Background information

This is a resubmission of the manuscript "Projecting the future of rainfall extremes: better classic than trendy" with reference number ADWR_2019_585, submitted on 22/7/2019. The initial version of the manuscript was reviewed by two Reviewers and the Associate Editor, and a decision was reached on 25/8/2019. After the Reviewers' assessment, the Associate Editor and the Editor Paolo D'Odorico recommended rejection with an invitation to resubmit the manuscript in case the Reviewers' criticisms could be addressed by substantial revisions. Below we present a detailed response to the review comments.

Reply to the Editor

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 reviewers' assessment. I regret to inform you that I have decided to reject your manuscript from publication in AWR. However, I will be happy to consider a resubmission if you think that with substantial additional work you would be able to address the reviewers' criticism.

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

We thank the Editor Paolo D'Odorico and gratefully appreciate the opportunity to resubmit the manuscript. We are submitting a thoroughly revised version of the initial work along with a detailed response to the Reviewers' criticism. We are grateful to the Reviewers as both their criticisms and suggestions helped us improve the paper substantially. While the core results of our analysis have not changed, we acknowledge that it was a certain lack of clarity, ambiguous assumptions, as well as redundancy in the analyses that triggered most of the criticisms on the initial version. The manuscript we are submitting is a substantial revision of the original work, comprising extensive additional analyses and a more concise and comprehensive presentation of the methodology and the motivation of the analysis. Therefore, we wish to thank again the Editor, the Associate Editor and both Reviewers for the constructive feedback and the opportunity for revisions.

Replies to comments from the Associate Editor and the Reviewers:
-Editor

- Associate Editor: Two Reviewers, who are experts in the field, have commented on this manuscript. While they both think that the topic is of interest, they found that the work would require a significant amount of work to make it publishable in AWR, as detailed in their extensive and insightful feedback. I recommend rejection with an invitation to resubmit in case the Authors feel that they can satisfactorily address of the comments by the Reviewers.

We thank the Associate Editor for considering our work and giving us the opportunity to resubmit a revised version of it. We are providing a detailed response that we believe sufficiently proves that the work presented stands against these criticisms. However, we acknowledge that there were several parts of the manuscript that needed improvements, and hence we have made numerous changes to the initial manuscript. A summary of the major changes follows:

1. The BIC analysis triggering concerns regarding the distributional properties of the data and criticized by both Reviewers, has been removed and instead the concept of parsimony in relation to predictive accuracy is theoretically discussed in Section 4.3.2.
2. Analysis of related literature has been significantly expanded, comprising a thorough discussion in the main text (Section 3.1) as well as a quantitative literature review on the use of trends in hydrology, based on results from Google Scholar. The latter is presented in Appendix I.
3. The scheme for calibration and validation of the models has been extended and better clarified in section 3.2 including a better explanatory figure, along with a literature discussion on related validation concepts.
4. The global trend model is also considered as a candidate model for prediction, together with the global mean, the global trend and the local trend, as introduced in Sections 3.2 and 3.3.
5. A comparison of the least-squares approach to robust regression techniques is now included in Appendix III. This was a major point of the Reviewers.
6. Experimental results from analysis of persistent models have been added (Section 4.3) replacing the comparison of the observed to the shuffled data, since the latter were deemed irrelevant by Reviewer 1.
7. Presentation of the results has been shortened decreasing the number of figures to 10 in the main text.
8. The manuscript has been restructured and various language improvements have been implemented.

-Reviewer 1

- This paper rightly criticises a simplistic but widespread approach to the quantification of long-term variability in precipitation time series, and proposes what the authors call a "multi-model" framework to overcome the deficiencies of such an approach. The features of the authors' proposed approach are (a) an analysis based on a comparison of different statistical modelling approaches (b) use of independent validation periods to compare the performance of different trend scenarios when extrapolated beyond the period used for their calibration. These features are

entirely sensible and, indeed, are more or less in line with the way that a competent and informed modern statistician would think about the problem.

As such, the paper perhaps doesn't offer anything particularly novel. Indeed, there have been several other papers with similar messages published over the last decade or so, of which the authors seem unaware - see below. Moreover, there's a long way to go before the present paper can be regarded as indicating "best practice". My overall view is that it has potential, but the authors need to broaden their horizons quite considerably before it can be published in a reputable journal.

By way of justification: my interpretation is that the paper is primarily about statistical methodology. In view of this, the lack of up-to-date statistical references in the bibliography is surprising: apart from some papers on the (ab)use of p-values, the only "credible" statistical references (in my view) are Akaike (1974) and Priestley (1981). Both are several decades old. Any paper seeking to improve the quality of statistical analysis should be better informed by the statistical literature, and also by the growing number of papers in the climate and hydrology literature that showcase the applications of modern statistical methods in related application areas (notably with respect to the assessment of trends and changepoints). In particular, the 2011 textbook "Statistical Methods for Trend Detection and Analysis in the Environmental Sciences" by Chandler and Scott covers many of the authors' points and is informed by a wide-ranging awareness of currently available techniques.

We thank the Reviewer for finding interest and potential in the paper but also for the constructive comments provided. We appreciate both the time devoted to our work and the thoughtful comments of the Reviewer, which indeed help us improve the paper. Before moving to the specific comments, we briefly discuss some of the points made above.

The main criticism of the Reviewer is that these features are entirely in line with the approach of a ‘competent and informed modern statistician’, suggesting that there are ‘several other papers with similar messages’. First off, we want to emphasize that we are addressing this paper to the hydrological community and not to modern statisticians. Most likely the Reviewer has a strong background in statistics, but a good hydrologist is not always in line with the methods a competent modern statistician uses; many papers published in reputable journals in hydrology are knowledge transfer from the statistical domain. Second, we understand that the aim of the work was not properly communicated in the previous version. The paper is not about statistical methodology per se, and does not really intend to suggest best practice, assuming we could uniquely define one. The paper proposes a shift in evaluation of trends by focusing on *predictive accuracy* of trends, rather than in-sample inference. By including a brief review of statistical methods in terms of in-sample and out-of-sample evaluation, we have clarified that there is a clear distinction between the two approaches, and this serves the different objectives of each study (Section 3.1). Of course, we do not claim that in-sample evaluation is irrelevant, neither that out-of-sample evaluation based on predictive performance is a new concept. On the contrary, we start by acknowledging Vit Klemes’s “split sample” scheme (who by the way writes in his abstract “The scheme contains no new and original ideas”). We argue that predictive performance has not been systematically explored for trend modelling, where typically the focus is placed on in-sample measures, but there is theoretical and practical basis for investigating it

because it is strongly linked to expectations for the future. To place the work in a wider statistical context, as the Reviewer suggests, we have included a brief but thorough review on the concepts of in-sample and out-of-sample evaluation, based on recent statistical literature, presented in Sections 3.1 and 3.2. We believe that this review has indeed clarified the paper's motivation and improved the context of the analysis.

We also wish to point out that the sparse critiques on trends (which we understand is what implied by the Reviewer by 'papers with similar messages') existing in the literature are conducted from a different viewpoint than the one we are using. They could be considered 'similar' perhaps in terms of a very general conclusion towards deterministic trends, but certainly neither on the methods used nor on the perspective. We deem that our approach is novel in terms of empirical evidence, since more and more hydrological papers are focusing on trends by in-sample analysis (see our review in Appendix I). Thank you for the comments referring to amelioration of the methods and analysis used, which we have taken into account and we discuss next.

We also thank the Reviewer for suggesting this relevant book, which we have consulted and found helpful.

Following from this overall assessment: one of the key parts of the paper is the critique of statistical significance in Section 3.4. Although I agree with some of it, I think that this critique represents an extreme position. In particular, I disagree that "statistical significance is now considered a poor and outdated scientific method for model evaluation". It is true that some have called for the concept to be abandoned entirely, and the majority of competent statisticians would surely agree that its "blind use" is inappropriate. Used appropriately however, it is a perfectly legitimate instrument in the analyst's toolbox. By "used appropriately", I mean that whenever a hypothesis test is conducted, the analyst must be clear as to what null hypothesis is being tested; moreover, this null should be defined in such a way as to be of scientific interest. The Akaike quote in lines 176-180 is somewhat in this spirit, although it is framed in terms of the statistical techniques that were available half a century ago: the discipline has evolved since then, and much more nuanced testing procedures are now available that allow us to test null hypotheses that are scientifically relevant. In particular, we now have test procedures that are designed precisely for discriminating between competing models - in a similar spirit to the authors' proposed approach perhaps, but as an alternative to indicators such as BIC that come with their own problems.

In principle we agree with and thank the Reviewer. But certainly, the critique on hypothesis testing is not a key part for our analysis and perhaps we did not properly communicate that in the previous versions. We have downplayed some of the relevant statements, without impacting our key premise that is the distinction between in-sample and out-of-sample measures. This section is entirely updated now and portrays the message more clearly. Yet we have kept some discussion on the misuse of hypothesis testing, which we think is important.

A further concern, which at first glance seems semantic but is actually quite important, is the authors' use of the word "trend". The title and abstract suggest "on the whole, there is little

evidence for trends in precipitation properties around the globe" - and this perception is reinforced by statements such as "Results consistently disfavour trend modelling for all rainfall indices" (line 496). What the analysis actually shows, however, is that long-term precipitation variability cannot be captured *using linear trends fitted using least squares* (please note the emphasis). As far as I'm concerned, this is not news - and issues such as the potential for negative values (see line 277) are barely worth discussing in reputable journals because they are discussed in first-year undergraduate statistics modules (take logs and the problem is solved). I know there's a lot of rubbish in the peer-reviewed literature, but it *does* seem to me that the authors are cherry-picking their criticisms to suit their message. In particular, they seem unaware that their "local mean" is actually a (very dated and imperfect) nonparametric estimator of a nonlinear trend. The discussion in (for example) Section 4.1 suggests that their interpretation of "trend" is very limited in scope. It is true that one person's trend is another's noise (see the Chandler and Scott text, referred to above, for more on this), but I doubt that many people would subscribe to the idea that trends can only be linear.

We are thankful for the comments. The Reviewer is right that we should be more specific to our use of the word, since indeed it may take a number of possible meanings. We have clarified our use of the word in Section 3.1 and downplayed most of the statements the Reviewer is referring to. In the same spirit, we believe that both the distinction we are drawing between explanatory and predictive performance, as well as the inclusion in the analysis of 'global trends' are more or less in line with the Reviewer's call for a wider scope in definition of trends. The application is necessarily limited in scope to some extent, since we cannot compare all conceived regression approaches. Yet the practice of fitting trends, and linear trends is widespread (Appendix I) and therefore, we believe that the comparison is fundamental and relevant. As for links to non-parametric statistics, please see added lines 253-258:

> "We note that these two seemingly simplistic predictive models, i.e. the linear model fitted with least-squares and local average, can be found in a variety of theoretical results in statistical sciences, for instance use of (temporally) local data constitutes a central concept in the *k*-nearest neighbors technique, as discussed in Hastie et al. (2005), as well as in local regression as discussed in Chandler and Scott (2011)".

We have also removed some phrases that were deemed trivial by the Reviewer; e.g. with respect to negative values (even though we believe they were didactic and fun). On the use of linear trends and least squares, please refer to next comments.

As a final general comment: many of the results are completely unsurprising to anybody who understands the underlying statistical principles. A lot of space could be saved, therefore, by focusing on a clear and correct exposition of these principles: many of the figures could then be removed without detriment (e.g. the ones where you compare the 2-parameter linear trend model to the 1-parameter L-mean and G-mean models for shuffled data, where it's known that all models are correct but the trend is zero and therefore the linear trend model is overfitted).

Thank you for the suggestions. Indeed, we have restructured the paper and reduced a lot the number of figures, and also removed some sections. At the same time, we have expanded ones focusing on ‘a clear and correct exposition of statistical principles’.

More detailed comments are as follows:

- Lines 98-99: are the WDAV and PD indices also calculated at an annual time scale?

Yes, we have added relevant clarifications in various parts of the text, where the indices are mentioned.

- Line 111: delete "stationary" here, it is not necessary (and, indeed, if the process of interest is known to be stationary then nobody should be fitting global linear trend models to it). Actually, no discussion of stationarity is needed at any point in a paper that is focused entirely on an empirical comparison of different "models".

Done.

- Line 113: the notation O for the forecast issue time is potentially confusing and will be read by many as 0 (zero). Call it t_O or something? Although, to be honest, why don't you just stick with i?

Done.

- Lines 121-122: the local mean model considered here is a very simple case of what modern statisticians would describe as a nonparametric regression model, albeit with undesirable properties compared with (say) local linear regression. To some extent there is value in demonstrating that a local mean provides more relevant information than a global linear regression (although it is quite depressing that the quality of so much science is poor enough to have to make the point); but here is a place where a better awareness of modern statistical techniques would be very helpful. This point becomes much more important later, in terms of how the conclusions of the paper are framed.

We have thoroughly revised this section (now Section 3.3). Regarding links to non-parametric regression we have added relevant context in Lines 253-258, as explained before. Obviously, this research could be reframed from a regression approach, favoring local regression, as the Reviewer suggests. But we are not convinced that this would make the paper more relevant to the hydrological community. In fact, we believe that the most relevant part is the distinction between out-of-sample and in-sample performance.

- Lines 125-6: this comment about global and historical means is a bit imprecise, for a variety of reasons - one being that you don't define what you mean by "global mean" and another, more important, that you don't state what convergence you're talking about. Stationarity isn't a sufficient assumption for convergence either: strictly speaking, you need ergodicity. The easiest way to deal with this is to remove the clause "although ... grows larger": this would not detract from your main point.

We now define what we mean by global mean by a better introduction of the calibration and validation scheme, clear wording and the new explanatory figure (Fig. 2) and we expect this to be much clearer; we have also deleted the lines the Reviewer was referring to.

- Line 161: what is "this step"? (I assume you mean "this second scheme").

Yes. This section is now entirely rephrased.

- Lines 162-3: I don't understand the "less than 5% of consecutive missing values" point. For a 30-year window this equates to 1.5 years rather than 3 as implied by "at least 27 valid indices" - and I also don't understand what is meant by *consecutive* missing values. Nor do I understand how the 5% criterion translates to "at least 60 time windows" (line 166).

We apologize that confusion arose due to our improper use of the word 'accordingly' in the previous version. This is deleted now and the section better explains the criteria and reads:

> "In general, most records have low percentages of missing values (Table A1), which in most cases are clustered in the beginning of the records. A few records have consecutive missing periods which might imply a change of instrumentation or relocation of the gauge. To avoid possible artefacts in trend estimation in static validation (in backward validation) that may arise from such cases, we analyze periods containing less than 5% of consecutive missing values of the yearly indices. For the dynamic calibration and validation scheme, we fit the models only if there exist at least 27 valid indices in each of the 30-year periods of calibration and validation."

In other words, the 5% consecutive missing values criterion refers to the choice of a record subseries (e.g. if in a 100-year record the first 2 years exist, but the next10 years are missing, we would entirely omit the first 12-year period, and start the analysis from year 13). This is relevant for the static validation scheme in backward validation mode. Three missing values are allowed in moving window calibration and validation, irrespectively if they are consecutive or separated in time. The 60 time-windows criterion is now omitted because we show the whole RMSE distribution, instead of computing single statistics.

- Lines 187-189: in the context that the authors describe, the independence assumption is indeed one possible reason for rejecting a null hypothesis. It's not the only reason, though: strictly speaking, the only thing you can conclude from these (dated) testing techniques is that the data are either not independent or not identically distributed.

Right, in the vast majority of trend studies what is rejected is the "identically distributed" assumption, without even considering the case that independence is not an appropriate assumption. So, in our paper we wish to stress the latter possibility which has not been given emphasis in literature, or is perhaps massively ignored.

- Lines 190-192: as noted in my general comments above, I think the idea of "regarding trends as

models" is absolutely right. It is not true, however, that statistical hypothesis testing cannot be used to discriminate between models: it can. It has its drawbacks even in this context, of course; but so do the model selection criteria used here.

The Reviewer is right; we have actually restructured the entire section and made the distinction between statistical in-sample inference and predictive performance explicit.

- Equation (2): this definition of BIC depends on an assumption that the data (actually the residuals from whatever model is fitted) are independent and normally distributed. In general, the first term should be -2log L, where log L is the log-likelihood for the fitted model. A corollary to this is that model comparisons based on BIC calculated using equation (2) are not appropriate unless the residuals *are* (roughly) independent and normally distributed. Have you checked this? A similar comment applies to the definition of AIC in line 204. I am highly confident that the normality assumption will fail spectacularly for series of annual maxima - this seems to be confirmed by plots such as Figure 3 which show that there are a few extreme outliers in many stations. Incidentally, this itself implies that the fitting of models via least squares is inappropriate: at the very least, a robust regression technique is needed. I am sure that this is the primary explanation for your results (e.g. lines 272-3) that the worst performance of the linear trend model is for annual maxima. **NB** I agree with you 100% that linear trends are silly in this context, I just don't think your criiticisms of them are the right ones.

Thanks for these comments. We decided to omit the BIC analysis for various reasons. First, backing up distributional assumption through tests is quite a problem, because for most schemes, the model is calibrated on 30 years, and thus the uncertainty regarding the distribution is large. Indeed, annual maxima have the most asymmetric distribution as far as all the record is concerned, but the effect is not as evident in the small subseries. What is more, we felt that the Reviewer is right (in the next comments) that it is atypical to use BIC for validation data. Our motivation was to judge parsimony of the models in case of them yielding an ‘equal’ prediction. Now instead we discuss this concept only on theoretical/philosophical grounds in Section 4.3.2. Finally, our major conclusion does not change with respect to RMSE, it is only that BIC magnified this conclusion by penalizing the extra parameter in prediction. All in all, we realize that both distributional and theoretical concerns may arise due to the use of the BIC in this context, that are cumbersome to mitigate, while as stated before this analysis is not central to obtaining our result.

As for robust regression, we have actually performed this in the analysis in Appendix III, and results have not improved (also stated in Lines 240-245). Further, we stress that the mere application of least-square fitting does not involve any assumptions on the distribution or the residuals.

- Lines 204-211: as a description of the differences between AIC and BIC, this is OK. Initially however, I thought that it missed the fact that the two criteria are designed to answer different questions. I'm going to choose my words carefully in what follows. Akaike's motivation in developing AIC was to answer the question "among a bunch of wrong models, which one will give me the best predictions in the future?". The strongest theoretical argument usually

advocated for BIC (at least, if you're not going to get into Bayes factors) is that it is a consistent model selection criterion: if you've got a bunch of models and the data were generated from one of them then, given enough data, BIC will correctly identify it. This distinction is rarely articulated in the literature. I note, however, that the authors use Ye et al. (2008) to justify their use of BIC. I was surprised by this, so I looked at the Ye et al. paper which led back to Cavanaugh and Neath (https://doi.org/10.1080/03610929908832282). The development in *this* paper boils down to the fact that given enough data, BIC selects the model that is *a posteriori* most probable, if you're prepared to concede *a priori* that all models are equally likely to have generated the data. This is interesting, but it doesn't contradict what I wrote earlier about (a) Akaike's motivation (b) the strongest theoretical argument usually advocated for BIC. At any rate, the justification for using BIC in lines 210-11 is not adequate.

We are thankful for pointing out this advantage of BIC, which however, as we have explained in the previous comments, we did not use in the revised paper.

- Lines 212-215: for "out-of-sample" evaluation, of course it's fine to compute the RSS and sample size from the data and predictions actually being compared. But I'm not sure about using BIC in this context, essentially because the penalty term in BIC is designed to compensate for the fact that models of varying complexity are being fitted and compared on the same dataset. If you introduce an "independent" validation dataset however, this issue disappears: if you overfitted to the calibration data, this will become apparent as soon as you calculate a measure such as RSS on the validation data. I am therefore not convinced by the use of BIC as a performance measure on the validation data.

We agree that it is atypical to use BIC on validation data, although there are reasons that could support it. Yet since the Reviewer finds that unconvincing, and keeping in mind the fact that the use of the BIC introduces distributional assumptions which are not easy to check in small samples, while it is not really necessary to obtaining the core result, we reiterate here that we decided to remove the relevant sections entirely.

- Line 235: the phrase "historical 'future' performance" is quite hard to interpret! Also, I don't understand what you mean by "all future time windows" for the static validation scheme illustrated in Figure 2(a).

These are now deleted. Static validation is better explained in a new figure, Fig. 2, where prediction with respect to the past is called ‘backward validation’.

- Lines 237-8: it would be helpful to indicate your motivation for calculating the standard deviation of the RMSE over time windows. I suspect you are thinking about consistency of model performance, but I'm not 100% sure.

In fact our aim is to summarize the model’s historical performance as we obtain a sample of RMSE values from the dynamic validation. Now we also plot the empirical distribution for all stations (Fig. 5-8), therefore this is clearer, and then we also include the plots where these are summarized by means of average RMSE and standard deviation of RMSE (Fig.9). The latter indeed is regarded as a measure of the consistency of the model performance.

- Lines 241-2: here you suggest that the second type of evaluation looks at cases where "a model is best for the present but is eventually outperformed by another one in the future". But earlier, you suggested that you're focusing on cases where the present model is *not* outperformed in the future. Please can you clarify?

This section is removed now, because it is less relevant after removing the BIC analysis. For clarification, we were referring to cases where the model is selected at present and tracked its performance on the future. When it was outperformed though, we showed by which models this was done.

- Lines 250-265: it's an interesting idea to shuffle the data, and almost reinvents the permutation test. However, this does not isolate the effect of dependence on model selection as the section heading suggests: it also destroys any trend that may have been present in the original data. This is acknowledged in the last paragraph of the section; but in comparing the metrics for the shuffled and original series you are committing precisely the same error for which you criticise statistical hypothesis testing: the shuffled series are, by definition, white noise and therefore provide a scientifically uninteresting comparator.

The comment made us think that a more interesting scientific comparison would be one with respect to persistent series. This is relevant not only due to the presence of persistence in most hydrological variables but also because segments show indeed 'trend-like' behavior, therefore fitting a trend without being aware of persistence could have been done. This analysis is introduced in new Sections 3.5 (Predictability of climatic changes under natural variability) and 4.3 (Models' performance under natural variability).

- Line 275 "climatic trends quickly reverse": this is well known as a phenomenon of decadal-scale variation. In any statistical characterisation of change, the solution is to work within a framework that acknowledges the potential for variation on this kind of timescale *alongside* longer-term variation.

We agree but this is actually rarely done in practice. Most importantly, typically tests taking this into account use autocorrelation metrics, but long-term variability is not sufficiently captured by autocorrelation. In any case, we agree with the Reviewer but we disagree that this is well-known; we rarely encounter competent discussion of decadal-scale variation in trend studies. In any case, we have now clarified that we are referring to 'empirical evidence'. The relevant phrase now reads:

> "A visual examination of the plots of the 60 long-term stations, provided in the Appendix figures (A4-A7), suggests a positive answer to the opening question, providing empirical evidence that climatic trends fluctuate and in fact, abruptly reverse."

- Table 1: I don't understand what the "Past validation" column here refers to. You might also point out, either in the caption or in the text, that the linear trend model is fitted by minimising the residual sum of squares and is therefore guaranteed to outperform any other model according to RMSE in the fitting period. Hence all the zeroes in the first column.

We have now deleted this table. ‘Past validation’ is now referred to as ‘backward validation’.

- Lines 308-9: it is unsurprising that the local mean model is not preferred in the shuffled data because, as noted above, this model is a (poor) nonparametric estimator of a nonlinear trend - and the shuffling destroys any such trend so that the global mean, which uses all available data, is guaranteed to do the best job. As such, it is no surprise that the global mean model is preferred when the data are shuffled (lines 312-313). Moreover, the "direct explanation" on lines 313-314 is wrong - or, at least, incomplete. Similarly, the claims in lines 314-5 about misinterpreting dependence as trend have some basis but need to be much more nuanced. I'll refer to the Chandler and Scott text again.

Regarding shuffling, this has been removed in favor of showing the persistent series instead. Now a more explicit and nuanced discussion of persistence is included in Section 4.3. The statement is slightly modified with respect to the previous one and is moved to the conclusions (Lines 491-493). It now reads:

> “Persistent processes show enhanced variability and a user unfamiliar with their properties may misinterpret segments of their timeseries as trends, which perhaps explains why trend claims have been that common lately.”

‘Direct explanation’ is removed as well and a discussion in terms of consistency of empirical and theoretical results is included.

- Lines 352-354: this comment about impact on selected models for the shuffled data demonstrates a lack of awareness. The shuffled data are white noise in all cases, and the modelling techniques and metrics used here only depend on the first and second moments. The result is entirely expected, therefore.

We have not claimed that this is a discovery neither something unexpected; the text simply states that it is ‘evident’. Probably, the Reviewer intends to imply that it is trivial to even explain this, but we would like to stress that we were not using a standard technique here. Rather we were introducing results from one we devised ourselves. In this context, this statement was reiterating the motivation behind using shuffled data. We believe that it was not redundant for a reader, perhaps less experienced with the concept of dependence in time. Anyway, this section is removed from the revised version because we use series from persistent models instead.

- Line 357: what am I supposed to be looking at to find "the false trend discovery rate by RMSE"? Also, what are "the random data"?

We have removed these along with the section.

- Pages 26-29: these figures are hard to read / interpret, and it's not really clear what message we're supposed to be taking from them. Are they all necessary? Similarly pages 32-35.

The Reviewer is right, we have shifted to showing the empirical distribution of the RMSE for all stations and therefore there is no need for these plots anymore.

- Section 4.3: the authors note that the comparison here can only be performed for the case where BIC is used for model selection. I have already expressed my concerns about the use of BIC in a validation context: I will not comment further here, therefore, except to say that the analysis seems to contain an inbuilt bias against the linear trend model because it has two parameters. I should perhaps clarify, if needed, that in my view the linear trend model should not even be contemplated in the way that it's used here (although I'm aware that this is often done): my criticism, therefore, is of the authors' reasoning rather than their final message.

We have dropped the relevant sections as explained before.

- Line 445: what do you mean by "clustering of errors"?

We were referring to the temporal propagation pattern of the error. Now we have rephrased this section as well.

- Lines 448-450: from the results presented here, you cannot possibly conclude that the performance of the L-mean model is due to Hurst-Kolmogorov dynamics (also known as long memory). The local mean is a nonparametric trend estimator. Essentialy therefore, you're trying to distinguish between the models y[t] = mu[t] + e[t] (nonlinear trend, independent errors) and y[t] = z[t] (no trend, long memory errors). Such discrimination, on purely empirical grounds, is effectively impossible: indeed, it depends on your definition of "trend" (Chandler and Scott again, if I remember correctly).

Indeed, the Reviewer is right, our wording should have been more carefully since this type of scientific inference refers to induction. We have rephrased the relevant sections by suggesting ‘consistency’ and ‘similarities’ with findings from the added experiment on synthetic persistent processes, e.g.:

> “We note that the behavior observed in the N=100 plots is qualitatively consistent with the one observed from the rainfall records..”
>
> “Results from the synthetic records show qualitative similarities with the ones from empirical rainfall records,..”

What is more, the purpose here is defined entirely with respect to prediction, rather than discriminating between underlying processes. We have clarified this.

- Lines 450-456: these criticisms of the linear trend model all stem from the fact that it's basically wrong in a really bad way. It isn't *just* the presence of extreme observations. I think the authors probably agree with me on this: so why labour the point?

Thanks for the comment, we have rephrased these lines, making the discussion more explicit, and providing more reasons than the presence of ‘extreme observations’. Also, this analysis is now presented as an example of the methodology (Section 4.2.1). The relevant lines now read:

"The error propagation pattern of the models is reflective of their performance. For the majority of time, the mean models are at the lower front of the errors, with the local mean model showing slightly superior performance. The local linear trend model results in higher errors and its predictions may quickly deteriorate, taking longer to converge to the mean models in areas of lower errors (Fig. 4). This is attributed to the fact that the trend model projects sensitive features of the calibration period, i.e. extreme observations or 'trendy' behaviour, which do not have a high chance to survive the end of the calibration sample. The more parsimonious structure of the mean model encapsulates minimal but robust knowledge of the process behaviour, which is more likely to characterize its future evolution as well. In the absence of an underlying global trend and as the sample grows larger, the global trend model converges to the predictions of the mean models, but its performance remains slightly inferior even towards the end of the record."

- Line 476: is it true that trend modelling is *increasingly* encountered? "commonly" encountered, yes; but I would hesitate to claim that the present generation is sinning more frequently than our predecessors.

This question motivated us to quantify the change; please refer to the quantitative analysis in the Appendix I.

- Lines 477-483: these lines overstate the novelty of the paper, due to the authors' apparent lack of awareness of other relevant literature. Elsewhere, Section 5 more or less repeats the (perhaps unconscious) biases that are evident elsewhere in the paper: in my view, therefore, it needs substantial revision in line with the suggestions above.

In terms of awareness of the literature, we hope that the Reviewer will be satisfied by our additions (both quantitative and in discussions in the text). As for the unconscious bias, we have reread the whole manuscript keeping in mind the Reviewer's criticism, we have downplayed some of our critiques that may have appeared superfluous and kept only the most relevant parts. By including in the analysis the global trend models and removing the BIC part, the conclusions' section has been almost entirely rewritten and the novelty revisited as well. On the whole, we sincerely thank the Reviewer for the extensive and constructive feedback and we eagerly expect a positive reevaluation.

-Reviewer 2

- Comments by Reviewer 1 (I am pasting what Reviewer 1 sent me because he/she had issues with submitting their comments; I put neutral to the questions above)

Comments to the authors

The paper Projecting the future of rainfall extremes: better classic than trendy by Iliopoulou and Koutsoyiannis presents a comparison between different models for the projection/forecasting of

rainfall extremes. In particular they compare models based on trend over models based on local or global mean, finding that using a trend often gives poorer performance in terms of some suitable model performance metrics (BIC and RMSE). The presentation of the paper could in some places be simplified: some sentences are quite long and not always easy to parse, but overall the manuscript is well structures and gives a good presentation of the methods and results. My main concern, as mentioned below, is the choices made by the authors for the modelling of the indicators they decided to include in the study. It appears form how the BIC is presented that the normal distribution is assumed for all variables under study, although this might be an incorrect assumption, and as any assumption should be validated and investigated in the modelling. The choice of distribution plays a key role in the calculations of the BIC value used by the authors, and before dismissing any modelling framework as not useful I think we should try to find the best possible model for the data at hand, including the choice of distribution.
More comments blow.

We thank the Reviewer both for the positive comments on our work and the constructive criticism. Regarding the BIC, we agree the distributional assumptions introduced are first, hard to assess (since most samples are small; 30 year values) and the end are not central to the message of our analysis. Therefore we have deleted this section. We have also simplified the presentation in various points inside the manuscript.

Page 7 - line 125, so the global average is all observation up to time O (x_1, X_O)?

We have now simplified the notation by dropping the symbol of time O and adding a new explanatory figure (Fig.2). Note that we have also included the global trend in the analysis as well. Relevant section now reads:

> "According to the followed calibration scheme, fitted to block-moving (local) 30 years or to all the known (global) period, the trend model is termed local trend (L-Trend) and global trend (G-Trend), respectively, and likewise, the mean model, is termed local mean (L-Mean) and global mean (G-Mean). In the local models, the period [i – 59, i – 30] is used for calibration and the [i – 29,i] for validation, while in the global models, the period [1, i –30] is used for calibration and the [i – 29, i] period for validation as in the former scheme."

Page 8: the presentation of the validation schemes could maybe refer more to the figure and to the (O-29,O) and other O-based notations used in the previous paragraph which could be added to the Figure? It is just not so easy to follow how validation schemes fits with the notation used before and with the Figure.

Thanks for the suggestion, we have majorly revised the presentation of the validation scheme, changed the notation and added a new figure (Fig.2) which hopefully is much more comprehensible.

Line 137-141 two "also" are included and the sentence is very long. Maybe split to give more importance to each sub-objective.

Done.

Page 11: Given that the RSS is the same quantity as the RMSE - maybe the BIC can be written as a function of RMSE, or the RMSE explained as a function of RSS? Further, if it is true that in the AIC for large n the likelihood component dominates the AIC value, using RMSE essentially corresponds to using AIC.

As already explained, BIC is removed now. As for AIC and RMSE correspondence, this is true for the large samples (e.g. applicable in the backward validation in the 'static validation' scheme).

More importantly though, the BIC as written assumes the normality of the distribution: the likelihood part of the formula is derived under the square loss typical of the normality distribution. Nevertheless there is little exploration of the validity of the assumption in the manuscript. The assumption is surely not true for the probability-dry variable, which is bounded in [0-1] and should be modelled possibly by a binomial distribution by means of logistic regression, and is highly dubious for annual maxima, which tend to follow heavy tailed distribution. These two variables are the ones for which trends are found to be the least satisfactory model, but possibly this is due to a poor trend model choice rather than the trend itself.

Indeed distributional assumptions are an issue with BIC. Strictly speaking this assumption should be checked in the small samples from which BIC is computed, i.e. typically 30 years (except for the calibration in the global-moving scheme). Therefore, deviations from normality (or another distribution) would not be that prominent in these small samples, but in any case, the Reviewer is right that there is heavy dependence upon this assumption. After revisiting our scheme, we have decided that this is not necessary to our results, as evaluation via RMSE is more straightforward and suffices to reach a conclusion. Regarding dependence on choice of least-square fitting, please refer to next comments.

For many parts of the manuscripts the authors discover something which has been known since the beginning of regression models: extrapolation beyond the limits of the observed explanatory variable is not advisable, especially not very far beyond the observed range of the explanatory variable (time in this case). I suspect that if prediction intervals were added to the trend lines in the Figures 3-6 they might in many cases encompass the observed data. There is nothing striking about the fact that in some cases extrapolating far in time will give poor results. I appreciate that it is important to make this more prominent in the literature, but it not surprising. As per the point made by the author that extrapolation even produce negative rainfall, I believe this could be fixed by using a different distribution than the normal which could be bounded at 0.

Indeed, this might be expected theoretically (particularly, caution regarding extrapolation is advocated in well-informed statistical books) but it has not been shown in empirical records of rainfall, to the best of our knowledge, particularly with respect to the better performance of the

local mean. Additionally, our new analyses show that persistent series also share similar properties in terms of predictive performance (Section 4.3); this is a new conclusion. More so, the revised paper attempts to draw a clear distinction between evaluating explanatory and predictive performance. Overall, we deem that it is important to make this point given the rise in trend studies (the relevant analysis is included in Appendix I).

Regarding the point for negative values, this is now removed as mentioned in a previous response to Reviewer 1.

Section 4.4: why not analysing only Radcliffe and Prague? I appreciate the Korean record is longer, but it is slightly confusing to mix and match stations/indices

Agreed. The Prague station is now analysed instead and used as an example of the application of the methodology (Section 4.2.1: An examination of one of the longest records).

I understand that when fitting trends the authors use a least square approach (assuming a normal distribution for the likelihood part of the BIC calculations). OLS is known to be non-robust to outliers, and it could be that some of the spurious results are due to this. Do the slopes obtained with robust regression or Thiel-Sen estimators look any different than the Least Square slopes? How about a trend model on top of a AR(1) error component to take into account the temporal dependence that the authors show plays an important role?

We note that OLS is a theoretically established method and as we use it (without formal hypothesis tests and not relating it to BIC) neither distributional nor independence assumptions are involved. We have followed though the Reviewer's suggestions and applied robust regression techniques, namely least-absolute deviation and Theil-Sen estimator. These are presented in Appendix III. Prediction results are not improved by these techniques, if not got worse. A short discussion is included in the Appendix as well.

As for including a dependent error term, we would like to keep our models simple in order to render them relevant to the greater part of hydrological literature on the subject. The goal is to judge the predictive performance of a widely used model, and not to predict the climatic regime of rainfall. If we intended to do so, obviously we would have to refine our models substantially and employ probabilistic prediction as well.

Apart from the quantitative review of the Appendix I, in order to illustrate the established use of linear trends for rainfall extremes, we include here a selection of relevant papers examining *linear trends in rainfall extremes* published in reputable journals:

- Hänsel, S., Ustrnul, Z., Łupikasza, E. and Skalak, P., 2019. Assessing seasonal drought variations and trends over Central Europe. *Advances in Water Resources*, 127, pp.53-75.
- Papalexiou, S.M. and Montanari, A., 2019. Global and Regional Increase of Precipitation Extremes under Global Warming. *Water Resources Research*.
- McKitrick, R. and Christy, J., 2019. Assessing Changes in US Regional Precipitation on Multiple Time Scales. *Journal of Hydrology*, p.124074.

- Dai, A. and Bloecker, C.E., 2019. Impacts of internal variability on temperature and precipitation trends in large ensemble simulations by two climate models. *Climate dynamics*, 52(1-2), pp.289-306.
- Bishop, D.A., Williams, A.P. and Seager, R., 2019. Increased Fall Precipitation in the Southeastern United States Driven by Higher-Intensity, Frontal Precipitation. *Geophysical Research Letters*, 46(14), pp.8300-8309.
- Ajaaj, A.A., Mishra, A.K. and Khan, A.A., 2018. Urban and peri-urban precipitation and air temperature trends in mega cities of the world using multiple trend analysis methods. *Theoretical and applied climatology*, 132(1-2), pp.403-418.
- Zittis, G., 2018. Observed rainfall trends and precipitation uncertainty in the vicinity of the Mediterranean, Middle East and North Africa. *Theoretical and applied climatology*, 134(3-4), pp.1207-1230.
- Zhang, W., Brandt, M., Guichard, F., Tian, Q. and Fensholt, R., 2017. Using long-term daily satellite based rainfall data (1983–2015) to analyze spatio-temporal changes in the sahelian rainfall regime. *Journal of hydrology*, 550, pp.427-440.
- Donat, M.G., Lowry, A.L., Alexander, L.V., O'Gorman, P.A. and Maher, N., 2016. More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, 6(5), p.508.
- Asadieh, B. and Krakauer, N. Y.: Global trends in extreme precipitation: climate models versus observations, *Hydrol. Earth Syst. Sci.*, 19, 877–891, https://doi.org/10.5194/hess-19-877-2015, 2015.
- Zhang, X., Zwiers, F.W., Hegerl, G.C., Lambert, F.H., Gillett, N.P., Solomon, S., Stott, P.A. and Nozawa, T., 2007. Detection of human influence on twentieth-century precipitation trends. *Nature*, 448(7152), p.461.
- Wentz, F.J., Ricciardulli, L., Hilburn, K. and Mears, C., 2007. How much more rain will global warming bring?. *Science*, 317(5835), pp.233-235.

Note that we have not included these references in the paper because it is not our purpose to criticize specific papers.

We thank again the Reviewer for the constructive comments and helpful suggestions, most of which we have adopted and certainly, contributed to improving the manuscript.