Interactive comment on "A global scale evaluation of extreme events in the *eartH2Observe* project" *by* Toby R. Marthews et al.

Anonymous Referee #2

Received and published: 27 March 2019

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018- 622, 2019.

Review of Marthews et al. "A global scale evaluation of extreme events in the earth2observe project"
The authors use model simulations from the earth2observe project to study the sources of uncertainty in simulated runoff and evapotranspiration (ET). Model simulations from this project are well chosen for this purpose as they are performed with (i) different precipitation forcing datasets and (ii) different land surface and hydrological models. Analysing these simulations, the authors compared the relative importance of the pre- cipitation forcing uncertainty with that of the model uncertainty for resulting runoff and ET extremes.

Recommendation: I think the paper should be rejected.
While the research question is interesting and relevant, and the model simulations are well suited for the purpose of this study, the applied methodology is too complex and hard to understand such that I am not sure about the robustness of the resulting conclusions.

-- RESPONSE -- Although this is a disappointing recommendation, thank you very much for the review and we hope very much that you will consider the responses below and, hopefully, we can convince you of the merit of this paper and our results.

General comments:

(1) As mentioned above I do not understand (the purpose of) the methodology applied in this study, even after carefully reading it many times.

While the focus on extremes is not explained or motivated, I also do not see why/how 10% of a 14-year time series can already be considered extreme. Also, there is no indication to what extent the final conclusions depend in this arbitrary choice. Further, the definition of 'uncertainty' in extremes is only explained in the caption of Figure 2, and I wonder why such great complexity is needed after all. Why not simply analysing the very highest/lowest monthly precipitation, runoff and ET sums at each grid cell, across models and forcing datasets?

-- RESPONSE -- Firstly, we apologise that we hadn't fully motivated the approach that we took for this analysis. We have amended the first paragraph of the introduction to include reference to the IPCC Special Report on extreme events and have highlighted the importance of looking at these events.

Secondly, extreme events exist on a continuum so some kind of definition is always required in a study like this (heavy rainfall in the UK would be considered normal in the Philippines, etc.). It is very standard to choose 10% as a threshold (a Q10/Q90 method) for extreme events (e.g. "The Intergovernmental Panel on Climate Change (IPCC) suggests that "rare" means in the bottom 10% or top 10% of severity for a given event type in a given location" on https://www.encyclopedia.com/environment/energy-government-and-defense-magazines/extreme-weather) so we have added a reference to IPCC (2004) to Section 2.1 where we specify this (it was not a focus of this paper to try to quantify the uncertainty related to the choice of 10% here). The text clarifying our definition of uncertainty has been taken out of the legend to Fig. 2 and added as a sentence to Section 2.1 as well.

Finally, "simply analysing the very highest/lowest monthly precipitation ..." is unfortunately simply not appropriate in an analysis at the global level: precipitation distributions do not only change in terms of mean and variance from place to place, but also change in terms of the shape of the distribution, i.e. skewness and bimodality. In order to carry out an analysis that covers all biomes from rainforest to desert, as we have done here, we need to use statistical methods, and the techniques we

have used are no more complex than used in comparable studies: in fact, although the use of ensemble methods brings in some complexity, the actual basic stats involved is nothing more complicated than a standard deviation of occurrence data.

Moreover, it remains unclear if absolute values or anomalies (i.e. with removed sea- sonal cycle) are used. In the case of absolute values, high ET extremes will necessarily occur in summer and while this is not always the case for extreme precipitation, this would lead to a (unwanted) de-coupling of the variables with this analysis design.

-- RESPONSE -- We used neither absolute values nor anomalies: we have been clear throughout the paper that our analysis was based on *occurrence data*: in any particular gridcell we get the distribution of e.g. precipitation from MSWEP (which gives us a baseline) and then instead of considering an absolute value (e.g. 50 mm rainfall) or anomaly (e.g. 50 mm minus the mean for that gridcell), we compare to the normal distribution and note whether an extreme event has occurred (1 or 0). It is then the occurrence numbers that are analysed/averaged. We believe this is the best way to analyse data that comes from widely disparate biomes with differing distributions of precipitation, ET or runoff. The analysis was also deliberately carried out month by month (e.g. comparing to a baseline calculated from all the Februaries in the 14 year MSWEP dataset) in order to exclude any spurious matching of e.g. winter months to summer months, which accounts perfectly for the de-coupling mentioned here.

Concerning the low extremes, I am not sure how much sense this makes for precipi- tation. Lets say in a dry grid cell precipitation is zero in most of the analyzed months, does it make sense then to determine such months as low precipitation extremes?

-- RESPONSE -- Please note in Section 2.1 we state that we masked out all gridcells with extremely low rainfall exactly to exclude this possibility.

I do not agree with referring to MSWEP as a 'gold standard', and with statements like 'the best global

evapotranspiration products (Martens et al. 2017)' or 'simulation results from the earth2observe project [...] driven by the best available published precipitation observations'. While these products are certainly state-of-the-art, I doubt that they will be 'the best' (based on what measure?) in all regions and at all times. As for the reference precipitation used in this study, it could be a more fair alternative to use the ensemble mean across the considered precipitation products.

-- RESPONSE -- We stand by these statements: we use "best" to mean "best available product/model at the current time" (which we do not feel implies "best in all regions at all times"). We do not believe that it would be better to take a simple mean across all the considered precipitation products: they are not all comparable because they have varying levels of processing, as we have highlighted in Table 2. The high level of sophistication in the MSWEP reanalysis and the thought and consideration they have put into producing that product and correcting for observational problems that occur in other products we believe quite legitimately support our phrase "gold standard" (although please note this is our phrase and does not come from the MSWEP team).

(1) I think the linearity assumption made in Figures 4-6 is not justified, such that the linear regressions are no suitable way to analyze these point clouds. Further, displaying the point cloud envelopes is misleading, as these envelopes is likely dominated by outliers/extremes, and do not necessarily reflect actual relationships. Instead, why not use a 2D density plot here, and climate-regime-based moving average lines to summarize the results?

-- RESPONSE -- We initially did use 2D density plots here, but the extremely large number of points (and substantial overlap) served to obscure the message that we were trying to communicate with these figures. Although we do accept that displaying the envelopes draw attention away from mean values towards the extremes, we feel that in a paper focused on extreme event analysis that this is not an inappropriate approach to take.

We do accept that applying a linear fit to these data is simplistic, and a number of alternatives were experimented with during the course of the analysis we carried out in this paper. However, applying more

sophisticated methods did not seem to be legitimate given that the only conclusions we were drawing from these figures was whether or not the trend was an increase or a decrease moving from left to right. We certainly do not contend at any point that the distribution of points is linear in theory: we just included these lines to indicate the trend, which is not clear to the eye from the envelopes (because they don't show the point cloud) or from the point clouds themselves (because they overlap too much and would have had to have been separated into individual plots, which for space reasons we didn't want to do)

Specific comments:

- section 3.1, line 13, and caption of Figure 3, and elsewhere: the authors sometimes refer to 'increases' while also decreases are found in some regions -- RESPONSE -- We have checked these statements and they are correct: please note that when we say alpha "increased with precipitation", this means it correlates positively with precipitation, which is unrelated to areas of blue versus green on the associated maps in the same figure.
- epsilon is used twice, in section 2.1, line 21, and then in section 2.2, line 10 -- RESPONSE -- Thank you for spotting this! Corrected.
- section 2.1, line 22: '20 mm annual precipitation' - does this refer to multi-year means, or to individual years -- RESPONSE -- This is indeed the MSWEP multi-year mean (we have now added this information in parentheses - thanks)
- section 2.1, line 23: abbreviation SD not defined -- RESPONSE -- "standard deviation" has been added in
- section 2.1, line 25: replace 'runs' with 'simulations' -- RESPONSE -- Thank you for spotting this! Corrected (and one occurrence of "runs" in the discussion too)
- section 2.2, line 2: 'simulator' is not defined -- RESPONSE -- "simulator" replaced with "simulator model"
- section 2.2, line 18: I think here you mix up i with j (?) -- RESPONSE -- Thank you for spotting this! Corrected.
- results section, and figure captions: instead of using X as subscript and then referring to ET or runoff,

you could replace the X with Q or ET -- RESPONSE -- In an earlier draft we did try this, but the large number of "Q or ET"s that necessarily have to occur in the text we felt obscured the message we were trying to write.

- Figure 2: numbers on color bar are very small -- RESPONSE -- Colour bar size increased by 10%
- Figure 3, caption: you mention a 'run' here, but these are just precipitation products and no model simulations -- RESPONSE -- Thank you for spotting this! Corrected.
- Figures 4-7: legends missing -- RESPONSE -- We do state in the legends "Points on the scatter plots are coloured according to latitudinal zones (Fig. 1)", which we hope is sufficient and saves having Fig. 1 as an inset on each of these figures.