Referee report on manuscript HESS-2020-44 :
"Spatial Dependency in Nonstationary GEV Modelling
of Extreme Precipitation over Great Britain "
(by Han Wang & Yunqing Xuan)

Main comments

The submitted paper proposes, in the context of flood frequency analysis, to assess that non-stationary GEV models are more suitable than stationary ones for modeling the extreme rainfall in the UK mainland territory. In my opinion, this is a relevant research question, that this paper tries to address.

Here are the problems I found while examining this manuscript. The main issue is the lack of clarity in the description of the methodology used.

1. An important issue with this manuscript is that *the authors do not provide a clear description of the type of data they are working with*.

   First of all, the reader has a hard work (I did) finding in the text the description of how the UK territory is separated in different zones, and the shape of these zones. Why considering only the mainland of UK ? Apparently, the territory is cut in non-rectangular zones , but *why not simply using rectangular zones, since only the very mainland part of UK is studied ?* (this is, in my opinion, an important issue with the paper). All the paragraph of the top of page 7 is rather obscure : what does the sentence "the focus is on the impact of location only" mean ? What do the lines 173-174 mean ? Do they mean that the zones have a common shape (that of Figure 1a ? Why this one ?) ? Why considering "randomized locations" ? It is not normal that such crucial description of what the study is about, is so badly described. And the reader can only wonder why a non-negligible part of the UK territory is not covered by the study (the seashores, and all the space between the different zones...).

   Second, the concept of "simulated samples" is not well described. If I understand well : (*i*) they start from real data from the GEAR dataset, and for each area, each model (stationary and non-stationary) and each estimation method, they produce estimates of their parameters (either (μ, σ, ξ) or $(\mu_0, \mu_1, \sigma_0, \sigma_1, \xi)$) ; (*ii*) with these estimates, they produce simulated datasets with these estimated parameters as inputs, one for each area/model/method ; (*iii*) they produce "indicators", which are the basis for selecting what they call the "best" fit for each area.

   If the simulation process is the one described above, why don't the authors present it in a clear manner in their manuscript ? This will enhance greatly its readability and accessibility.

   Finally, concerning this data description issue, in step (*ii*) above, I suppose that the authors produce *several* simulated samples (an not just a single one), but the text does not provide this precision... Only Figure 7b suggests that several simulated samples are generated (each one producing a quantile estimator for each return period). This example of lack of precision is symptomatic of the unclear way the methodology is presented in the paper.

2. Concerning the indicators which are used to compare the models (S or NS) or the estimation methods, I have some real concerns with them.

   Concerning the "Diff" indicator, its definition is unclear : line 153 of page 6, what do y and y' precisely refer to ? Examination of Figure 7b gives some hints of what the authors are doing, but Lines 151-153 of Page 6 are very mysterious and imprecise...

   The use of the Kolmogorov-Smirnov statistic in this paper is a real issue for me, particularly for the non-stationary modeling. The authors say, page 6, *"The test is carried out by comparing the empirical cumulative probability distribution with the GEV cumulative probability distribution"*. However, what does "*the* GEV" refer to ? If $X_1, \ldots, X_n$ ($n = 113$) denote one simulated sample for one area, for the non-stationary (NS-GEV) model, then the elements of this sample are precisely constructed to <u>not</u> follow the same probability distribution ! So, if non-stationarity is strongly present, one cannot imagine that the "sample" $X_1, \ldots, X_n$ will fit a *single* GEV distribution. It is thus curious that the authors report (in lines 178-179) high p-values for the non-stationary model...

I also have some concerns with the stationary situation : in this case, it is true that testing that the sample is issued from <u>a</u> GEV distribution here makes sense. However it is very "surprising" that the p-values are all reported as being close to 1. *If* these p-values were correctly acquired and computed, and if most samples were correctly fitting a GEV distribution, then it is "common" statistical knowledge that the p-values should be uniformly distributed on the interval $[0, 1]$ (and if some samples were inconsistent with the GEV family, then the corresponding p values would be *small*, not close to 1...). They don't here, according to the authors. My analysis is that the authors certainly make an unfortunately very common mistake, which is to compute p-values for the KS test when the target distribution was *estimated*, by using the KS test distribution when the target distribution F_0 is *known* and is not estimated from the data. This mistake makes the non-rejection (line 179, page 6) of the hypothesis "the AMDR follows the GEV distribution", not very reliable (note that, anyway, this sentence is strange : what does the article "the" mean ?).

Moreover, in line 147 of page 6, the authors do not precise (i) what sample is fitted to a GEV distribution ? (the simulated one(s) ? or the initial one from the GEAR dataset ?) ; (ii) on which sample is based the estimation of the parameters of the GEV distribution to which the sample (of (i)) is fitted ? This is an important point, which should probably explain why the p values are found to be (artificially) so close to 1.

The authors might think that I have a rather severe judgement on their manuscript, however in my opinion it is normal to await a clear methodology description in a statistically based research work, if one wants the scientific findings to be accepted.

3. Another concern is that the reader has no idea of the influence of the choice of the region shape, or of the location of the centers; *how can one know whether the final findings of the paper will not (or will !) change if this common shape and/or choice of the location centers are changed ?*

4. Page 7 line 185, on which basis can the author say that the NS-GEV model "works better" than the S-GEV model ? (*i.e.* on which basis is the choice between the cross, and the squares, in Figure 3a, based ?). This is a crucial point of the paper .

   How was the choice between $\xi > 0$, $\xi = 0$ and $\xi < 0$ made in Figure 3b ? On the basis of which sample(s) ? On the basis of a statistical test ? In particular, what situation can lead to the choice $\xi = 0$?

   Figure 3d is interesting, however I do not understand why the values of μ_1 and σ_1 (which are estimates) are multiples of 0.01 : have they been rounded ? If yes, why ? If no, please explain this curious situation...

   Finally, in Figure 3c, what does σ_1 exactly refer to ? Does it refer to the estimate of σ_1 based on the initial GEAR data for the zone ?

5. In Figure 7b, how were the more extreme quantiles (return period of 50 or 100 years) estimated ? *And what do they mean / how are they defined in the non-stationary framework ??*

I will stop here for the general comments. If I admit being a non-specialist of the FFA topic, I nevertheless consider that the comments above are sufficient for not recommending publication of this manuscript in its present form, in HESS or elsewhere : many points need to be clarified.

Minor comments

- P1, lines 9-10 : sentence without verb.
- P1 , line 14 : what does "The most frequent AMDR as represented by the location parameter tend to be increasing..." mean ??
- P2, lines 49-50 : "In addition, ... nonstationary models" could be reconsidered, in regard to my general comment number 2 above...
- P2 , lines 63-67 : these sentences are not correctly written ("it then follows the introduction", "The specific focus on the spatial dependency of the methods tested the highlights of this study") or not very clear . Please rephrase them, since they are crucial for convincing the reader of the benefits of this work.
- P2, lines 31-45 : I am not sure the proposed bibliography of lines 31-45 is sufficient and/or sufficiently relevant, or should be completed.
- P3 , line 77 : the start of this subsection should be rephrased : in particular, the GEV cdf of equation (1) does not provide the "threshold value" X_n as a function of the return level F_n, but the converse.
- P3 , line 85 : not sure the reference Nascimento et al is the most appropriate here... Prefer a well-known textbook.
- P3, line 86 : what do the authors mean by "independent parameters" ?
- P3 , line 87 : several problems (including "hence the name." : ??)
- P4 , line 89 : "involve" instead of "contain"
- P4 , line 91 : what does the "r-th random variable X" refer to ? An order statistics ?
- P4 , line 93 : β_r is not an estimator, but the true value of the PW moment...
- P4, line 102 : this is absolutely not $Pr(X \leq x)$, but the empirical cdf $F_n(x)$...
- P4 , line 104 : the notations $\sigma 1$ and $\mu 1$ are curious, why not using σ_1 and μ_1 ? These are numbers, not vectors...
- P5 , line 110 : uppercase X_t, or lowercase x_t ?
- P5, line 112 : "where $f(\cdot)$ is the univariate density function" is unclear . In the stationary setting, it is the presumed underlying GEV density. But in the non-stationary setting, it should be f_t, because the density depends on the value of the parameter $\theta = (\mu_0, \mu_1, \sigma_0, \sigma_1, \xi)$, <u>and</u> on the value of the year t ! Can this "notation mistake" be considered as relevant of some sort of lack of comprehension of the backstage statistical theory by (one of) the authors ? Or lack of careful re-reading ?
- P5 , line 119 : the data is not described yet, so $n = 113$ is difficult to understand.
- Pages 5 and 6 , section on the B-MCMC method : this section is difficult to read, the notion of (random) step length is not defined, nor is the notion of "current state S_t". Either the authors should provide a preamble explanation of what a MCMC algorithm is run (and is about !), and be clearer in their explanations (for instance in an appendix), or they should provide an adequate reference for the statistical method that is run here. There is also some confusion between the state t in line 121,p5, and the state S_t...

- P8 , lines 198-199 : please rephrase, unclear
- P9 , line 235 : this sentence is not clear. What does "dividing the AMDR values by their associated probability P into four levels" mean ?? Rephrase this part.