Discrimination between statistical distributions for hydro-meteorological frequency modeling

Ismaila Ba

MSc Student, Department of Mathematics and Statistics

Université de Moncton

INTRODUCTION

- The identification of a statistical distribution to model the frequency of occurrence of extreme hydro-meteorological events is important.
- Two-parameter distributions such as the Generalized Pareto, lognormal, gamma or Weibull are useful in fitting datasets in areas such as POT extreme value modeling.
- Three-parameter distributions are also very important, such as for fitting annual maximum flood or precipitation series.

INTRODUCTION

- We recommend some methods of discrimination between distributions.
- The discriminations considered are between: Generalized Pareto (GP) and Kappa, Gumbel and some alternative frequency models, and model pairs belonging to the group {generalized extreme value (GEV), Pearson type 3 (P3), generalized logistic (GLO)}.
- Four discrimination methods are compared by Monte Carlo Simulation in terms of their discrimination power and discriminaton bias.

DISCRIMINATION STATISTICS

- Anderson-Darling statistic (AD)

$$A = A(X_n) = -n - \sum_{i=1}^{n} \frac{2i-1}{n}\left[\ln \hat{F}\left(X_{(i)}\right) + \ln\left\{1 - \hat{F}\left(X_{(n+1-i)}\right)\right\}\right]$$

Gives more weight to observations in the tails of the distribution as compared to Cramér-von Mises (CvM) and Kolmogorov-Smirnov (KS) statistics.

- Ratio of maximized likelihood statistic (RML)

$$T = T(X_n) = \ln\left[\frac{L\left(X_n; \hat{\theta}_0\right)}{L\left(X_n; \hat{\theta}_1\right)}\right]$$

Most widely investigated method.

DISCRIMINATION STATISTICS

- Transformation to normality followed by the application of Shapiro-Wilk GoF statistic (TN.SW)

$$Z_i = \phi^{-1}\left(\hat{F}(X_i)\right)$$

$$S = \left[\sum_{i=1}^{n} v_i Z_{(i)}\right]^2 / \sum_{i=1}^{n}\left[Z_{(i)} - \bar{Z}\right]$$

- Transformation to normality followed by the application of the Probability plot correlation coefficient statistic (TN.PPCC)

$$Z_i^* = \Phi^{-1}\left(\hat{F}(X_i)\right)$$

$$W_i^* = \Phi^{-1}\left(p_i\right)$$

DISCRIMINATION STATISTICS

The TN.PPCC statistic is then calculated as follows:

$$R^* = \frac{\sum\left(Z^*_{(i)} - \bar{Z}^*\right)\left(W^*_i - \bar{W}^*\right)}{\left[\sum\left(Z^*_{(i)} - \bar{Z}^*\right)^2 \sum\left(W^*_i - \bar{W}^*\right)^2\right]^{0.5}}$$

- Note: When the two frequency models have the same number of unknown parameters, applying RML is equivalent to using AIC or BIC.

PARAMETER ESTIMATION METHODS

- In practice, the parameters of the model are unknown, so they need to be estimated from the data.

- We considered three parameter estimation methods:

- Maximum likelihood (ML)
- Moments (MOM)
- Probability weighted moments (PWM).

DISCRIMINATION STUDIES

The discriminations considered are between:

- GP and KAP models
- Gumbel and some alternative frequency models
- Model pairs belonging to the group {GEV, P3, GLO):

  GEV vs P3, GEV vs GLO and P3 vs GLO.

GP AND KAP

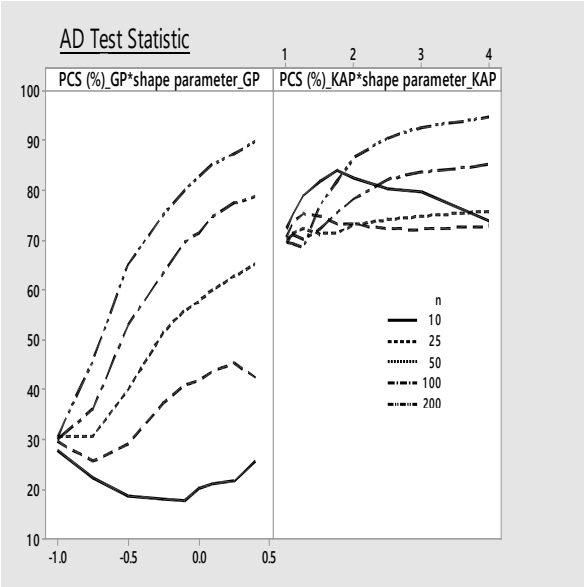


Fig. 1 PCS(%) using the AD statistic when GP is the true sampled distribution (left) and when KAP is the true sampled distribution (right).

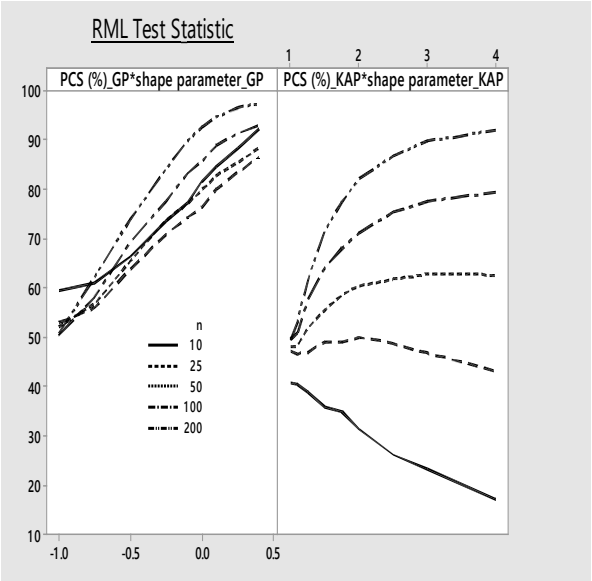


Fig. 2 PCS (%) using the RML statistic when GP is the true sampled distribution (left) and when KAP is the true sampled distribution (right).

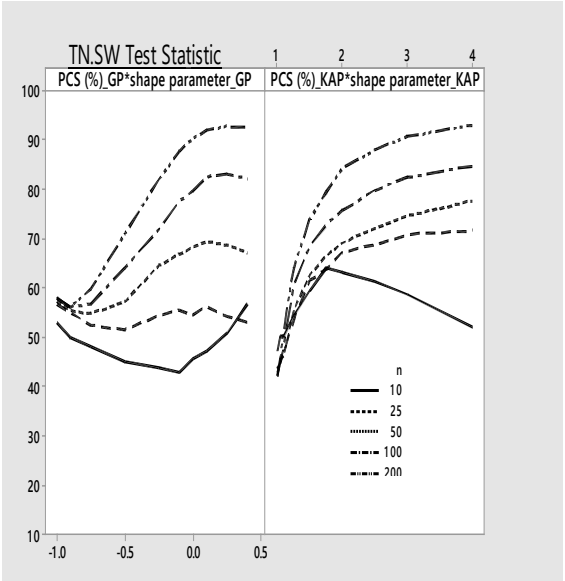


Fig. 3 PCS(%) using the TN.SW statistic when GP is the true sampled distribution (left) and when KAP is the true sampled distribution (right).

FloodNet NSERC

GP AND KAP

Application with Eight Hydrological Datasets

Table 1. Selecting between the GP and KAP distribution for fitting POT flood data at eight hydrometric stations

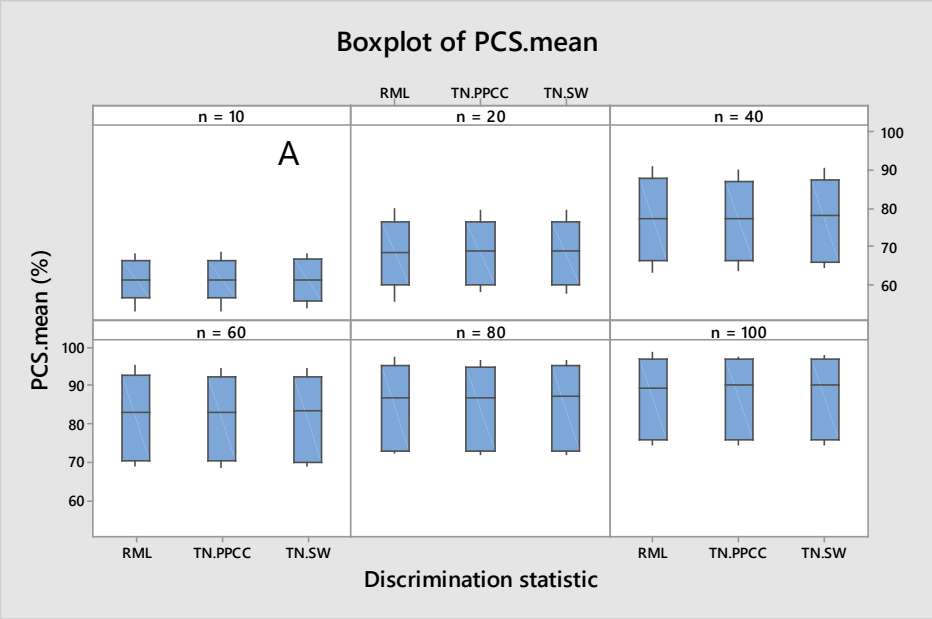
Station	n	GP Estimates (shape; scale)	KAP Estimates (shape; scale)	AD statistics (a_{GP}; a_{KAP})	TN.SW statistics (s_{GP}; s_{KAP})	RML statistics (t_{GP}; t_{KAP})
01AQ001	151	(-0.25; 19.98)	(1.95; 18.83)	(0.77; 0.23)	(0.989; 0.994) KAP is selected	(-1.88; 1.88) KAP is selected
01BL002	64	(0.01; 13.72)	(2.77; 12.89)	(1.06; 0.38)	(0.969; 0.982) KAP is selected	(-2.06; 2.06) KAP is selected
02FC002	159	(0.11; 119.5)	(2.80; 110.4)	(0.40; 1.08)	(0.992; 0.986) GP is selected	(3.03; -3.03) GP is selected
01BJ007	52	(-0.04; 430.8)	(2.20; 386.9)	(0.40; 0.61)	(0.983; 0.976) GP is selected	(1.05; -1.05) GP is selected
01AF007	47	(0.26; 28.49)	(4.00; 27.04)	(0.36; 0.24)	(0.985; 0.990) KAP is selected	(-0.33; 0.33) KAP is selected
04CA002	32	(0.03; 157.4)	(2.74; 156.2)	(1.00; 1.51)	(0.962; 0.959) GP is selected	(1.00; -1.00) GP is selected
02LB008	42	(-0.14; 41)	(1.92; 36.63)	(0.26; 0.41)	(0.977; 0.972) GP is selected	(0.81; -0.81) GP is selected
01BJ003	53	(0.13; 37.28)	(3.11; 35.65)	(0.27; 0.43)	(0.985; 0.987) KAP is selected	(0.52; -0.52) GP is selected

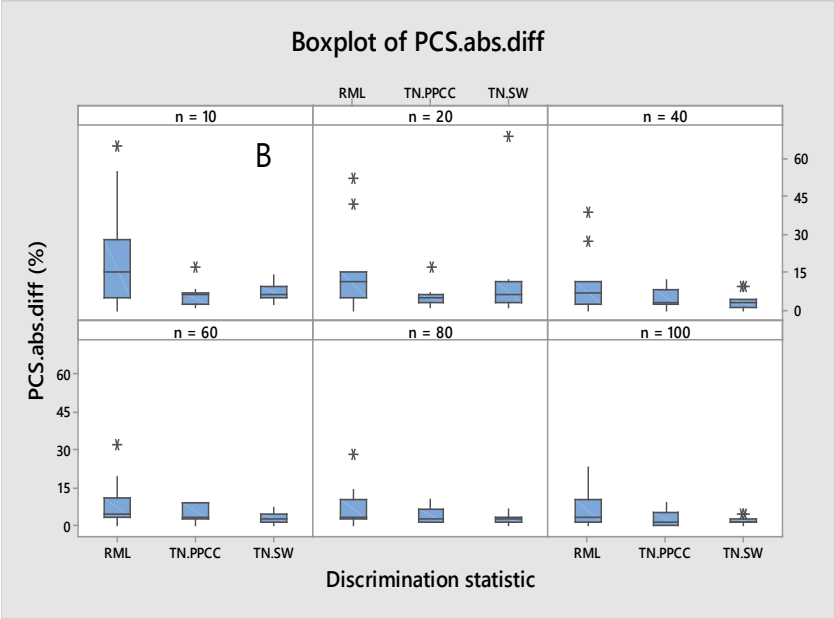
Gumbel and some alternative frequency models

The alternative models are:

- The normal: $N(\mu, \sigma)$
- The logistic: $LOG(\mu, \sigma)$
- Two student's t models to which a location parameter was added: $STU(\mu, \sigma, \nu)$
- Three models from the 3-parameter gamma family: $GAM3(\mu, \sigma, \gamma)$
- Four models from the GEV family: $GEV(\mu, \sigma, \gamma)$

Gumbel and some alternative frequency models





FloodNet
NSERC

GEV, P3 and GLO

Discrimination between GEV and GLO

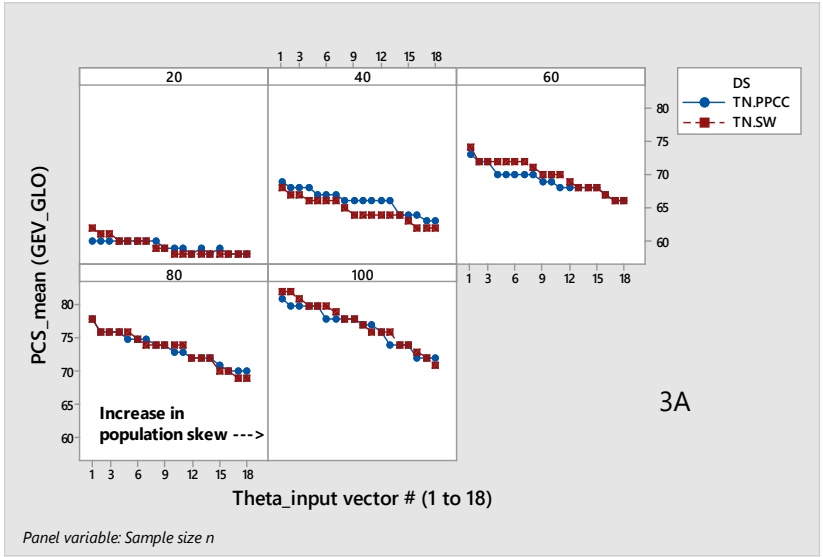


Fig. 5 PCS means for comparing TN.PPCC and TN.SW

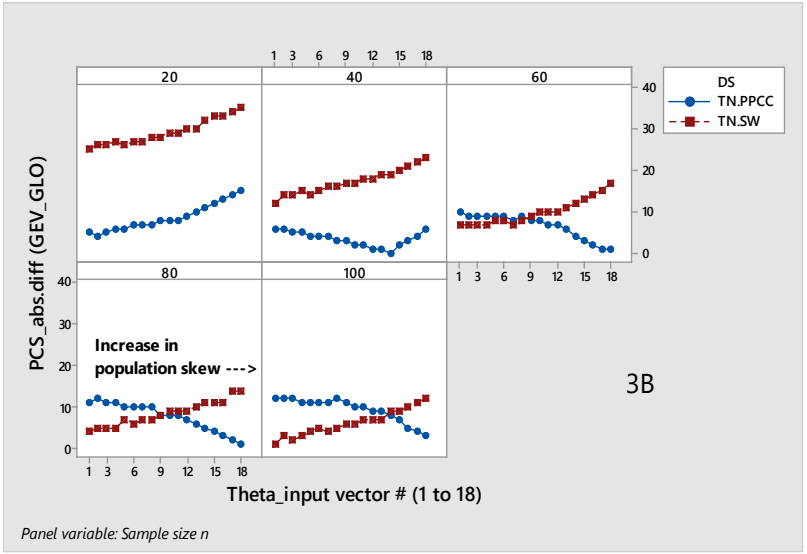


Fig. 6 PCS absolute difference for comparing TN.PPCC and TN.SW

FloodNet
NSERC

GEV, P3 and GLO

Discrimination between P3 and GLO

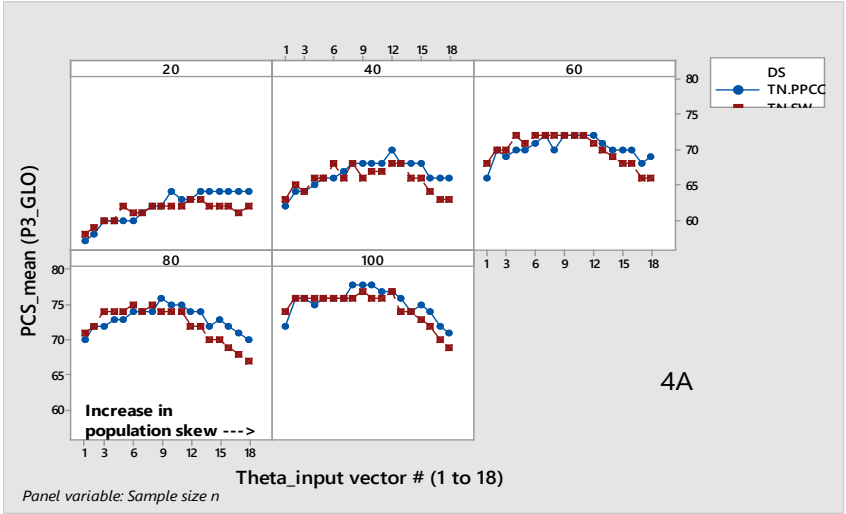


Fig. 7 PCS means for comparing TN.PPCC and TN.SW

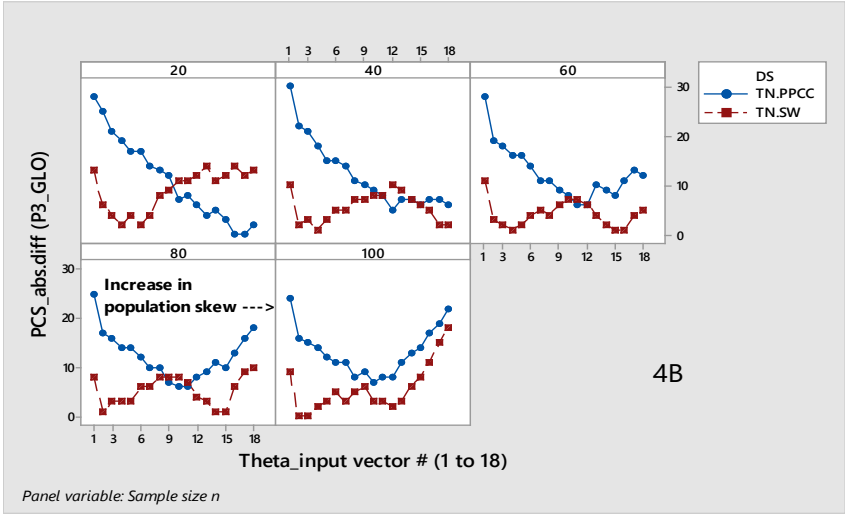
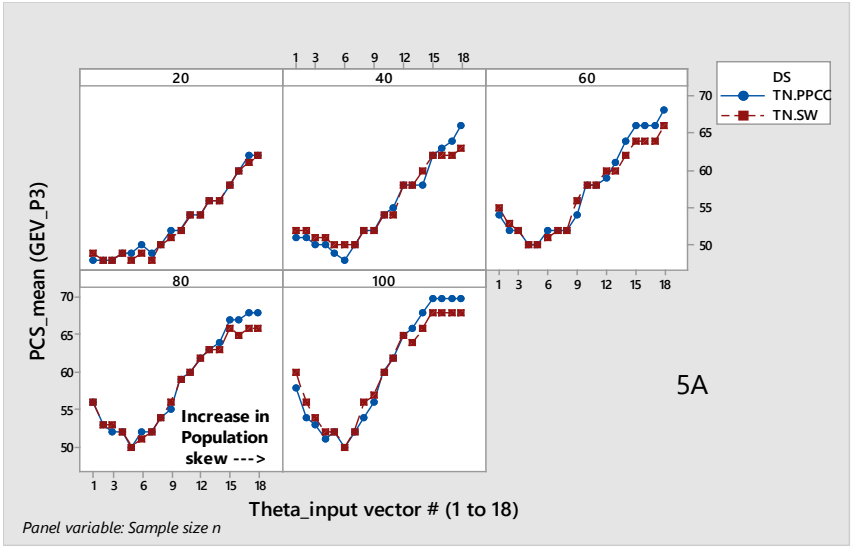


Fig. 8 PCS absolute difference for comparing TN.PPCC and TN.SW

FloodNet
NSERC

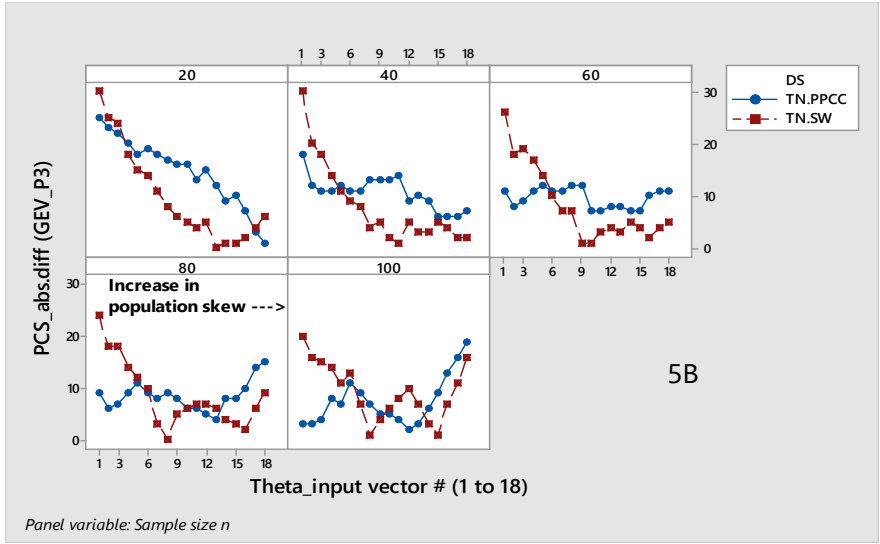
GEV, P3 and GLO

Discrimination between GEV and P3



5A

Fig. 9 PCS means for comparing TN.PPCC and TN.SW



5B

Fig. 10 PCS absolute difference for comparing TN.PPCC and TN.SW

FloodNet NSERC

GEV, P3 and GLO

Table 2. Use of the TN.SW statistic s and the TN.PPCC statistic r^* to choose between the GEV, GLO and P3 models for the 18 data series

ID	Sample size	s.GEV	s.P3	s.GLO	r*.GEV	r*.P3	r*.GLO	Chosen.Model
#1	52	0.98111	0.97965	0.97082	0.99298	0.99238	0.98853	GEV
#2	37	0.97856	0.97838	0.96846	0.99072	0.99062	0.98689	GEV
#3	100	0.98867	0.98887	0.98667	0.99488	0.99521	0.99473	P3
#4	45	0.97498	0.97569	0.96414	0.99007	0.99037	0.98531	P3
#5	42	0.98632	0.98535	0.98497	0.99129	0.99066	0.99273	*
#6	32	0.97078	0.97264	0.96182	0.98952	0.99029	0.98557	P3
#7	43	0.98013	0.98252	0.97143	0.9921	0.99269	0.98851	P3
#8	96	0.98883	0.98994	0.98219	0.99493	0.99539	0.99234	P3
#9	78	0.98073	0.9801	0.98983	0.98965	0.9893	0.99549	GLO
#10	48	0.97653	0.98013	0.96938	0.99057	0.99197	0.98753	P3
#11	50	0.98684	0.97144	0.98792	0.99247	0.98284	0.99445	GLO
#12	41	0.9914	0.96647	0.98944	0.99473	0.97907	0.99521	*
#13	86	0.99426	0.97187	0.99101	0.99709	0.98388	0.99627	GEV
#14	52	0.98132	NA	0.97269	0.99247	NA	0.98861	GEV
#15	24	0.95023	0.97909	0.94225	0.97953	0.99168	0.97576	P3
#16	42	0.97923	0.97641	0.98231	0.98922	0.98713	0.99207	GLO
#17	20	0.94534	NA	0.93812	0.97752	NA	0.97403	GEV
#18	40	0.97626	NA	0.97174	0.99075	NA	0.98878	GEV

CONCLUSIONS

- To discriminate between the KAP and GP models, use of the AD statistic leads to bias for one model over the other.
- The use of RML for discriminating between three-parameter distributions led to some serious numerical problems.
- The TN.SW and TN.PPCC statistics proved to be the most advantageous among those considered, they would be recommendable in practice for this reason.
- We found a difficulty in discriminating between the P3 and GEV models.