Responses to Reviewer #1 on "Does Nonstationarity in Rainfall Requires Nonstationary Intensity-Duration-Frequency Curves? By Poulomi Ganguli and Paulin Coulibaly

We thank Referee #1 for reviewing our manuscript and providing constructive feedback, which improves the quality of the manuscript. Our responses are embedded within the comments (in BLACK) in BLUE.

The manuscript presents an interesting topic, and discuss the crucial question of whether there is enough evidence of changes in hydrometric series to warrant a change in the IDF curves used for the design and maintenance of hydraulic structures. Although the topic discussed is interesting and worthy, the paper is quite inconclusive and does not manage, in my opinion, to provide a clear point of view on the matter. The authors have definitely done a lot of work and have looked very carefully at the data, but they fail to summarize their finding in any useful way and simply provide a lot (too much maybe) of information. The presentation of the methods and results is quite unclear and it has several opaque points**.** The statistical methods are often presented with some imperfections and in general the paper could greatly benefit from some proof-reading and re-organisation. In particular the authors should make more of an attempt to summarise their findings from all the non-parametric tests in a way that is more informative.

Response: Thanks for the feedback. The reviewer comments are well appreciated. In our case, a series of statistical tests are necessary to assess nonstationarity in design rainfall, as echoed in earlier literature (Sadri et al., 2016; Yilmaz et al., 2014, 2017). A single statistical test may not be reliable enough to detect signatures of nonstationarity in hydrometeorological time series. Further, we note that multiple tests allow a more rigorous assessment of overall trend in the time series since certain tests are complimentary to each other. Therefore, we explored various statistical tests, starting from testing auto-correlation, presence of monotonic (using trend tests) or abrupt change (using single point change detection algorithm) at different statistical significance levels in practice. Further, we have presented a flowchart of complete methodology in Figure 2 to comprehend the overall analysis. Now coming to statistical methods, we have significantly revised the manuscript to correct any miss perfections as pointed by the reviewer. While we are highly appreciative of the suggestions and comments by the reviewers, we do have one minor point to make which may come across as a slight disagreement with one set of comments. We sense a sentiment shared in one of the comments that our presentation of methods and results are quite unclear. We do not agree with this sentiment even though we agree that the various nuances were not clearly explained in the previous version of the manuscript. Since the focus of the work is insight driven, we have discussed methodologies thoroughly in Supplements to avoid distraction of audience by over-emphasizing the methodologies.

However, we have attempted to improve the presentation of methods and re-organized our manuscript in light of the reviewer's comments. As suggested we have made the following changes in the revised manuscript:

- We have expanded Section 3.2 in Methods to include rationale for the inclusion of multiple tests for detecting nonstationarity. We argue that some of the tests are complimentary to each other. Further, multiple tests allows a robust assessment of overall trend, shifts and nonstationarity in the time series as suggested in the literature (Sadri et al., 2016; Yilmaz et al., 2014, 2017).
- We have reorganized Section 3.3 to include mathematical formulations of GEV distribution and associated time varying covariates to model nonstationary GEV parameters.
- We have re-written the Methodology section and re-organized the Supplements into different sections to present it in a more coherent and clearer way to the readers.
- We have summarized the results of trend detection tests in detail in Page 12, line 17 – 29.
- We have included Bayes-factor criterion in addition to AIC statistics for small sample to evaluate fit of the nonstationary model.
- We have restricted our analysis to Bayesian fit for stationary and nonstationary model.
- We have recalculated 95% credible intervals for all sites from 0.025 and 0.975 quantiles of the simulated posterior samples.

The title of the manuscript indicate that IDF curves are the main topic, although the authors limit themselves to the (hard) task of fitting different frequency curves to the each series with different duration separately. This could result in non-consistent estimates eventually. The type of studies the authors perform is laudable and would be the first step to take to assess whether new IDF curves would need to be derived.

Response: Here we slightly disagree with the reviewer. First, we fitted both stationary and nonstationary frequency curves corresponding standard durations, commonly used in practice for infrastructure in design. We also test the hypothesis whether we need nonstationary frequency curves for the moderately and densely populated urbanized locations across Southern Ontario. We discussed motivation of our study in detail and extended literature review in the revision. Next, we compared the design storm estimates using simple z-statistics considering range of uncertainty as assessed by 95% credible interval to find out whether statistically significant differences exist between nonstationary versus stationary method. Finally, we presented updated IDF curves for all nine locations across Southern Ontario, which is of interest to stakeholders' of the region. We further compared updated versus EC-generated IDFs considering both nonstationary and stationary (Figures 6 and 7) conditions.

The authors do a lot (a lot!) of tests to the data series of each duration - definitely the issue of multiple comparisons arise and it is to be expected that some tests will turn out to be significant just by randomness.

Response: We appreciate the reviewer's point. However, we would stress that multiple tests are needed to detect presence of monotonic trends or abrupt shifts, and nonstationarity in the time series since a selected or cherry-picked number of tests may not be sufficient to detect plausible changes and nonstationarity in the time series. Multiple tests were also performed in earlier studies (Sadri et al., 2016; Yilmaz et al., 2014, 2017) to detect temporal changes in the time series. For example, we employ both Mann-Whitney and Pettitt method to find abrupt shift in mean in the time series, whereas Mann-Kendall test was employed to detect monotonic trend in the time series.

Previous studies (Xie et al., 2014; Yue and Wang, 2002) have found that the rank-based nonparametric Mann-Whitney test is not really distribution free and the power of test is often affected by the properties of sampled data. In practice, when real change point is unknown, often Mann-Whitney test in general does not work well, and the Pettitt method can yields plausible change point location along with its statistical significance. However, significance of the Pettitt test can be obtained using an approximated limiting distribution. As shown earlier, the p-value associated with the test statistics is evaluated following an approximate estimate (Xie et al., 2014). Further, it is also important to note that presence of nonstationarity may not be evaluated merely on the basis of trends or abrupt shifts in the time series, even if the increasing or decreasing trends are statistically significant (Yilmaz et al., 2014). Therefore, we also employed three statistical tests, namely Augmented Dickey-Fuller (ADF), Kwiatkowski–Phillips–Schmidt–Shin (KPSS) and Priestley Subbarao (PSR) test to further investigate nonstationarity in the time series. Both ADF and KPSS tests are based on autoregressive nature of time series. However, Yilmaz et al. (2014) did not observe presence of any significant nonstationarity in short-duration extreme rainfall time series in the city of Melbourne even after employing these tests. Therefore as an alternative, we employed frequency-based PSR test, which is able to capture nonlinear dynamical nature of hydrological system than the former two tests (Ali and Mishra, 2017; Hamed and Rao, 1999). We have incorporated these points in the revised version of the manuscript in appropriate places (Page 8, lines 20-24; Page 9, lines 2 - 6).

I have to say it is difficult to follow the authors in all their testing, there is very little effort made to summarise the finding in any useful way and the results are simply presented/dumped as they are in the SI.

Response: We agreed. In the revised manuscript we provided a more detailed description of the results:

- In page 11, lines 21-26, we provided results of skewness and kurtosis in Annual Maxima (AM) time series. We move results of skewness and kurtosis analysis in the form of Tables (Tables 2 and 3) in main manuscript. We have added following sentences:

  "The skewness is a measure of the asymmetry in the AMP distribution. Positive values of skewness indicate that data are skewed to the right. The skewness of sub-hourly precipitation extremes varies between 0.22 and 4.45, with highest being 30-min AMP record at Hamilton and least being at Oshawa respectively (Table 2). Likewise, for hourly extremes, the skewness ranges between 0.54 and 2.54, with least being 1-hour AMP at Oshawa and highest is 1-hour AMP at Hamilton respectively (Table 3)."

- In page 12, lines 17 – 29, we summarized results of nonstationary trend detection tests. We have added following sentences in the revised manuscript:

  "We find statistically significant monotonic increase and abrupt step changes, both in mean and variance in Oshawa and Trenton respectively (Table S6 and S10), whereas London show (significant) decrease (Table S9) from duration of 6-hour and more. Windsor, Kingston and Stratford show (significant) step changes as confirmed by Mann-Whitney and Mood Tests (Tables S7, S8 and S11). On the other hand, Toronto, Hamilton and Fergus Shand Dam (Tables S4, 4.1; S5, 5.1; S12) do not exhibit any statistically significant gradual or abrupt changes in the AMP time series. The ADF tests show presence of nonstationarity in all durations across the sites. To further validate results of ADF test, KPSS and PSR tests are employed. The KPSS test detects presence of nonstationarity at 3 out of 9 sites for 24-hour rainfall extreme at 5% significance level, whereas the results of PSR test indicate nonstationarity across 5 sites in 24-hour rainfall extremes. While KPSS test alone could not detect presence of nonstationarity in any of the extreme series in Oshawa and Stratford respectively, the results of PSR test did not indicate nonstationarity in any of the short-duration rainfall extreme in Windsor. Both of these tests taken together detect presence of nonstationarity in rainfall extremes across 6 out of 9 sites".

- We have incorporated results of the nonstationary versus stationary model fit of selected airport sites, such as, Toronto, Hamilton, Windsor and London in Tables 4 – 7 in the main manuscript and explained the results in page 13, lines 9 – 14.

- We have revised the result section to include more thorough explanation of each of the findings.

I am not sure whether the results are reliable given the authors have p-values larger than 1.

Response: Here we briefly explain computation procedure of Pettitt Test (Xie et al., 2014), which we have appended in Supplements (SI 2).

When a sequence of random variables is divided into two segments represented by $x_1,...,x_{t_0}$ and $x_{t_0+1}, x_{t_0+2},...,x_{t_0}$, if each segment has distribution functions, $F_1(x)$ and $F_2(x)$, where $F_1(x) \neq F_2(x)$, then change point is identified at t_0. Thus the null hypothesis of the test is "no change", $H_0 : \tau = T$ against the alternative of "change" $H_1 : 1 \leq \tau < T$. The test is based on following statistic (Serinaldi and Kilsby, 2016; Xie et al., 2014)

$K_T = \max_{1 \leq t \leq T} |U_{t,T}|$, where $U_{t,T} = \sum_{i=1}^{t} \sum_{j=i+1}^{T} \text{sgn}(X_i - X_j)$

Where $\text{sgn}(x) = 1$, if $x > 0$, 0 if $x = 0$ and -1 if $x < 0$. The p-value associated with K_T is approximately evaluated as (Xie et al., 2014), $p = 2\exp\left(\frac{-6K_T^2}{T^2 + T^3}\right)$. Given a certain significance level α, if $p < \alpha$, we reject the null hypothesis and conclude that x_τ is a significant change point at level α. Since the associated *p-value* is computed following an approximate estimate of *p-value*, in few cases it exceeds the value 1, which we sense is due to analytical intractability of the estimate. In that case, we have kept the table value blank simply putting a hyphen, and added a footnote indicating the calculation of *p-value* is analytically intractable in those cases.

Response to Further Remarks by Reviewer 1

Comment 1 The beginning of Section 3.3 is very messy and should be rewritten. Distributions do not contain parameters, they are characterised by parameters. Line 25, " a value of the shape parameter equal to zero". Line 28: "In the case of a negative shape parameter, the distribution is a Weibull". Note that the Frechet is also a bounded distribution, except it has a lower bound. Overall I would write down the whole thing in a formula, specifying the limits of the distribution for the different values of the shape.

Response: Agreed. We have added following sentences in the revision:

"The GEV distribution is characterized by three parameters, the location, the scale and the shape of the distribution, which describes the center of the distribution, the deviation around the mean

and the shape or the tail of the distribution (Katz et al., 2002; Katz and Brown, 1992). The cumulative distribution function of stationary (time invariant) GEV model is given by (Coles et al., 2001):

$$G(z)=\exp\left\{-\left[1+\zeta\left(\frac{z-\mu}{\sigma}\right)\right]_{+}^{-1/\zeta}\right\} \qquad \sigma>0,-\infty<\mu,\zeta<\infty \tag{3.1}$$

Where, $y_{+}=\max\{y,0\}$, and

$z\in\left[(\mu-\sigma)/\zeta,+\infty\right)$ when $\zeta>0$; $z\in\left(-\infty,(\mu-\sigma)/\zeta\right]$ when $\zeta<0$; and $z\in(-\infty,+\infty)$ when $\zeta=0$

μ is a location parameter, σ is a scale parameter and ζ is a shape parameter determining the heaviness of the tail. The shape parameter ζ, determines the higher moments of the density function and also the skew in the probability mass. The '+' sign indicates positive part of the argument. The Eq. (3.1) encompasses three types of DFs based on the sign of the shape parameter, ζ: (i) the Fréchet, with a finite lower bound of $(\mu-\sigma)/\zeta$ and an unbounded, heavy upper tail, ($\zeta>0$), (ii) the Weibull, unbounded below and with a finite upper bound of $(\mu-\sigma)/\zeta$, ($\zeta<0$) and (iii) the Gumbel, unbounded below and above with a light upper tail $\zeta=0$, formally obtained by taking limit as $\zeta\to 0$. The Gumbel distribution is described by an unbounded light tailed distribution and the tail decreases rapidly following an exponential decay. The Fréchet distribution is a heavy-tailed distribution, and the tail drops relatively slowly following a polynomial decay (Towler et al., 2010). On the other hand, the Weibull distribution is a bounded distribution".

Comment 2 Page 2 line 13. It is often the case though that IDF curves are derived not only from at-site data but using a pooled set of stations see for Svensson and Jones (2010, doi:10.1111/j.1753-318X.2010.01079.x) for a review of methods used in several countries.

Response: Agreed. The approach can be implemented locally (at Site; or SFA) or regionally (RFA or pooled). The regional frequency analysis is used when available record length are short or at locations where no observed data are available (Castellarin et al., 2012; Komi et al., 2016). However, various RFA estimation methods have certain drawbacks, such as Index flood method is sensitive to the homogeneity assumption and formation of regions; in Bayesian method of regionalization, the prior distributions of parameters are often not precise enough and do not add precision to the estimates; in Hierarchical approach, the method may produce abrupt changes in

the parameters from one site to another. Komi et al. (2016) summarizes the limitations and advantages of some of the widely used RFA techniques. In our case the available records across all sites ranges between 47 and 66 years, which are more than the climatology (often over time periods of 30-years) of a region. Hence, we employ SFA method in our study. The rationale of incorporating at-site frequency method to derive IDF curves in the present study is discussed briefly in page 3, lines 17 – 25. This also allows a consistent comparison with the EC-IDFs that have been used in practice in the study area.

Comment 3 Page 3 - line 8-9: the authors seem to imply that the Gumbel distribution is symmetric - which is not the case, as it is easy to see by plotting the pdf of a Gumbel distribution.

Response: We agree. This was a mistake. We revise the sentence as follows:

EV1 distribution has certain limitations, such that it is a non-heavy tailed distribution and characterized by constant skewness and kurtosis coefficients.

Comment 4 Section 3.1: I think the information of the percentage of missing values of each station/duration should be given somewhere - ideally in the main text and not in SI. I can not judge whether the MCR technique is the most appropriate one, as this is too far away from my area of expertise.

Response: We agree. We have moved Table S1 from Supplement to main manuscript as Table 1. We have also added an extra column in Table 1 indicating information of missing years and durations at each station.

Comment 5 Page 8 - line 8: if the 5% and 95% quantile of the posterior samples are taken then a 90% credibility interval is constructed. A 95% interval is taken to be one that contains 95% of the distribution.

Response: Agreed! As suggested we have re-analyzed our data to incorporate 2.5% and 97.5% quantiles of the posterior sample to construct a 95% credible interval.

Comment 6 Section 3.3: it is not clear to me why the authors go through the trouble of fitting both an ML and Bayesian fit for the stationary model if they only use a Bayesian model for the non-stationary models. Just use the Bayesian methods and embrace Bayesian Inference.

Response: We appreciate the reviewer's comment. As suggested, we have presented the results only using Bayesian model and exclude ML method.

Also, seeing in Table SI16-S24 that the more complex non-stationary model GEVII is often selected I wander whether the authors have tried to only fit models with the scale taken as the only varying function?

Response: We have revised our results in light of the above comments. However, from revised set of results we noted that in a few cases GEV II model (nonstationary in location and scale parameter), performed better than GEV I model (nonstationary in location only). The above results are not uncommon given the highly nonstationary nature of precipitation extremes as observed from the Figure 3. Similar findings were also noted by (Gu et al., 2017) in a flood frequency analysis of Pearl River basin in China, where the author have analyzed 28 stream gauge locations. The results of their analysis suggested in 5 out of 28 sites GEVII performed better as compared to the stationary and GEV I models.

Lastly, why not to formally test stationary/non-stationary model is better by using a Bayesian factor or some pre-set rule on the 95% credibility interval not-containing zero?

Response: Agreed! We have incorporated Bayes factor, AIC statistics for small sample and probability-probability (*P-P*) plot to evaluate model fit.

Comment 7 Section 3.3: what do you do with the results of the Pettitt test? One could use it to build a model with a step-change rather than a continuous function of time. In general, why doing all the non-parametric test AND the parametric models? What is the use of the non-parametric tests exactly?

Response: This is indeed a good point raised by the reviewer. Here, we used three different tests, Pettitt, Mann-Whitney and Mood tests to identify abrupt step changes in the time series, which is different from monotonic or gradual trends in the time series. We have implemented a series of statistical tests since a single statistical test may not be able to capture full ranges of nonstationarity in highly nonlinear dynamical system, such as short-duration extreme precipitation. As we discussed earlier, the rank-based nonparametric Mann-Whitney test is not really a distribution free and the power of the test is often affected by the properties of sampled data. In practice, when real change point is unknown, often Mann-Whitney test in general does not work well and the Pettitt method can yield plausible change point location along with its statistical significance. However, the significance of the Pettitt test can be obtained using an approximated limiting distribution. Therefore, above tests were needed in the current setting.

Further, we applied nonparametric tests due to their robustness to non-normality, which usually appears in the hydroclimatic time series. Further, in order to reduce the number of underlying assumption required for testing a hypothesis, such as presence of specific kind of trend or change

point in the data set, nonparametric tests were employed. We discussed each of these issues in the revised manuscript.

Comment 8. Page 8 - line 14: it is very good that the authors verify the goodness of fit by using PP, but it is unclear to me how they "select the model with fewer parameters as the best model when two models have comparable performances.". This is exactly what the AIC should do, so even if the AIC does not indicate that a simpler model should be used the authors might cull a non-stationary model out if the stationary model give a better fit in the PP plot?

Response: We have reanalyzed the data and new results are different from the previous ones.

Comment 9. Page 8 - line 25-26: a positive skewness is just an indication of an asymmetric/skewed distribution, it doesn't necessary indicate a change in the distribution. I mean "extreme values are more frequent in the time series" compared to what?

Response: We have revised this sentence in page 11 (line 22-23) as follows:

Positive values of skewness indicate that data are skewed to the right.

Comment 10. Page 9, line 29: Bayesian measures of uncertainty are normally called credibility and not confidence intervals. Also as I mentioned above - unclear if the 95% or the 90% intervals are derived.

Response: We appreciate reviewer's feedback. As suggested we have replaced the word with credibility interval wherever it is appropriate. We have constructed 95% credibility intervals from the 2.5^{th} and 97.5^{th} percentiles of the simulated posterior samples.

Comment 11. Page 10/Figure 4: how are the DSI calculated for the non-stationary models? Is the last value of the parameters used to compute the quantiles? Why do you show boxplots of the posterior sample and not a 95% credibility interval? As I said I would drop the estimation using ML completely, but if you do use it, you could show confidence intervals based on the delta-method (see Coles, 2001).

Response: We estimated parameters using Bayesian inference (BI) coupled with Differential Evaluation Markov Chain (DE-MC) simulation as in (Cheng and AghaKouchak, 2014; Cheng et al., 2014). DE-MC is an adaptive Monte Carlo Markov Chain (MCMC) algorithm (Ter Braak and Vrugt, 2008; Ter Braak, 2006), in which multiple chains (here, we fix chain length '*n*' as 5) are run in parallel. The resulting MC simulations are then run to an equilibrium (often referred to as the *burn-in* period). It is a standard practice to discard the initial iterations of simulated samples since they are strongly influenced by starting values and do not provide usable

information of the target distribution. Here we run DE-MC simulations for 3000 iterations and kept the 2001-3000th iterations of each chain. The convergence of MC simulation is checked by the "potential scale reduction factor ($\widehat{R}$)" as in (Gelman et al., 2011), which suggests the value of $\hat{R}$ should remain below the threshold value of 1.1. The post burn-in random draws from posterior distribution is then used to construct predictive distributions. For annual maxima time series of each duration, the mean and associated 95% credibility intervals of parameters $(\mu(t), \sigma(t))$ are derived by computing 50th (the median), 2.5th and 97.5th (bounds) percentiles of post *burn-in* random draw (for example, 50th percentile of $\mu(t_1), \ldots., \mu(t_{100})$). The derived model parameters are then used to compute corresponding design rainfall quantiles at *T*-year return period and corresponding credibility interval. We calculated median value of design storm by computing 50th percentiles of the post-burn in simulated posterior quantiles for the nonstationary model. We have constructed 95% credibility intervals from the 2.5th and 97.5th percentiles of the posterior samples. In the manuscript, the boxplots are shown for 95% credibility interval and not with posterior samples. To avoid further ambiguity we have revised corresponding figure caption (Figure 4) as, "DSI estimates of median (horizontal line within the box plot) and 95% credible intervals for 100-year return periods of stationary versus nonstationary models across nine sites (a - i). The boxplots indicate the uncertainty in estimated DSI using Bayesian inference". As suggested we have dropped ML method completely in the revised manuscript.

Comment 12. Page 11/Figure 6: has any assessment been done on whether the stationary version of the fitted curves has a good overlap to the EC-curves? Surely if these two curves are very different, any mis-match between the non-stationary results and the EC-curves could be due to the fact that the EC curve doesn't fully fit the data of a site. This links to a comment on the statements in page 13 between line 20-25: you are saying that from the comparison of stationary to non-stationary models there seems to be no indication of a need to update DSI, but when comparing the outputs of a non-stationary model to the EC-curves (obtained assuming stationarity) then the evidence is that we should update the DSI. This points in the direction of the EC-curves being different from the at-site stationary curves.

Response: As suggested we have compared stationary version of the fitted curve with EC curves. Associated results are presented in Figures 8 and S15. We discuss following results in the revision:

"In order to distinguish between stationary and nonstationary method of analysis, we also present updated IDF assuming stationary condition relative to EC IDF in the same plot (in top panel). The comparisons of remaining sites are presented in Figure S15. Thus we made the first attempt to compare the results of updated versus EC-generated IDFs considering both nonstationary and

stationary conditions, which are part of contemporary Design Standards and widely used by the stakeholders and practitioners. Overall, the updated IDFs closely follow the pattern of trends analogous to EC-generated IDFs, except for the 100-year return period. The difference being more pronounced considering nonstationary condition, especially at Toronto International Airport (Figure 8), Oshawa WPCP and Stratford WWTP (Figure S15). At longer durations and higher return periods, stations in metropolitan areas (such as Toronto International Airport, Hamilton Airport, Oshawa WPCP and Windsor Airport) show large differences in DSIs, whereas moderately populated locations such as, Kingston P. station and Fergus Shand dam show relatively smaller changes. Considering, nonstationary condition, the maximum increase in Furgas Shand dam is noted as 18.7% for the 2-hour storm duration and 100-year return period, whereas an increase of around 44.5% is shown for 12-hour storm duration at Toronto Airport".

Comment 13. Page 14: I don't understand what the last sentence of the paper means.

Response: We have revised the sentence as follows:

"Given that these findings are for the current period (e.g. historical extreme rainfall time series), we recommend a careful extrapolation of the findings with regards to future climate projections, in which frequency and magnitude of extreme rainfall are expected to intensify (Mailhot et al., 2012; Deng et al., 2016; Fischer and Knutti, 2016; Prein et al., 2016; Pfahl et al., 2017)". Further work should consider nonstationary methods for deriving future IDFs in Southern Ontario.

Comment 14. SI3: I would give the lower and upper bound of the GEV in a formula to give a simpler indication of the effect of the value of the shape parameter.

Response: Agreed, we add following expressions to indicate effect of shape parameter in GEV distribution:

$$G(z) = \exp\left\{-\left[1+\zeta\left(\frac{z-\mu}{\sigma}\right)\right]_{+}^{-1/\zeta}\right\} \qquad \sigma > 0, -\infty < \mu, \zeta < \infty \tag{3.1}$$

Where, $y_+ = \max\{y, 0\}$, and

$z \in \left[(\mu-\sigma)/\zeta, +\infty\right)$ when $\zeta > 0$; $z \in \left(-\infty, (\mu-\sigma)/\zeta\right]$ when $\zeta < 0$; and $z \in (-\infty, +\infty)$ when $\zeta = 0$

μ is a location parameter, σ is a scale parameter and ζ is a shape parameter determining the heaviness of the tail. The shape parameter ζ, determines the higher moments of the density function and also the skew in the probability mass. The '+' sign indicates positive part of the argument. The Eq. (3.1) encompasses three types of DFs based on the sign of the shape parameter, ζ: (i) the Fréchet, with a finite lower bound of $(\mu-\sigma)/\zeta$ and an unbounded, heavy upper tail, ($\zeta > 0$), (ii) the Weibull, unbounded below and with a finite upper bound of $(\mu-\sigma)/\zeta$, ($\zeta < 0$) and (iii) the Gumbel, unbounded below and above with a light upper tail $\zeta = 0$, formally obtained by taking limit as $\zeta \to 0$.

Comment 15. SI3.1: why using ML in one case and Bayesian methods for another?
Response: Agreed. As suggested we have excluded the results of ML estimate.

Comment 16. SI3.1, paragraph after equation 3.8: p(y|λ, x) does not give information on the parameters. The formulation of the sentence seem to imply that the likelihood p(y|λ, x) gives information on the parameters under non-stationarity, which is not the case.
Response: Agreed. To avoid any ambiguity, we have revised the sentence as:
The posterior distributions, $p(\omega|y)$ and $p\left(y|\boldsymbol{\lambda}, x\right)$ indicate likelihood functions, which infer parameters $\boldsymbol{\omega} = \{\mu, \sigma, \zeta\}$ considering stationarity, and $\boldsymbol{\lambda} = \{\mu_1, \mu_0, \sigma_1, \sigma_0, \zeta\}$ assuming nonstationarity conditions, respectively.

Comment 17. SI4.1 - the definition in eq 4.1 for the Akaike information criterion is not correct (or better it is correct for a normal model, but not for a GEV). AIC is generally defined as AIC = −2log(L(ω, x)) + 2m . That's how the two references cited by the authors define the AIC as well. From what I understand from the explanation of the observed/expected values the authors are doing a model selection using AIC based on the quantiles, which is not made explicit in section 3.3. If that's the case, which quantiles are used?

Response: Here we cannot concur with the reviewer. We also point to the reviewer that we have used a least square version of Akaike Information Criterion (*AIC*), which is calculated as the largest deviation between the observed (empirical in this case, obtained from rank-based plotting

position formula) and modelled cumulative distribution. This form of *AIC* is widely used in hydrology in general and multivariate statistics in particular (Dawson et al., 2007; Deepthi Rajsekhar et al., 2015; Ganguli and Reddy, 2012; Hu, 2007; Janga Reddy and Ganguli, 2012; Karmakar and Simonovic, 2007, 2009). Further, we point that this form does not correspond to a normal model. For calculation of *AIC* statistics, we consider median of the DE-MC sampled parameters, which can be considered as an average or expected value of risk in the historical observation. We have added this in detail in section 3.3 as suggested by the reviewer.

Comment 18. Equation 5.1 and 5.2, what happens if $\zeta = 0$?

Response: When $\zeta \to 0$, the GEV distribution reduces to Gumbel distribution (or Extreme Value Type I). In that case, the return period is obtained by calculating frequency factor. We add following sentences SI 4, page 40 in the revised version of the manuscript:

"When $\zeta \to 0$, the GEV distribution reduces to Gumbel distribution (or Extreme Value Type I). It should be noted that Gumbel Extreme value distribution has been commonly usto estimate design storm by Environment Canada (CSA, 2010). The Gumbel probability distribution has following form (Wang et al., 2015)

$$q_p = \mu + K_p \sigma$$

Where K_p denotes frequency factor depending on the return period *T*, which is obtained using following relationship (Wang et al., 2015)

$$K_p = \frac{-\sqrt{6}}{\pi}\left[0.5772 + \ln\left(\ln\left(\frac{T}{T-1}\right)\right)\right]$$

Environment Canada uses this method to estimate rainfall frequency at a given duration and obtain nationwide IDF curves".

Comment 19 Table S7, 24-hours, the p-value for the pettitt test is larger than 1 - this cannot be right. (see also S9 30min, S11 15min to 2hr, S14 15min to 2hr, S15 12hr)

Response: Agreed. As explained before, the significance of the Pettitt test can be obtained using an approximated limiting distribution, the p-value of certain durations could not be computed accurately due to analytical intractability. We have kept those places as blank (-) in the revised manuscript. We have added a footnote at the end of Table S4 explaining this point.

Comment 20 Table SI16 - not sure if the red and blue are right in all stations.

Response: We have revised our analysis and revised results are different from earlier.

Comment 21 Pg 14 Supplement : the definition of return level has the word expected in the wrong place. ... often referred as return level in the literature is the expected value to be exceeded on an average once in every... should be ... often referred as return level in the literature, is the value which is expected to be exceeded on an average once in every... - see Coles, 2001 - end of section 3.1.3 (pg 49 in my edition).

Response: Agreed. We have revised the definition in current version as suggested.

Comment 22 I also find some of the Figures - and in particular their captions - could be improved.

Response: Agreed. We have revised captions of the figures wherever appropriate to enhance clarity. By doing so, we have also incorporated changes as suggested by the reviewer.

Comment 22.1 Figure 3 caption

- Durations higher than an hour are also shown I would say "Spatial distribution of trends, change points and non-stationarities in rainfall extremes of several durations in nine urbanized locations, Southern Ontario"
- Drop the information on the population - it's in Figure 2 and in the text (several times)
- Drop the information on the tests performed or at least reduce it since it's given in the text (for example drop the references)
- Include information on the color coding in the legend.
- If tests are performed at 5% and 10% - what is considered statistically significant? p-values < 0. 05 or p-value < 0.1?

Response: Agreed and incorporated in the revision. Further, p-values < 0.1 is considered to be statistically significant. The same has been incorporated in the revision.

Comment 22.2 Figure 4 caption: drop the list of the name of the station - it is given in the plot.

Response: Agreed and incorporated in the revision. Also we have revised the figure caption in light of comment no. 11.

Comment 22.3 Figure 5 caption: add the information on the cyan shading representing the site with significant autocorrelation in the legend and drop from the legend. The second last sentence grammar is not correct.

Response: Agreed and incorporated in the revision. We have revised the grammar of the second last sentence.

Comment 22.4 Figure 7: I would include the information on solid/dotted lines in the legend.

Response: Agreed and incorporated in the revision.

Comment 23 The paper has several grammar mistakes, with articles missing or appearing in the wrong place and several sentences which have non-concordant subject and verb. I list here a minuscule sample of the typos/mistakes I found

Response: We have thoroughly checked the manuscript, corrected all typos. We have revised the manuscript in places as they were suggested.

Comment 23.1 Page 3, line 16 slowly or varying are not antonyms. Line 18-19 does should have a singular subject (not signal). Same in line 25-26.

Response: Agreed. We have revised this to gradual or monotonic changes. We have revised the sentence in line 18-19. We have revised the grammar in line 25-26.

Comment 23.2 Page 3 line 23-24: The structure of the sentence is confusing. It is not the signatures that necessitate IDF. Maybe use "...make necessary the use..."

Response: Agreed and incorporated in the revision.

Comment 23.3 Page 5: line 4-5 more repeated twice.

Response: Agreed and we have revised the sentence as suggested.

Comment 23.4 Page 8: Line 27-28: the sentence is not complete.

Response: We apologized for this. We have corrected all incomplete sentences including this one in the revision.

Comment 23.5 Page 10 - line 16: less uncertainty (not lesser).

Response: Agreed and incorporated in the revision.

Comment 23.6 Page 11 - line 17: More genralLY - and the sentence has a singular subject so line 19 should be is not are.

Response: Agreed and incorporated in the revision.

Comment 23.7 Page 12 - line 2: smaller, not lesser.

Response: Agreed and incorporated in the revision.

Comment 23.8 Page 12 - line 17: It? I think you need a "We"?

Response: Agreed and incorporated in the revision.

Comment 23.9 Page 13 - line 6: does/is?

Response: Agreed and incorporated in the revision.

Comment 23.10 Page 13 - line 12: several studies HAVE.

Response: Agreed and incorporated in the revision.

Comment 24. Further inconsistencies I identified:

- **Comment 24.1** Page 4 Line 10: the ref to Jien and Gough is missing in the reference list and I think is not needed since it states a basic fact about the geography of Canada.
  Response: Agreed and the citation is excluded from the revised version.

- **Comment 24.2** Page 9, Line 28 - ξ, instead of ζ used in the SI, for the shape parameter of the GEV.
  Response: Agreed and incorporated in the revised version of the manuscript.

- **Comment 24.3** Reference list: Cheng, L. and AghaKouchak, A. 2014 - just give the doi, not the ncbi link.
  Response: Agreed and incorporated in the revision.

- **Comment 24.3** Supplement references: Coles and Tawn (1996) cited in text missing in the ref. Anyway, for that formula Coles, 2001 is probably enough as a citation.
  Response: The citation Coles and Tawn (1996) is included in the revised version.

- **Comment 24.4** The citation to Coles 2001, An introduction to statistical modelling of extreme values, Springer in the supplementary material is wrong, as it has additional authors other than Coles.
  Response: Agreed and incorporated in the revision.

References

Ali, H. and Mishra, V.: Contrasting response of rainfall extremes to increase in surface air and dewpoint temperatures at urban locations in India, Sci. Rep., 7(1), 1228, doi:10.1038/s41598-017-01306-1, 2017.

Castellarin, A., Kohnová, S., Gaál, L., Fleig, A., Salinas, J. L., Toumazis, A., Kjeldsen, T. R. and Macdonald, N.: Review of applied-statistical methods for flood-frequency analysis in Europe, Available from: http://nora.nerc.ac.uk/19286/, 2012.

Cheng, L. and AghaKouchak, A.: Nonstationary precipitation intensity-duration-frequency curves for infrastructure design in a changing climate, Sci. Rep., 4, doi: 10.1038/srep07093, 2014.

Cheng, L., AghaKouchak, A., Gilleland, E. and Katz, R. W.: Non-stationary extreme value analysis in a changing climate, Clim. Change, 127(2), 353–369, 2014.

Coles, S. G. and Tawn, J. A.: A Bayesian Analysis of Extreme Rainfall Data, J. R. Stat. Soc. Ser. C Appl. Stat., 45(4), 463–478, doi:10.2307/2986068, 1996.

Coles, S.: An introduction to statistical modeling of extreme values, Springer, 2001.

CSA (Canadian Standards Association): Technical Guide – Development, Interpretation and Use of Rainfall Intensity-duration-frequency (IDF) Information: Guideline for Canadian Water Resources Practitioners, 2010.

Dawson, C. W., Abrahart, R. J. and See, L. M.: HydroTest: A web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts, Environ. Model. Softw., 22(7), 1034–1052, doi:10.1016/j.envsoft.2006.06.008, 2007.

Deepthi Rajsekhar, Vijay P. Singh and Ashok K. Mishra: Hydrologic Drought Atlas for Texas, J. Hydrol. Eng., 20(7), doi:10.1061/(ASCE)HE.1943-5584.0001074, 2015.

Deng, Z., Qiu, X., Liu, J., Madras, N., Wang, X. and Zhu, H.: Trend in frequency of extreme precipitation events over Ontario from ensembles of multiple GCMs, Clim. Dyn., 46(9–10), 2909–2921, 2016.

Fischer, E. M. and Knutti, R.: Observed heavy precipitation increase confirms theory and early models, Nat. Clim. Change, 6(11), 986–991, 2016.

Ganguli, P. and Reddy, M. J.: Probabilistic assessment of flood risks using trivariate copulas, Theor. Appl. Climatol., 111(1–2), 341–360, doi:10.1007/s00704-012-0664-4, 2012.

Gelman, A., Shirley, K. and others: Inference from simulations and monitoring convergence, Handb. Markov Chain Monte Carlo, 163–174, 2011.

Gu, X., Zhang, Q., Singh, V. P., Xiao, M. and Cheng, J.: Nonstationarity-based evaluation of flood risk in the Pearl River basin: changing patterns, causes and implications, Hydrol. Sci. J., 62(2), 246–258, 2017.

Hamed, K. H. and Rao, A. R.: A modified Mann-Kendall trend test for autocorrelated data, J. Hydrol., 204(1), 182–196, 1998.

Hu, S.: Akaike information criterion, Cent. Res. Sci. Comput., North Carolina State University. Available from: http://www4.ncsu.edu/~shu3/Presentation/AIC_2012.pdf, 2007.

Janga Reddy, M. and Ganguli, P.: Application of copulas for derivation of drought severity–duration–frequency curves, Hydrol. Process., 26(11), 1672–1685, doi:10.1002/hyp.8287, 2012.

Karmakar, S. and Simonovic, S. p.: Bivariate flood frequency analysis. Part 2: a copula-based approach with mixed marginal distributions, J. Flood Risk Manag., 2(1), 32–44, doi:10.1111/j.1753-318X.2009.01020.x, 2009.

Karmakar, S. and Simonovic, S.: Flood Frequency Analysis Using Copula with Mixed Marginal Distributions, Water Resour. Res. Rep. Available from: http://ir.lib.uwo.ca/wrrr/19, 2007.

Katz, R. W. and Brown, B. G.: Extreme events in a changing climate: variability is more important than averages, Clim. Change, 21(3), 289–302, 1992.

Katz, R. W., Parlange, M. B. and Naveau, P.: Statistics of extremes in hydrology, Adv. Water Resour., 25(8), 1287–1304, 2002.

Komi, K., Amisigo, B. A., Diekkrüger, B. and Hountondji, F. C.: Regional Flood Frequency Analysis in the Volta River Basin, West Africa, Hydrology, 3(1), 5, 2016.

Mailhot, A., Duchesne, S., Caya, D. and Talbot, G.: Assessment of future change in intensity–duration–frequency (IDF) curves for Southern Quebec using the Canadian Regional Climate Model (CRCM), J. Hydrol., 347(1), 197–210, 2007.

Prein, A. F., Rasmussen, R. M., Ikeda, K., Liu, C., Clark, M. P. and Holland, G. J.: The future intensification of hourly precipitation extremes, Nat. Clim. Change, advance online publication, doi:10.1038/nclimate3168, 2016.

Sadri, S., Kam, J. and Sheffield, J.: Nonstationarity of low flows and their timing in the eastern United States, Hydrol Earth Syst Sci, 20(2), 633–649, 2016.

Serinaldi, F. and Kilsby, C. G.: Stationarity is undead: Uncertainty dominates the distribution of extremes, Adv. Water Resour., 77, 17–36, 2015.

Ter Braak, C. J. and Vrugt, J. A.: Differential evolution Markov chain with snooker updater and fewer chains, Stat. Comput., 18(4), 435–446, 2008.

Ter Braak, C. J.: A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces, Stat. Comput., 16(3), 239–249, 2006.

Wang, X., Huang, G., Liu, J., Li, Z. and Zhao, S.: Ensemble projections of regional climatic changes over Ontario, Canada, J. Clim., 28(18), 7327–7346, 2015.

Wang, X., Huang, G., Liu, J., Li, Z. and Zhao, S.: Ensemble projections of regional climatic changes over Ontario, Canada, J. Clim., 28(18), 7327–7346, 2015.

Xie, H., Li, D. and Xiong, L.: Exploring the ability of the Pettitt method for detecting change point by Monte Carlo simulation, Stoch. Environ. Res. Risk Assess., 28(7), 1643–1655, 2014.

Yilmaz, A. G., Hossain, I. and Perera, B. J. C.: Effect of climate change and variability on extreme rainfall intensity–frequency–duration relationships: a case study of Melbourne, Hydrol. Earth Syst. Sci., 18(10), 4065–4076, 2014.

Yilmaz, A. G., Imteaz, M. A. and Perera, B. J. C.: Investigation of non-stationarity of extreme rainfalls and spatial variability of rainfall intensity–frequency–duration relationships: a case study of Victoria, Australia, Int. J. Climatol., 37(1), 430–442, doi:10.1002/joc.4716, 2017.

Yue, S. and Wang, C. Y.: Power of the Mann–Whitney test for detecting a shift in median or mean of hydro-meteorological data, Stoch. Environ. Res. Risk Assess., 16(4), 307–323, 2002.