Reply to Anonymous Referee #2

September 11, 2015

Referee's comment I dont think I can teach other scientists how to write a paper, but to me the technical language, even if mostly appropriate, is not clear and lacks definitions, explanations on the meaning of the statistics, information on what the author is trying to do in the different chapters and why. This undermines the readability of the work.

Authors' reply: We are sorry to read that because we tried to make the paper as clear as possible. We'll do our best to account your comments in next draft (see below for details).

Referee's comment The methods that the authors are using are mostly taken from the literature, but with some modification. I suggest to highlight what are the novelties also in the methodology introduced by this paper and discuss it.

Authors' reply: MEWP model is not new. It was previously proposed in [Garavaglia et al., 2010]. However, it has never been extensively compared to other methods. So the main novelty of this article doesn't rely in the employed theory but in the fact that we make for the first time a global and extensive evaluation of MEWP. To do this, we make comparisons with other methods relying on EVT with different levels of complexity (with or without seasons, with or without weather subsampling, ...). Also, for the first time we introduce the extension of MEWP for heavy-tailed distributions, the MGPWP model and compare results of MEWP and MGPWP. We thank the reviewer for highlighting this point. We will make it clearer in the introduction of the next version of the article.

Referee's comment I think that the last part, where you observe a trend applying the statistics to two different periods, is out of place in the current paper. It can be an interesting observation, but it is not treated appropriately (as you state in your conclusions "should be taken as a motivation for such an analysis of trends") and I feel like it is not very much related with the rest of the analysis. I think the paper is much more coherent without this section or, if you want to keep it, you should point out how it is related with the rest of the analysis.

Authors' reply: We agree this is slightly off topic of this paper. However there has been very few analysis of trend in extreme rainfall in Norway, so this is a ”cheap” analysis but giving already new results. As noted, it motivates future works. We would be tempted to keep the section but we let the editor decide whether it should be deleted or not.

Referee’s comment Right now I think that the conclusions are poor. In section 4 you present the results, and then without discussion you move to the conclusion, which are very short. You simply state what you have observed in the results and then described the trend (which I think should be removed). I think you should discuss why the model performs better with subsampling by season and weather pattern, the relation with literature (is it the same result observed in other regions? Was it expected?).

Authors’ reply: The reason why the model performs better with subsampling by seasons and WP is that by subsampling we make groups of rainfall that are more alike (in statistics we say there are better identically distributed). It is thus not surprising that similar rainfall values are more easily fitted to a given distribution than those that represent different parent populations. The seasonal and WP influence on rainfall can easily be seen by computing empirical statistics. This is shown in Figure 1 for one of the Norwegian stations randomly chosen for illustration. The top row of the figure clearly shows monthly/seasonal patterns of rainfall (left), as well as the influence of WPs (right). The results of our analysis show that, at least at regional scale, it is better to split data into both seasons and WPs, giving the 16 subclasses shown in the bottom row (left: the 8 classes of the season-not-at-risk; right: the 8 classes of the season-at-risk). This study is the first extensive evaluation of MEWP, so its better performance with regards to more usual methods without subsampling has not been observed in other region of the world, or at least not in published articles. However, as mentioned p. 3545 l. 24, MEWP has already been applied successfully in other regions of the world (in France, Austria, Canada and Norway, see e.g. [Brigode et al., 2014]) so we intuite that the same conclusion should apply in these regions if the same analysis would be lead.

Referee’s comment What part of this analysis do you think could be generalized and where do you think these results do not hold?

Authors’ reply: The modeling in itself could be, in principle, applied in any region of the world (and it has already been used in France, West Canada and Austria, see [Brigode et al., 2014]). The only preliminary requirement is to build a classification of days into WPs derived from an analysis of extreme rainfall in relation to atmospheric circulation patterns.

Referee’s comment Does the use of central rainfall compromise the comparison with other cases in literature?

Authors’ reply: Since all yearly maxima are central rainfall, the comparison with any method for annual maxima can be done without restriction.

Referee’s comment Could a similar analysis apply to floods?

Authors’ reply: A WP sampling approach is not as relevant for extreme flood estimation because several meteorological and hydrological processes are

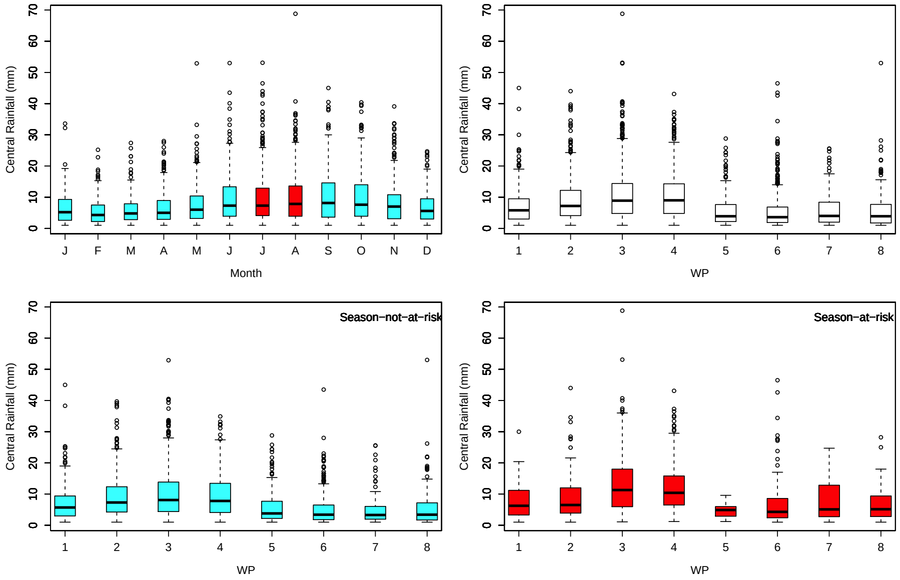

Figure 1: Boxplot of central rainfall for one of the Norwegian stations. Top left: when data are split into months. Red months correspond to the season-at-risk. Top right: when data are split into WPs. Bottom: when data are split into both season and WPs (left: for the season-not-at-risk, right: for the season-at-risk).

involved in flood generation (extreme rainfall of course, but also antecedent rainfall, soil saturation, snowmelt etc.) and their complex interaction can not be summarized (or illustrated) by the WP of the day of the flood only, or in other words, the WP sampling will provide very few benefits in terms of sample homogeneity for a statistical modeling. In the SCHADEX method (Paquet 2013), the WP analysis is applied for extreme rainfall estimation, and is completed by a rainfall-runoff stochastic simulation for extreme flood estimation.

Referee's comment P3551 L2: what is q+?
Authors' reply: q_α^+ is defined p. 3549 l.22; it is the largest threshold: $q_\alpha^+ = \max_{s,k} q_{\alpha,s,k}$.

Referee's comment P3551 L9: I think the reason behind the division of data in two subsamples is not explained clearly. Also I suggest adding an explanation on what you want to do in this paragraph and how.

Authors' reply: Here, as mentioned, we follow the split sample evaluation of [Garavaglia et al., 2011] and [Renard et al., 2013] where more details may be found. Splitting the data to evaluate a model is actually quite common. The idea is to fit the model on two separate samples, and then i) compare how alike the two fits are (SPAN), ii) assess how well the second sample (validation data) is fitted by the model estimated on the first sample (calibration data) (FF and N_T). In other words, as written p. 3551 l. 9-11, "our goal is to test the consistency between validation data and predictions of the estimates, and the accuracy and stability of the estimates when calibration data change."

Referee's comment P3551 L18: missing number of the equation
Authors' reply: Since we do not refer to this equation later on, we don't think the numbering is necessary here.

Referee's comment P3551 L18: is it C1i instead of C2i?
Authors' reply: Absolutely, this is a mistake, it should be $\mathrm{C}_i^{(1)}$.

Referee's comment P3552: the explanation of FF (lines 9-16) is not clear. Please add a definition and make a clearer description. In general for each of the three statistics I would add at the begin a brief definition, with what exactly they describe.

Authors' reply: As noted p. 3551, the scores FF, N_T and SPAN, are not new. They have been proposed in [Garavaglia et al., 2011] and extensively studied in [Renard et al., 2013] where more details may be found. It's hard to explain FF with words but the idea of FF and N_T is to evaluate goodness-of-fit of the tail of the fitted distribution. For this FF evaluates the probability of occurrence of the maximum observed in the validation sample according to the model fitted with the calibration sample, while N_T counts how many times a prescribed return level (e.g. the 10-year return level) is exceeded by the validation data.

Referee's comment P3553: I understand you develop case 1 and 2 to prove your point, but to make it clearer I suggest trying to connect case 1 and 2 to hypothetical situations in your data, and to explain what is the problem connected with judging the two cases as different.

Authors' reply: Departure of scores from the uniform case is sometimes not easy to interpret. However case 1 corresponds usually to a tendency towards an overestimation of the largest observation, while case 2 corresponds to a tendency towards overfitting the largest observation. With the evaluation based on the CDF, one would tend to prefer a model that overfits than a model that overestimates (area of case 2 $<$ area of case 1). There is actually no reason for this preference, and this is why we prefer the evaluation based on the density which indeed gives area values very similar for cases 1 and 2.

Referee's comment P3554: Again, I think the explanation in this page and begin of pag3555 is very technical, but lacks in clear explanations, examples and definitions.

Authors' reply: Actually our wish here is only to remind the definition of N_T which has already been proposed in [Garavaglia et al., 2011] and [Renard et al., 2013]. Unfortunately we think there is no room in this article for examples since this is not new.

Referee's comment P3556: in your list of exponential model you name 4. Next page you say there are 6 (k = 4 and k = 8). In figure 5 you compare 4, which are different from the 4 you listed before. Please improve the cohesion.

Authors' reply: We apologize for the confusion. There is indeed a mistake here. There are 12 models, not 8. Cases $(S, K) = (1, 4)$ and $(1, 8)$ are missing in Figure 5. That figure should be replaced by Figures 2 and 3 below. We will improve the text as suggested.

Referee's comment P3557: again, the description of the method used to estimate the season at risk is not clear enough to me. I suggest to improve the description at L12 including a schematic explanation of what you are going to present next.

Authors' reply: Rather than a schematic explanation, we propose to make clearer the procedure by replacing p. 3557 l. 14 to p. 3558 l. 5 by:
"In detail, the procedure is as follows:

Step 1 Compute the 12 mean monthly maxima of central rainfall.

Step 2 Set $M = 2$.

Step 3 Compute the mean of these values over moving windows of size M months.

Step 4 Select the M consecutive months corresponding to the highest of these values. These M months define the season-at-risk. The remaining months define the season-not-at-risk.

Step 5 Fit the considered model (e.g. MEWP$(0.5, 2, 8)$) with this seasonal definition.

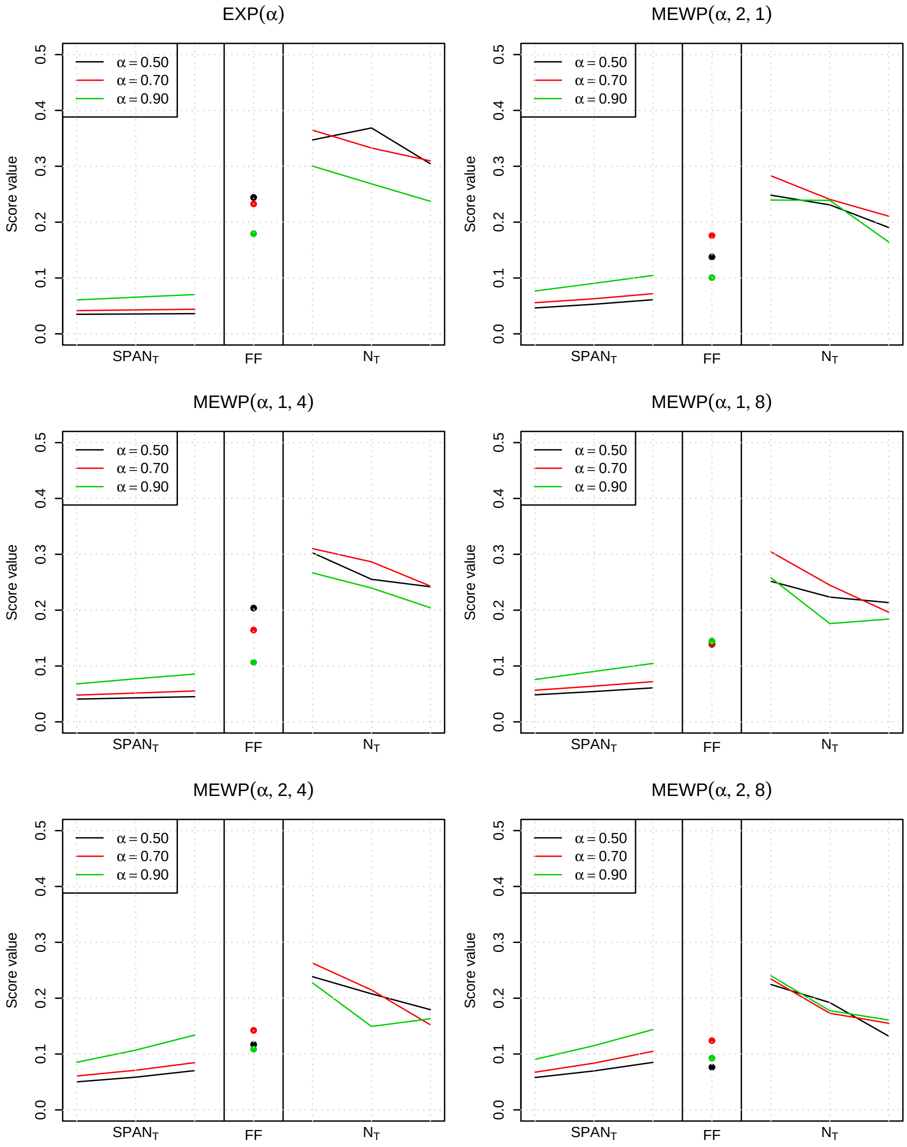

Figure 2: Scores of evaluation for MEWP models, for $\alpha = 0.5$, 0.7 and 0.9. Better scores have values closer to 0. Scores of SPAN$_T$, for $T = 20, 100, 1,000$ year return periods, are the mean scores of (8), while scores of FF and N_T, $T = 5, 10, 20$ years, are based on the density areas (9).

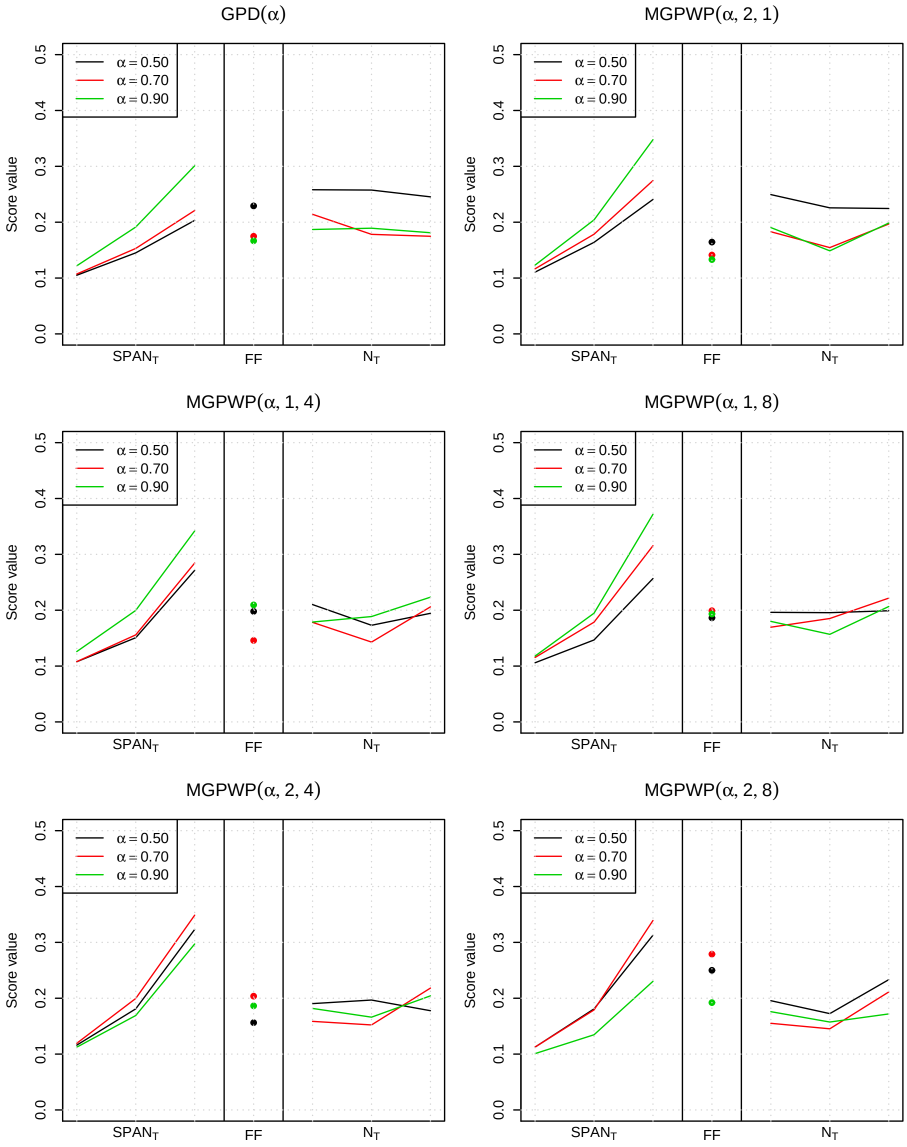

Figure 3: Same as Fig. 2 for MGPWP models.

Step 6 Compare the monthly fits to the monthly empirical distributions. This comparison is made with KGE score (Kling-Gupta efficiency [Gupta et al., 2009]), which if computed, for a given month m, as

$$\mathrm{KGE}_m = \left\{\mathrm{corr}(\tilde{F}_m, \hat{F}_m) - 1\right\}^2 + \left\{\mathrm{std}\left(\frac{\tilde{F}_m}{\hat{F}_m}\right) - 1\right\}^2 + \left\{\mathrm{mean}\left(\frac{\tilde{F}_m}{\hat{F}_m}\right) - 1\right\}^2,$$

where $\tilde{F}_m$ and $\hat{F}_m$ are respectively the empirical and fitted distributions of month m. It should be noted that the KGE criterion is not the only score which could be used here, and was not necessarily developed for scoring distributions. However, the final result (i.e. the seasonal split selected) is not particularly sensitive to the score used.

Step 7 Compute a global KGE score as a weighted mean of these 12 KGE scores, with weights proportional to the mean monthly maxima, in order to force the model to have the best fits for the months with the highest risk.

Step 8 Set $M = 3$ and apply steps 3 to 7.

Step 9 Set $M = 4$ and apply steps 3 to 7.

Step 10 Compare the three global KGE scores obtained respectively for $M = 2, 3, 4$. Select the seasonal definition corresponding to the lowest of these scores.

This procedure is applied for each station and each model separately..."

Referee's comment P3559 L6: instead of starting the paragraph with a list of details for the comparison I suggest to explain what are you going to do. (why do you use different T for different statistics? Why do you compare for different alpha? What do you want to show?). Instead of "the closer to zero the better the score", which does not explain the reason why you are using three statistics, it would be better to include a short summary with the meaning of the statistics.

Authors' reply: The reason for using different T in N_T is to evaluate different part of the tail of the distribution. With large T we assess the very tail of the distribution while with small T we assess the bulk of the distribution. Comparing different α allows us to select the right threshold to be used. This is a bias-variance tradeoff as explained p. 3550 l. 11 : the higher the threshold, the better the approximation of the tail (smaller bias), but at the same time, the higher the variance of the estimated parameters because a smaller number of exceedances are available. Finally, we use three statistics to get a full evaluation of the fits. The statistics are complementary: as explained p. 3551, SPAN assess stability while FF and N_T assess reliability. So none of them assess both reliability and stability, –this is why we have to use several.

References

[Brigode et al., 2014] Brigode, P., Bernardara, P., Paquet, E., Gailhard, J., Garavaglia, F., Merz, R., Mićović, Z., Lawrence, D., and Ribstein, P. (2014). Sensitivity analysis of SCHADEX extreme flood estimations to observed hydrometeorological variability. *Water Resources Research*, 50(1):353–370.

[Garavaglia et al., 2010] Garavaglia, F., Gailhard, J., Paquet, E., Lang, M., Garçon, R., and Bernardara, P. (2010). Introducing a rainfall compound distribution model based on weather patterns sub-sampling. *Hydrology and Earth System Sciences*, 14(6):951–964.

[Garavaglia et al., 2011] Garavaglia, F., Lang, M., Paquet, E., Gailhard, J., Garçon, R., and Renard, B. (2011). Reliability and robustness of rainfall compound distribution model based on weather pattern sub-sampling. *Hydrology and Earth System Sciences*, 15(2):519–532.

[Gupta et al., 2009] Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91.

[Renard et al., 2013] Renard, B., Kochanek, K., Lang, M., Garavaglia, F., Paquet, E., Neppel, L., Najib, K., Carreau, J., Arnaud, P., Aubert, Y., Borchi, F., Soubeyroux, J.-M., Jourdain, S., Veysseire, J.-M., Sauquet, E., Cipriani, T., and Auffray, A. (2013). Data-based comparison of frequency analysis methods: A general framework. *Water Resources Research*, 49(2):825–843.