Dear Ms. Iliopoulou,

Thank you for submitting your manuscript to Advances in Water Resources. Your paper has been evaluated by two reviewers and an Associate Editor who have advised against publication. After my own reading of your work, I concur with the reviewer's assessment. I regret to inform you that your manuscript will not be given further consideration for publication in AWR.

Please refer to the comments listed at the end of this letter for details of why I reached this decision.

We appreciate your submitting your manuscript to this journal and for giving us the opportunity to consider your work.

Kind regards,

Paolo D'Odorico
Editor
Advances in Water Resources

Comments from the editors and reviewers:
-Editor

- Associate Editor: The same two Reviewers who revised the original version of the manuscript have provided comments on the revised manuscript. While they have both appreciated the effort by the authors, they have raised a number of major issues, provided a detailed evaluation, and recommended rejection.

-Reviewer 1

I have read the paper "Projecting the future of rainfall extremes: better classic than trendy" by Iliopoulou and Koutsoyiannis as a complete new submission rather than treating it as revised version of the previous submission, since the authors did indeed made great changes to the original paper in a way which makes the comparison of the two manuscripts almost impossible. The authors investigate the performance of several modelling approaches using out-of-sample measures: the fact that the study focuses on the longest available rainfall records allows for a through investigation of the out-of-sample behaviour of the models. The authors find that the performance of trend-based estimation can be very poor and suggest that simpler/more parsimonious models are to be preferred. The manuscript is well organised and mostly reads well, although I find two major criticism in the exposition of the methods (and of the results as a consequence).

One item which still stand from my previous comments is that the authors use least square to quantify the (linear) trend in the raw indices series extracted from the rainfall series with no discussion at all given to the possibility of using different models based on more suitable distributions for the data at hand. Any serious investigation of trends in the proportion of wet days should at least explore/mention the fact that model based on the binomial (or negative

binomial) distribution might be suitable. Similarly annual maximina are typically analysed using a distribution which allow for excessive skewness (Log-Normal, Gumbel, GEV, LP-III). I understand the aim of the authors is to shed light on the dangers of blindly applying trend analysis and possibly extrapolate any detected trend, and I even sympathise with the argument, but the argument can not be won by carrying out inappropriate trend studies which do not use a suitable modelling strategy.

I find this a major issue which might influence the validity of the findings in the manuscript. I appreciate means and least squares can be estimated rather easily and enjoy optimal in-sample properties for large iid sample, but the authors do use the estimates to make predictions and in that case models which are more appropriate for the type of data under study should be employed.

I also have some major concerns regarding the paragraph starting at Page 7 line 115. First of all: statistical hypothesis testing are not a way to identify trends per se. Statistical hypothesis testing on some parameters of a specified model are a way to identify trends. The easiest way to identify trends is arguably to draw a smoother trough a data series and see how this looks like. Nevertheless to make a judgement on whether this trend is in some way relevant, statistical hypothesis testing can be used (for example testing the significance of a linear trend term or using a likelihood ratio test between a model using only the mean and a model using a cubic regression spline basis). But more importantly, traditional frequentist statistical hypothesis testing DOES NOT "estimate the probability that an alternative hypothesis may hold true". Citing Gauch (2003, Ch. 7 page 273 in the edition I found) "[The] p-value is the probability of an outcome as extreme as or more extreme than the experiment under the assumption that the null hypothesis is true". A similar definition can be found in the several papers which the authors cite which discuss the mis-use of p-values in applied sciences. I think anybody criticising the use of p-values should define them correctly. Further, one could construct two models which includes ways to model the persistence in the data series and the lack of independence which can undermine results for a least square estimate and have a trend/no-trend component for which a statistical hypothesis test can be constructed. The issue the authors underline is the fact that the trend models used in many cases are not suitable for the data they are applied to, but the issue is in the modelling approach not in the hypothesis testing mechanism.

Some smaller comments:

Page 8, line 135: In the sentence "With a stronger focus on modelling power rather than confirmatory analysis" it is unclear to me what "modelling power" and "confirmatory analysis" mean

Page 7, line 145 ... argues against the concept ... Which concept? [I assume statistical significance, but the last mention of it is quite far in the paragraph]

Page 10, line 181/182: the past performance of models is a good indicator of their future performance if the data generating process is not undergoing radical changes at a very fast pace,

which is what some fear is happening under climate change. If (and I stress if) rainfalls indicators have been stable for most of the 20th century and in the last years are going trough some radical shift, the past performance of models would not be indicative of the future performance. The problem is of course we don't know what the future will look like and the authors make their point that it is best to use simpler models to describe the behaviour of rainfall indices. I think though this underlines a more pervasive attitude in the paper: I feel the authors omit to discuss/acknowledge that the reason why there are so many trend studies is the urge to investigate the possible impacts on the natural system of the increase in the global temperature. There is an unspoken causal assumption underlying many trend studies, that is that (man-made) climate change is affecting several parts of the climate system including rainfall intensities and amounts, and while I can agree that fitting linear trends to any series under the sun is not a good way to move scientific understanding forward, I feel the authors are a bit reticent on the fact that the interest in trends is linked to the interest in impacts of climate change.

Page 13, line 247: "present climatic period": should this not be calibration period? Present indicate the latest one. The sentence in line 248 is somewhat unclear ("according to the followed calibration scheme" is quite clunky)

Page 16: if the trend found is very mild, one would imagine it would be deemed not significant in a statically hypothesis testing and the simple mean would be preferred as a model. Any insight on the difference in RMSE for series in which the trend was found to be quite strong (and likely to be significant)?

I think Section 4.3.2 is an important piece of discussion which is often lacking in statistical modelling of hydrological series: simpler models are indeed to be preferred. I would have maybe recalled here that Information criteria such as AIC discussed earlier in the paper are indeed constructed to favour parsimonious models. Moreover one could also add pointers to more modern discussions about ways in which parsimonious models are of importance. For the Bayesian view I am thinking, among other, about penalised complexity priors (arXiv:1403.4630), spike and slab priors (arxiv:0505633), while more "frequentist" friendly approaches like the Lasso, the non-negative garrote or simply ridge/regularised regression could also be mentioned. The topic is complex and challenging, and I perfectly understand the authors can not do it full justice in a subsection of a manuscript which deals with something else.

-Reviewer 2

- This version of the paper is substantially revised, and a lot of work has clearly gone into it. According to the authors' response to our previous comments, the major changes are (a) an emphasis on out-of-sample prediction as a criterion for distinguishing between models (b) replacement of the "shuffling" experiment with an experiment based on simulations of long-memory ("persistent") stationary series (c) changes in the way that performance is measured e.g. by using RMSE throughout instead of BIC.

Unfortunately, despite the extensive changes and the authors' attempt to justify what they're doing in their response letter, I remain unconvinced. This is partly because to a casual reader, the

main message of the paper will still come across strongly as "there's no evidence for trends in rainfall time series": indeed, this is more or less the implication of the title, which gives no clue that the main point is actually to focus on out-of-sample performance as the authors claim. Moreover, the criticisms of trends, and the poor performance of the "trend" models in the paper, are all associated with *linear* trend models; but this is not emphasised sufficiently. After reading the paper carefully, I find that the only conclusion one can reasonably draw from the work is that your predictive performance will be poor if you fit a silly model. Despite the apparent lack of awareness of this in much published literature, I don't think it's a message that merits publication.

Now that the focus on out-of-sample predictions has been clarified, I have some secondary concerns about the way that these have been evaluated - or, rather, I'm concerned that the evaluation isn't as informative as the authors think it is. This is because it's well-known that to minimise the prediction RMSE, your prediction should be the conditional expectation of the future values given the observations. For any stationary process, this conditional expectation tends to the overall mean with increasing lead time, albeit slowly in the case of long-memory processes: the "global" and "local" mean estimates considered in the present paper can be considered as estimates of this overall mean - with the "global" estimate being more accurate. It's completely unsurprising, therefore, that the mean models outperform the "trends" in the synthetic example here. It is, moreover, not particularly surprising that the mean models outperform the trends in the analysis of the real rainfall series - although the argument here is slightly more sophisticated than the obvious statement that "it's unwise to extrapolate a linear trend beyond the range of the data to which it was fitted". The reason is that there are some nonstationary process for which the predictions are linear even though the generating process contains no deterministic trend: any process that is stochastically equivalent to an ARIMA(p,d,q) with d=2 has this property, for example - and some nonparametric trend estimates can be regarded as derived from models in which this equivalence holds. Similarly, any process that is stochastically equivalent to an ARIMA(p,d,q) with d=1 has the property that the predictions are constant - even though the process itself is nonstationary. The latter point is important here: the fact that constant predictions yield the best RMSE does *not* necessarily mean that the series is stationary. Indeed, this highlights a weakness in the authors' approach, because it shows that there are multiple processes, potentially with very different properties, that cannot be distinguished using RMSE. It is certainly true that predictive performance can be used to discriminate between models, but RMSE on its own is not an adequate measure of performance for this purpose: a measure that incorporates some element of prediction uncertainty is needed as well. I note that the other reviewer pointed out the need to consider uncertainty in their previous report: the authors would have done well to take this comment more seriously.

On the basis of these concerns, unfortunately my view now is that this paper is not suitable for publication. If the editor disagrees however, some further minor / detailed comments may be helpful. These are as follows:

- Lines 24-25: I agree that there's a lot of natural variation in the rainfall process so that long-term predictability is hard. Actually, this is consistent with the results quoted in many IPCC reports, where it is common to find that the direction of future precipitation changes cannot be predicted with high confidence.

- Lines 28-29: it is not surprising that the local mean model is favoured when applied to "persistence" time series - see general comments above.
- Lines 64-66 "in the field modelling ... as hypothesis testing": there's something wrong here, the sentence doesn't make sense.
- Line 71: for the first time, it becomes apparent that the criticism is of *linear* trends. This has not been clear until now, nor is it really clear from the abstract. If the paper is to be published anywhere, I think there should be a very clear statement somewhere that makes explicit that the conclusions can only be taken as providing evidence against the *linearity* of any trends that may be present. The title should also be changed, to reflect this (see general comments above).
- Line 118: this continues to assert that significance testing is a "dated scientific method for model evaluation", which is simply not true. Used appropriately, it is a perfectly acceptable part of the modern analyst's toolkit. The criticism should rather be of the inappropriate use of testing.
- Line 229: what do you mean by the "temporal propagation of the errors" here? Also in line 319. I wonder if "evolution" would be a better word than "propagation", because "error propagation" has a precise meaning in a different context.
- Lines 368-369: what's the rationale for claiming that "in terms of the standard deviation of the RMSE distribution, it is evident that the local mean model prevails"? What are we supposed to be looking at here, and why?
- Lines 395-396: what do you mean by "we plot the average ECDF of non-overlapping segments ..."? What *quantity* are you plotting the ECDF of?