diagnostic plots for the fitted GEV distribution (left) and independent assumptions (right)

In classic extreme analysis, Gumbel plots are extensively used because a straight line graph would be obtained on a Gumbel plot if the analysed data are Gumbel distributed. However, using Gumbel plot, heuristic plotting position formula are needed to estimate the probability of extremes based on their ranks. The used plotting position formula for the Saskatoon Quantile and Gumbel plot in Figure

6 is the Weiibull formula. In the left panel of Figure 6, the Quantile plot shows that the nonexceedance probability estimated from the Weibull formula is consistent with the GEV distribution except from the 2010 event. The middle and right panels of Figure 6 show the Gumbel plots with the 95% confidence bounds derived by the normal approximation and the order statistics respectively. The 2010 event are distinctly outside the 95% normal approximated confidence bounds of the two Gumbel plots.

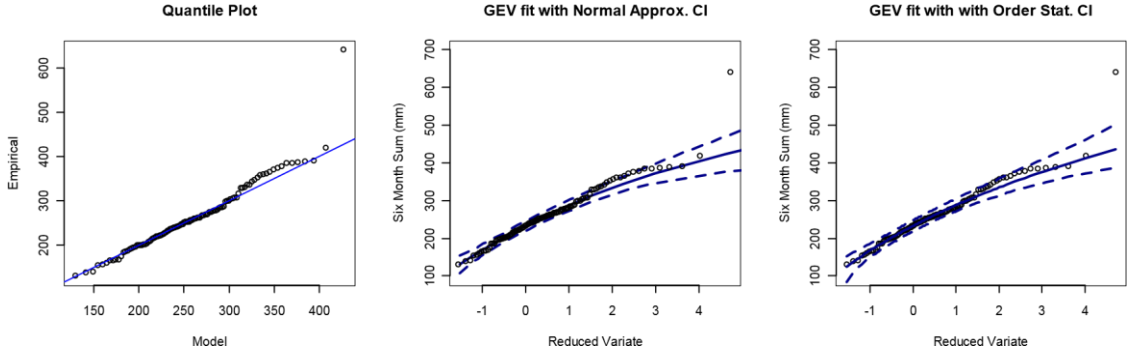


Figure 6. Diagnostics plots are based on the Weibull plotting position formula. The GEV confidence intervals (blue dash lines) in the middle and right panels are dived by the normal approximation

As increasing evidence for the nonstationary extreme pattern, the heuristic position formula needs to be used with caution and benchmark with other approaches. Instead of using heuristic plotting position formula, Chun (2011) proposed to plot extremes directly against the Rth largest order. Figure 7 shows the Saskatoon six month sums against their Rth largest orders. With the confidence bounds based on the order statistics, the left panel shows the GEV distribution fit and the right panel shows the Gumbel distribution fit. Apart from the 2010 event, the observations are generally fitted well with the either distributions and bounded in the estimated confidence bounds. It is interesting to note that the 2010 event is within the 95% confidence bound of the Gumbel distribution fit in Figure 7. The reason for the Gumbel distribution give better confidence bounds is that the estimated GEV distribution in Figure 7 has a negative shape factor and it has upper bound of possible extremes, whereas a Gumbel distribution has unbounded upper possible extremes.

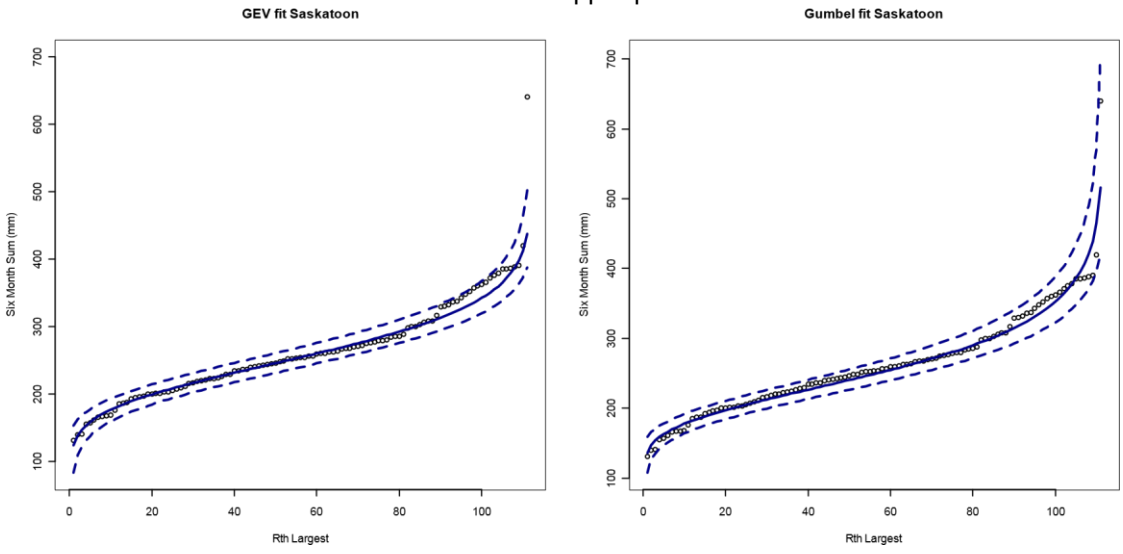


Figure 7. The Saskatoon GEV and Gumbel fit (solid lines) plots with the 95% confidence intervals (dash lines) based on the order statistic

Further to the results of Figure 7, extreme analysis is also performed for the six month precipitation sum shown in Figure 4. Figures 8 and 9 show the GEV and Gumbel fits for the four sites respectively. The results are generally consistent with Figure 7, aside from that the GEV confidence interval can bound the Pilger 2010 event. Overall, if the 2010 event is considered to be an unobserved event and

is predicted by a distribution fitted from the events before 2010, the Gumbel distribution is a better model than the GEV model because the confidence bounds of the Gumbel model include the 2010 event.

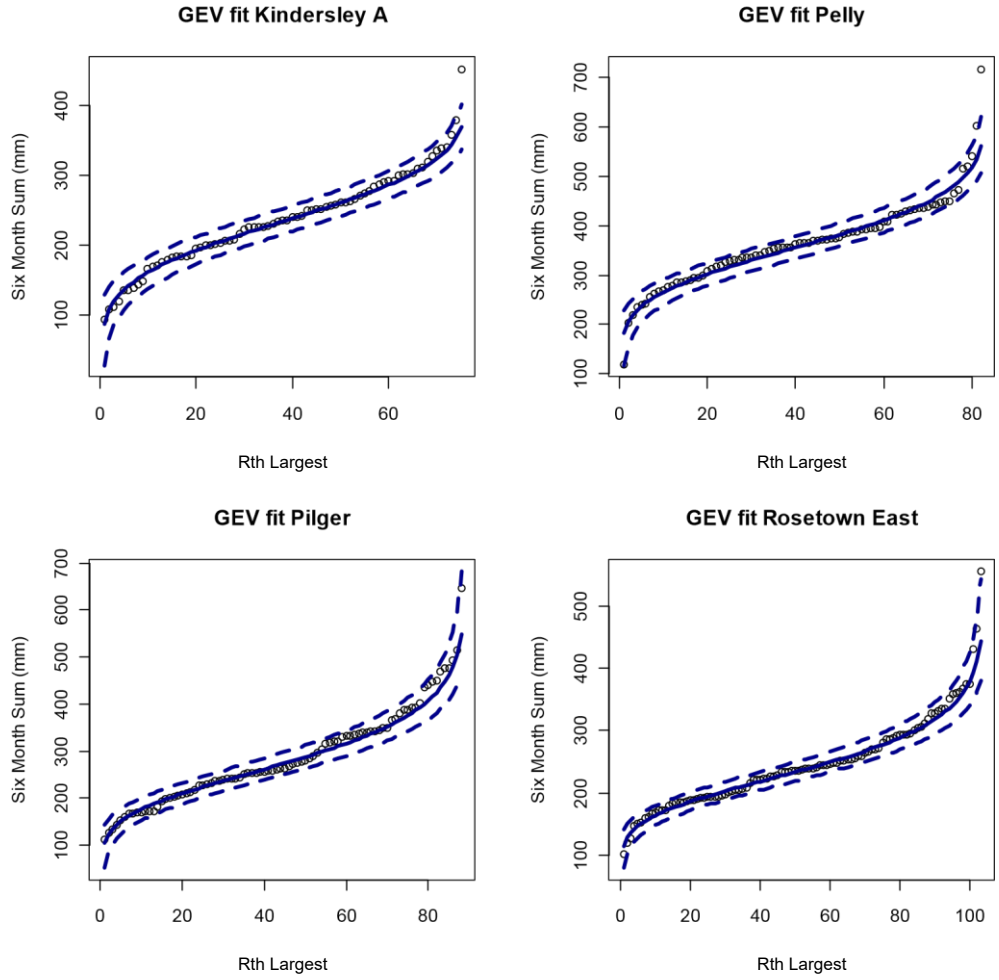


Figure 8. The GEV fit (solid lines) plots with the 95% confidence intervals (dash lines) based on the order statistic for the four stations near Saskatoon

The classic Gumbel plots for the Saskatoon Gumbel fit are shown in Figure 10. The left panel gives the confidence interval estimated by the normal approximation and the right panel provides the confidence interval derived by the order statistics. The result is consistent with the assertion (Chun, 2011) that the confidence intervals from the order statistics are better than those estimated by the normal approximation (Coles, 2001).

As a further investigation on how the 2010 event affects the GEV distribution parameters, Figure 11 provides a summary of the parameter evolution in months of 2010. In Figure 11, the three panels from left to right of each row present the location, scale and shape parameters for one of five studied areas. Based the historical events before 2010, the solid horizontal black lines are the parameter estimations and the dash black lines are the corresponding parameter confidence intervals. For the parameter estimation including the consideration of the 2010 event, a possible approach is to use the sum of historical median as the 'expected' 2010 event to estimate the distribution parameters. The blue horizontal dash lines are the estimate using the sum of historical monthly medians. All the blue dash lines on Figure 11 are very close to the black solid lines, and this indicates that the parameter estimation is not sensitive to an additional median event for the five investigated area as the stations have relative long historical record (>50 years).

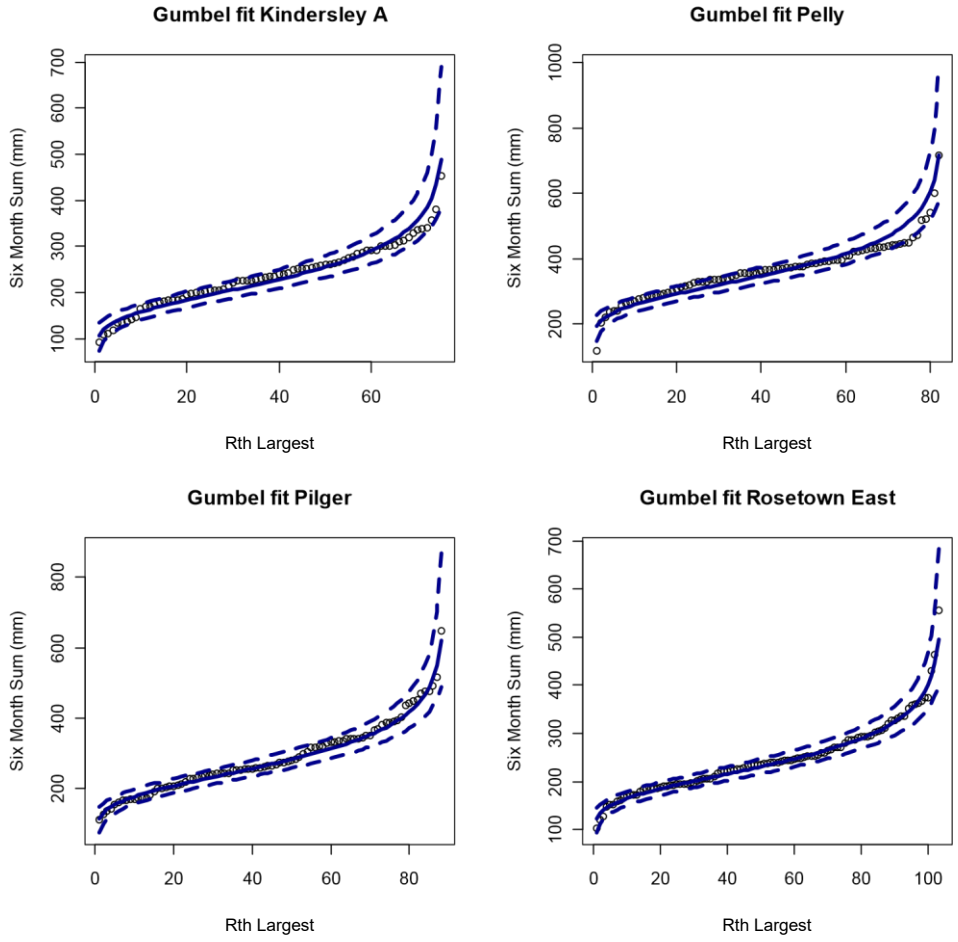


Figure 9. The Gumbel fit (solid lines) plots with the 95% confidence intervals (dash lines) based on the order statistic for the four stations near Saskatoon

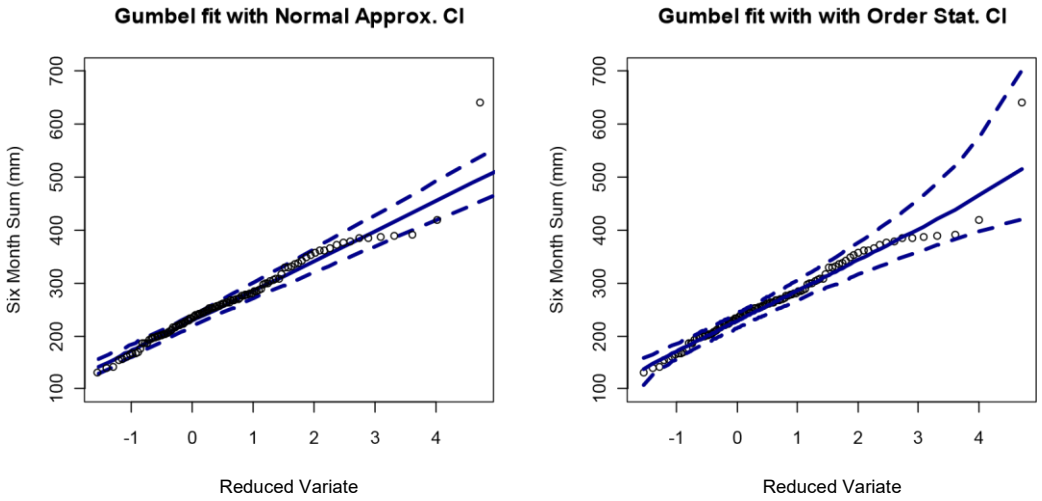


Figure 10. Confidence intervals for the Saskatoon Gumbel fit using the normal approximation (left) and the order statistic approach (right)

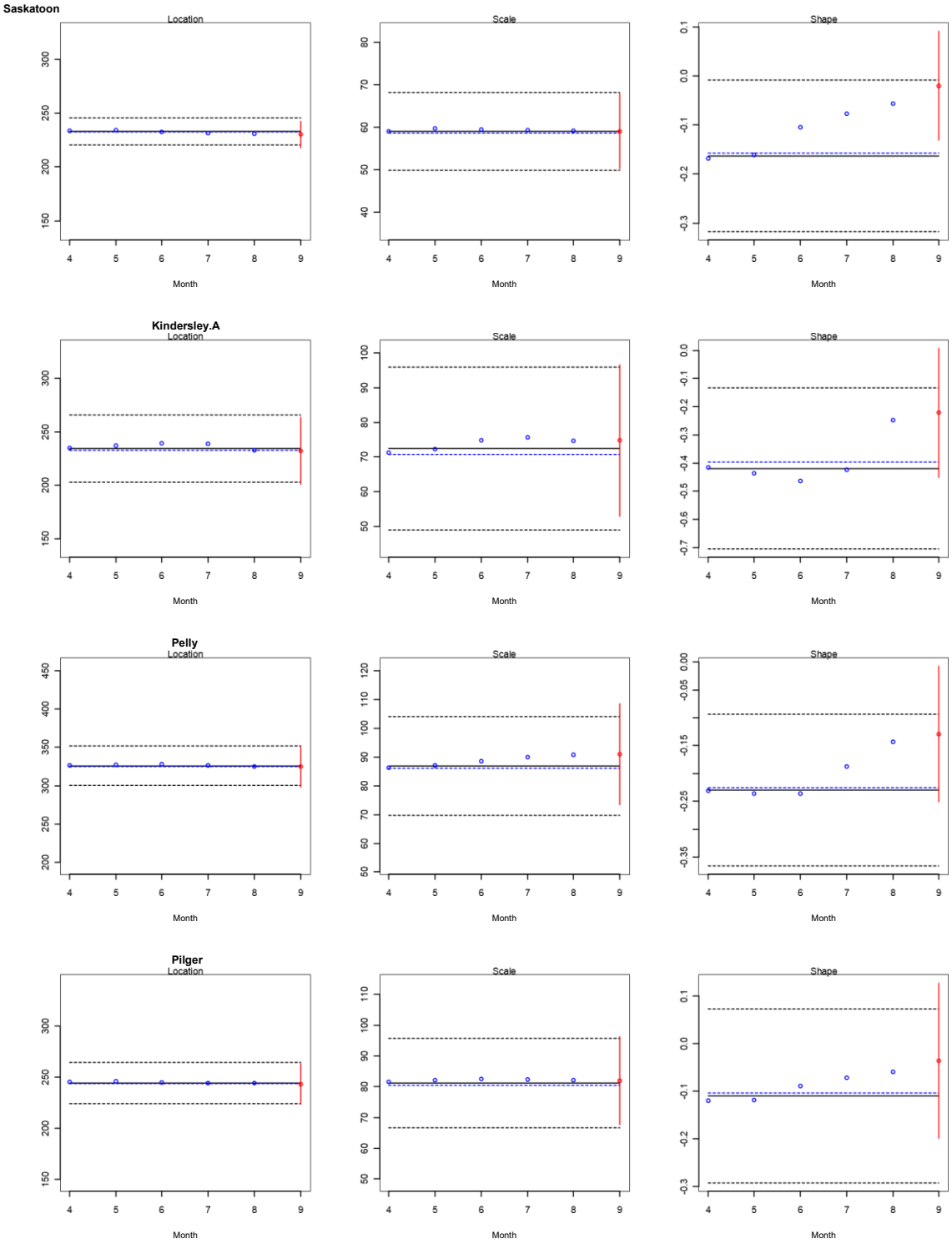


Figure 11. The evolution of the GEV parameters of Saskatoon, Kindersley, Pelly, Pilger and Rosetown East in the 2010 summer

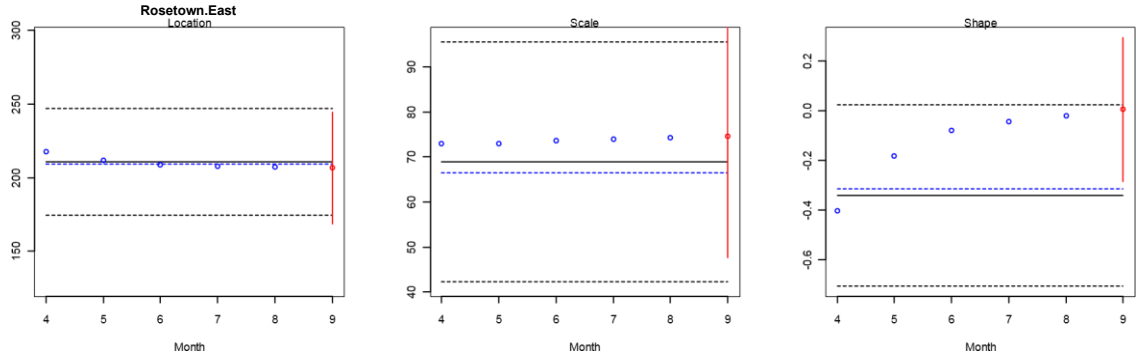


Figure 11 (continued). The evolution of the GEV parameters of Saskatoon, Kindersley, Pelly, Pilger and Rosetown East in the 2010 summer

The blue points in the Figure 11 are the parameters estimated from the sum of the 2010 observations until the end of the month shown on the x-axis and the monthly medians for the remain months. The red points and the red vertical line are the parameter and corresponding confidence intervals estimated using all the observations before 2010 and 2010. It is interesting to note that the location and scale parameters are not sensitive for the 2010 event but the shape parameters are sensitive. The shape factors are shifting from slightly negative to near zero and this change has important implication of characterising extremes. The shift of the shape parameters indicate that the 2010 event increase the uncertainty of the upper possible extremes, as the Weibull distributions (i.e. having negative shape parameters) has an extreme upper bound whereas the Gumbel distribution (i.e. having zero shape parameters) has no extreme upper bound.

The conjecture for the fitted distributions having negative shape factors based on the data before 2010 is that the missing data of the historical data give a wrong signal of the existence of a possible extreme upper bound. The missed historical data may be a result of equipment failures due to extremely high precipitation intensity instead of completely random missing data. Although it is generally believed that longer data series would provide better parameter estimation, using long historical record has to be cautious as the reliability of historical data is likely to decrease. The older records are more like to be plagued with systematic missing data. For example, the number of missing day is significant related to the Saskatoon six month sums (p-value for F-statistic = 0) and there are more missing days for the older record.

4. CONCLUSIONS

Several areas near the boreal forest in Saskatchewan experienced an exceptional wet year in 2010. Using GEV distributions with negative shape factor cannot properly characterise the 2010 event. Moreover, the 2010 event is within the Gumbel distribution confidence bonds estimated from the order statistics but the confidence bounds estimated by the normal approximation (Coles, 2001) fail capture this event. When the GEV distribution are compared to the Gumbel distribution, the GEV distribution has more feasible shape and a 'narrower' confidence bonds (uncertainty) but it seems to be more sensitive to missing data (less robust). Therefore, based on the above results, using Gumbel distributions with the order statistic confidence bounds appear to be more suitable than using GEV distribution to classify the six month cumulative extremes near the boreal forest in Saskatchewan.

One of the difficulties of the current study is that there are only limited long historical records over Saskatchewan. Moreover, incidentally missing data, shifting station location and changing instrument increase the difficulty of analysis and reduce the reliability of the results. Despite statistical inference alleviating these problems, continuous measurement, redundant measurements and good quality control are deemed important in Saskatchewan.

In some recent climate change extreme studies (e.g. Maraun *et al*. (2010, 2009a) and Rust *et al*. (2009)), shape parameters are assumed to be constant because it is deemed to be a common practice in extreme value statistics (Maraun *et al*. 2010). However, Figure 11 may raise a question to this assumption. Although current results do not strongly support that the shape factors of GEV for

the five studied stations are changing, Figure 11 show that shape parameters are sensitive to new observations and they are deserved more investigation.

The Bayesian extreme framework proposed by Coles (2003) is a possible approach that can be further applied to this study. From the results here, the prior distributions of the shape parameter can be defined to be a distribution which has its peak close to zero with a narrow spread. As a result, the shape parameter estimation can be constraint by the experience here or the knowledge of the shape parameter. Furthermore, the multivariate extreme framework (Coles, 2001) and the Generalised Linear Model (GLM) typed extreme approach (Maraun *et al*., 2010, Chun, 2011) may also be considered to allow climate variables to condition extreme value distribution and non-stationary extreme model.

ACKNOWLEDGMENT

We would like to thank Dr. Garth van der Kamp of Environment Canada for his discussion related to the Saskatoon precipitation series and Eva Mekis of Environment Canada for providing us the Adjusted and Homogenised Canadian Climate Data (AHCCD).

REFERENCES

Beirlant J., (2004) Statistics of extremes: theory and applications, John Wiley & Sons Inc.

Castillo E., Hadi A.S., Alegría J.M.S. and Balakrishnan N., (2005). Extreme value and related models with applications in engineering and science. Wiley Chichester.

City of Saskatoon, (2008). New Neighbourhood Design and Development Standards Manual. Canada: City of Saskatoon.

Coles S., (2001). An introduction to statistical modeling of extreme values. Springer Verlag.

Coles S., Pericchi L.R. and Sisson S., (2003). A fully probabilistic approach to extreme rainfall modeling, *Journal of Hydrology,* **273**(1-4), 35-50.

Chun K.P., (2011). Statistical Downscaling of Climate Model Outputs for Hydrological Extremes. PhD edn. UK: Imperial College London.

Fowler H. and Kilsby C., (2007). Using regional climate model data to simulate historical and future river flows in northwest England, *Climatic Change,* **80**(3), 337-367.

Gumbel E.J., (2004). Statistics of extremes. Dover Pubns.

Hong Kong Drainage Services Department, (2000). Stormwater Drainage Manual. Hong Kong: Government of the Hong Kong Special Administrative Region.

Hosking J.R.M. and Wallis J.R., (1997). Regional frequency analysis. Regional Frequency Analysis, by JRM Hosking and James R.Wallis, pp.240.ISBN 0521430453.Cambridge, UK: Cambridge University Press, April 1997, **1**.

Hughes A., Vounaki T., Peach D., Ireson A., Jackson C., Butler A., Bloomfield J., Finch J. and Wheater H., (2011). Flood risk from groundwater: examples from a Chalk catchment in southern England, *Journal of Flood Risk Management,* **4**(3), 143-155.

Jakeman A. and Hornberger G., (1993). How much complexity is warranted in a rainfall-runoff model?, *Water Resources Research,* **29**(8), 2637-2650.

Katz R.W., Brush G.S. and Parlange M.B., (2005). Statistics of extremes: modeling ecological disturbances. *Ecology,* **86**(5), 1124-1134.

Maraun D., Osborn T.J. and Rust H.W., (2011). The influence of synoptic airflow on UK daily precipitation extremes. Part I: Observed spatio-temporal relationships, *Climate Dynamics,* **36**(1), 261-275.

Maraun D., Rust H. and Osborn T., (2009). The annual cycle of heavy precipitation across the United Kingdom: a model based on extreme value statistics, *International Journal of Climatology,* **29**(12), 1731-1744.

Mekis É. and Vincent L.A., (2011). An Overview of the Second Generation Adjusted Daily Precipitation Dataset for Trend Analysis in Canada, *Atmosphere-Ocean,* **49**(2), 163-177.

Rust H., Maraun D. and Osborn T., (2009). Modelling seasonality in extreme rainfall: a UK case study, *Europ.Phys.J.Special Topics,* **174**, 99-111.

Semenov M.A., (2007). Simulation of extreme weather events by a stochastic weather generator, *Climate Research,* **35**(3), 203.

Wagener, T., Wheater, H. and Gupta, H.V., (2004). Rainfall-runoff modelling in gauged and ungauged catchments. Imperial College Pr.

Wheater H., 2002. Progress in and prospects for fluvial flood modelling, *Philosophical Transactions of the Royal Society of London.Series A: Mathematical, Physical and Engineering Sciences,* **360**(1796), 1409-1431.