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

Editor Decision: Publish subject to revisions (further review by editor and referees) (07 Oct 2019) by Patricia Saco

Comments to the Author:

We have now received useful comments from two referees. Based on my own careful assessment of the revised paper and the response letter, I agree with the reviewers that there are still some aspects that need to be addressed before this manuscript can be considered for publication. Though the authors have provided detailed responses to the referee comments, the manuscript will be further improved by integrating some of the material of these responses into a revised submission.

Thank you very much for this feedback and we hope very much that our resubmission addresses all outstanding concerns in full.

In particular, the revised manuscript could be improved by addressing some of the concerns of the referees:

1) Please include a more clear justification of the methodology emphasizing its appropriateness and its advantages over using a simpler methodology (this will address one of the concerns of re viewer #2, but please note that lack of clarity was also pointed out by the comments reviewer #1 on the original submission).

We have given detail in our responses to the reviewers, but we believe that we have much improved the presentation and justification of the methodological approach with our additional text, our new Figure (Fig. 2) and the worked example that is included in that figure.

2) Please analyze/discuss the robustness of results as suggested by reviewer #2, or alternative possible limitations of the study.

Please see added text and specific responses to all the individual reviewer comments below (Reviewer #3 first, followed by Reviewer #2 further below - all given as track changes on the previous, interactive review on *HESSD*).

3) Regarding the linearity assumption made in Figures 4-6. It would be good to add a very short discussion (a couple of sentences) similar to that included in the authors response, to clearly state

that the intent is not to suggest that the points in the figures follow a linear trend.

Thank you for this and we have followed this advice, adding in a sentence to each figure legend (based on the form of words given by the reviewer) stating clearly that we did not intend to imply that the underlying processes here are linear.

Though these are some of the main concerns that need to be addressed, please note that it is important to address all the reviewers' comments as this will help improve the contribution.

Thank you very much for the opportunity to make a second response to these reviewer comments. We have revised all sections of the paper as well as the reviewer comments (and we have even rechecked our responses to the first original reviewer as well). We hope very much that with these changes this article might still be considered for publication in *HESS*.

Very many thanks for your patience with this article and for your helpful feedback at all points.

Best regards,

Toby Marthews, Eleanor Blyth, Alberto Martínez and Ted Veldkamp.

30th October 2019.

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

Anonymous Referee #3 [= Anonymous Referee #2 from the first round of review]

Received 7 October 2019
Second review of Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-622, 2019.

The authors have implemented some of my suggestions in the revision of the manuscript. However, my main concerns with the complex methodology and my corresponding doubts regarding the robustness of the results and conclusions have not been addressed satisfactorily.

Therefore, I still recommend rejection of the paper.

Firstly, an apology to this reviewer: we can see that this response was submitted to *HESS* on 27th June, but we only received it on 7th October after the second reviewer response had also been received. We have responded as quickly as possible to the concerns raised and we hope that the delay will not count against us in this case.

We thank this reviewer for having given our manuscript due consideration both for the first review and also in this re-review, and we appreciate greatly that he/she has put in a considerable amount of time in formulating these comments with the motivation of helping us to improve this manuscript. In the light of these further comments, we have made many changes to the text to accommodate the suggestions raised and we hope very much that the quality of the paper has now been raised up to what is expected from a *HESS* article.

General comments:

This is not an easy decision to reach for me as I still appreciate the interesting research question, and the unique and very suitable dataset which the authors investigate. However,

(1) As mentioned above, I feel that none of the 3 major comments from my first review have been satisfactorily addressed in the revision of the manuscript. And this is even though I recommended rejection, which should already indicate that these are serious shortcomings in my opinion.

(2) I realize that some of my criticism in comment (1), namely the latter points on absolute values versus anomalies, and on low precipitation extremes, was maybe not fully justified as this was explained in the manuscript.
Thanks to the authors for pointing me to the corresponding paragraphs in the responses. Nevertheless, such misunderstandings could have been used as a motivation to try to further clarify these issues in the manuscript. Further, my main concerns in comment (1),

the complex methodology and the definition of extreme events, has not been addressed in the revision process. The authors could not convince me of the necessity of the cumbersome methodology.
Also, Comments (2) and (3) have not been much addressed overall.

(3) Even if the authors, for different reasons, disagree with most of my main suggestions, they could have done some sensitivity testing to show that the conclusions are robust with respect to my concerns. For this purpose, they could have added some results obtained with alternative methodologies (or reference precipitation).

The major issues referred to in the previous review are:

(1) *Our use of a 10% threshold to define extreme events and the suggestion that our method was overall more complicated than justified by the data and objectives of our analysis*

We justified before our use of the 10% threshold (the accepted standard of the IPCC) and for method clarification we have inserted a whole new figure, supporting text and a worked example to justify our analysis approach (the new Fig. 2). We very much hope that this will be enough to convince this reviewer that our approach was indeed justified by our data and the objectives of our research as stated.

(2) *Referring to MSWEP as a 'gold standard' and to other comments about 'the best' evapotranspiration products, etc.*

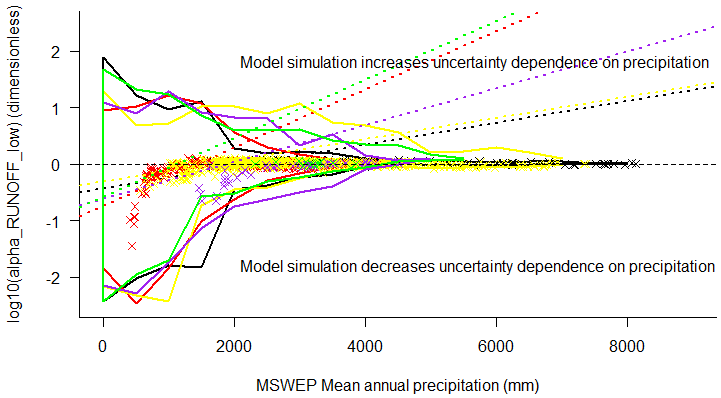
We have now removed all these statements about 'gold standard' and 'best'.

(3) *The linearity assumption made in our Figures 4-6 [now Figs. 5-7]*

We would like to emphasise that by applying regression lines to these plots we only intended to support the statement that there was a general trend in our data from left to right. The regression fit applied was not used in any subsequent analysis and was only intended for visualisation of the trend. One of the plots referred to by this reviewer comment is included below for clarity: this reviewer suggested to "use a 2D density plot here, and climate-regime-based moving average lines [instead of regression lines]", however the other reviewer suggested instead to leave the combination plot as it is and simply to "add a very short discussion (a couple of sentences) similar to that included in the authors' response, to clearly state that the intent is not to suggest that the figures suggest that the points follow a linear trend"

Having tried various options here, we would like to follow the second reviewer's suggestion for the following reasons: (i) firstly, space - each of Figs. 4,5,6 are currently composite 8-plot figures (with the first 4 plots of each being a scatter plot like the one below) so to split each scatter plot into 5 density plots (for each of the latitudinal zones of Fig. 1) would make each of Figs. 4,5,6 into a composite 24-plot figure and we believe that this would put us substantially beyond the length restrictions of *HESS* articles. Secondly, (ii) to replace the regression lines with moving-average lines, the righthand halves of these lines all disappear into the y=0 line and become indistinguishable and it is no longer possible to see the basic message communicated by the regression lines that there is a (statistically significant) general trend to the right. We would very much like to show more of our data in these plots by plotting the complete point clouds, but we believe that it is simply not

possible given the space constraints of this article, therefore we have opted instead to add in the sentence to each figure legend "Linear regression lines for each latitudinal zone indicate the trend as precipitation increases within each zone (all regressions were significant at the 1% level), although n.b. we do not contend in any way that the distribution of points shown is linear: these lines simply indicate a trend that is not clear to the eye from the envelopes displayed (which do not show the complete point cloud)" as requested by the second reviewer.



Instead of such actual changes/additions to the manuscript and/or the supplementary material, the authors are in many cases providing explanatory justification in their responses to my review comments.
And even in case of some comments which the authors decided not to consider as requested, they could still have included (some of) that justification into the manuscript such that all readers become aware of their arguments.

We like to believe that we have indeed now included the extra justification referred to here.

On a final point, we would like very strongly to thank this reviewer for raising all these issues. Including the smaller specific points at the end of the original review, we have benefited from a large number of well-chosen comments here and the manuscript has been changed throughout as a result of these insightful responses - many of which have brought up important aspects of the method and results that we failed to explain clearly in our original submission. Our manuscript is very much improved as a result of the raising of these concerns and it is now clear that we were indeed very much too brief on several parts of our method explanation. We apologise again that our lack of explanation caused this to be a more difficult paper to review than it could have been.

Best regards,

Toby Marthews *et al.*

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

Anonymous Referee #2 [reviewer RC2 on https://www.hydrol-earth-syst-sci-discuss.net/hess-2018-622/]

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 [now Fig. 3], 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 (2014) 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 [now Fig. 3] 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.

Second response: Making use of the suggestions given by the reviewer here (combined with similar comments on the same topic from the other reviewer), we have inserted a new figure Fig. 2 into the paper featuring a flowchart explanation of the quantities alpha, beta and epsilon, as well as a worked example of how these are calculated and combined in the paper. We have also added text to sections 2.1 and 2.2 that we hope explain much more clearly what we have done and why.

We hope now that the presentation of a worked example here makes it clear why we have had to consider quantities that, on the face of it, appear to be complicated: it is the specific format of the source data of our study that requires us to do so (assembled by the many collaborators of the eartH2Observe project), and the approach dividing clearly between data and model uncertainty recommended by our guiding textbook Oberkampf & Roy (2010).

When we submitted the paper we were very conscious of space limitations and therefore we only included the absolute minimum description of the method and analysis quantities used, however if we are allowed to have new Fig. 2 then we think it does outline the concepts of our uncertainty quantities and encapsulate their interrelationships in the briefest but unambiguous way possible.

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 [this point has now been emphasised more clearly in new Fig. 2]. 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 [this last sentence has now been added to section 2.1 to emphasise this point].

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.

(2) 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 --

Second response: We have now removed the reference to MSWEP as a "gold standard" and all occurrences of the word "best" in this context.

(3) I think the linearity assumption made in Figures 4-6 [now Figs. 5-7] 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% - now increased by a further 50% in the Supp Mat
- 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.

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

Anonymous Referee #2

Received 7 October 2019
Comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-622, 2019.

==========

I believe that the authors have addressed and/or provided a good rebuttal to the comments and concerns of the previous reviewers.

Thank you very much for this comment.

Though the authors have provided detailed responses to the referee comments, the manuscript would be further improved if some of these responses were added or discussed in the revised submission.

Thank you: we have indeed now added in a selection of the responses we made into the text itself where appropriate.

In particular, the revised manuscript could be improved by addressing some of the concerns of referees:
1. The selection of evaporation products. Please rephrase to clearly explain what you mean by best (i.e., not in all regions at all times), and I believe that "gold standard" is a bit of an overstatement and not needed in the context of the paper.

All references to "gold standard" and "best" have now been removed (see the other reviewer's General Concern #2).

2. The linearity assumption made in Figures 4-6. It would be good if the authors can add a very short discussion (a couple of sentences) similar to that included in the authors' response, to clearly state that the intent is not to suggest that the figures suggest that the points follow a linear trend.

Thank you for this comment and we refer to our response on the same point under the other reviewer's General Concern #3: we have followed exactly the advice given here.

Very many thanks for your time reviewing our manuscript: it is hugely appreciated.

Best regards,

Although not specifically required at this stage in the review, we have revisited our responses to Referee #1 below in the light of the second round of review responses and we would like to modify some of our responses below:

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

Anonymous Referee #1 [reviewer RC1 on https://www.hydrol-earth-syst-sci-discuss.net/hess-2018-622/]

Received and published: 25 February 2019
Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2018-622, 2019.

General comments The manuscript presents an analysis of a unique dataset that was produced from the eartH2Observe project. This dataset involves the simulation of sev- eral hydrologic variables from a number of state-of-art land surface/hydrologic models and using as forcing several satellite and reanalysis dataset. The scale of analysis is global and the focus is on the tails (i.e. low/high extremes) of evapotranspiration and surface runoff. Overall the work is very interesting and the dataset analyzed is very unique. Additionally, the fact that the analysis is performed at global scale provides im- portant information on the regional variability of findings. The manuscript is generally well written but there are certain parts (especially in the description of methodology) that require additional clarification and discussion. I provide some specific comments below that hopefully will help the authors to improve their manuscript.

-- RESPONSE -- Thank you very much for the review and we hope that the responses below are sufficient reply to these useful comments.

Specific comments

1. I believe that the title should be revised to better reflect the context of the paper. One of the main elements of you analysis is "uncertainty in identification of extreme events" but this is not reflected in the current title.
-- RESPONSE -- We agree and have added "uncertainty" to the title.

2. Abstract L17-19: I agree but given the focus of your analysis (i.e. identification of extremes) you should be more specific on what your results will allow to comment. For example, models can be quite robust in representing the main body of the distribution of hydrologic variables, which is actually very important for water resources applications. So I suggest to specifically refer again to representation of extremes.
-- RESPONSE -- We agree: we have now specified in the abstract that "Our results are important for highlighting the relative robustness of satellite products in the context of land

surface simulations *of extreme events*" as requested.

3. P3L3: For a multiregional evaluation of satellite precipitation over complex terrain, you may want to consider also Derin, Yagmur, et al. "Multiregional satellite precipitation products evaluation over complex terrain." Journal of Hydrometeorology 17.6 (2016): 1817-1836.
-- RESPONSE -- Thank you for this reference: it was indeed relevant and we have used it at two points in the revised text.

4. Information in Section 2.1 needs to further clarified. What do you mean by "base distribution"? Is this the reference for your EE/yr at each cell? Why you average the five runs and you don't consider each model independently? Do you repeat the same procedure for each product and then compare? Please clarify.
-- RESPONSE -- Thank you for pointing this out: we had not explained why a base distribution (or baseline distribution) is necessary. We have added to section 2.1 "Extremes for any particular variable may only be assessed in relation to an estimate of 'normal' conditions, and for this we took a baseline distribution of values ...". At this point we have not considered each model independently because we do that at a later point when comparing model uncertainty and data product uncertainty. Finally, yes we did exactly repeat the same procedure for each product (and each model) and compare: this is the basis of the definitions of alpha, epsilon and beta later on in the method.

5. It would be very useful to provide a graphical example to demonstrate the different uncertainty components that you describe in equations 1-3.
-- RESPONSE -- We have indeed included a graphical example of the different uncertainty components: this is Fig. 7 [now Fig. 8]. We have considered moving that figure to an earlier point in the paper, but it is difficult to do so because we need to explain more details of how it is calculated before presenting that figure.

Second response: Having revisited this response, we have come to the conclusion that it was not a fully adequate response and we apologise for this. The addition of Fig. 8 was good, we still feel, but it served only to clarify our results, not the description of our methodology. We have now also inserted a new Fig. 2 that presents the relationships between alpha, beta and epsilon in flowchart form and a worked example of how these quantities relate to each other. The definitions here are the product of a lot of thought based on methods outlined in Oberkampf & Roy (2010) and they are necessary for an uncertainty analysis carried out in a model ensemble context.

We apologise that something similar to Fig. 2 was not inserted into the paper before. Each step described there was actually implicit in the careful wording of our definitions of alpha, beta and epsilon before, however we do accept that we failed to draw attention to the full implications of these definitions for our analysis (e.g. the significance and interpretation of epsilon>1 and why a quantity like *DIU* cannot be used directly when considering more than one variable of differing units).

6. L24-28 are confusing. First, it is not clear why you consider $\epsilon_{x,j} > 1$ as an indicator of model amplification of uncertainty? Do you mean $\alpha_{x,j}$ instead? Also if you want to identify the relative contribution of the different sources of uncertainty, why don't you take the ratio of α/β ?

-- RESPONSE --

We have added some text to section 2.2 ("In summary ... product only") that we hope clarifies this point and the relationship between alpha, epsilon and beta.

Second response: With the insertion of the new Fig. 2, we hope that this has become a lot clearer and it is more visually presented now.

In response to the suggestion that alpha could be an indicator of model amplification rather than epsilon, we have inserted into the text new sentence "This augmentation comes from two sources: firstly, a model ensemble can produce outputs with higher sensitivity to input precipitation e.g. through a significant nonlinear relationship between *X* and precipitation in the majority of ensemble models (α), but it must not be forgotten that higher uncertainty in the outputs may also come from the differences in non-precipitation dependencies inside these models, which may also be larger in magnitude than *DIU* (β)", which we hope clarifies that alpha is only one aspect of model augmentation and in this analysis it is very important to consider both aspects: the influence of precipitation (which comes through in alpha), yes, but also the influence of non-precipitation factors (which generally come through in beta, as a dependence on 'choice of model'). Incidentally, this new sentence was itself borrowed (with thanks!) from the same reviewer's comment at #14 below where he/she brought up a similar point.

7. P6L6: "global average", why do you consider global average? It is not advisable since the average masks regional variability. Also "ET highs (58.1% vs 41.9%)", it is not clear what these numbers correspond to.

-- RESPONSE -- This is an introductory point at the start of section 3.1 which we then expand upon in more detail later. It is not irrelevant to point out that the alpha values are universally quite small and do seem to decline with increasing precipitation. the regional variability is displayed graphically in Figs. 4-6 [now Figs. 5-7].

8. P6L10 "$\alpha_{x,j}<-1$", I believe you mean $\log(\alpha_{x,j})$.

-- RESPONSE -- The reviewer was correct (many apologies!): The text has now been amended to "$\alpha_{X,j}<0.1$, $\log10(\alpha_{X,j})<-1$".

9. P6L19-23. Interesting findings, some additional comments are welcome here. For example, why "the magnitude of the increase reduced in wetter environments"?

-- RESPONSE -- We feel that it would be too speculative to include here any of the various theories that could explain why the magnitude of the increase is reduced in wetter environments. For example, there could be a saturation effect in the environment (but in the absence of soils or land use data we cannot be sure of this) or fast drainage could occur more often under more episodic rainfall (but we have no data on drainage patterns) or the

occurrence of convective cells might be very regionally specific (but these are not even visible on most remote sensing products). We have tried to be careful to stick to discussing points that are directly relevant to the results and data that we have presented and we hope in a later study to look at these trends in more detail, but we have omitted any discussion of this here.

10. P6L25: "The global mean value. . ..is a measure of variability". How can a mean value tell you anything about variability? Please clarify/revise.
-- RESPONSE -- If the quantity in question (alpha) is itself a measure of variability, then the mean of alpha will still be a measure of variability even though we agree it will not contain any information about the variability of alpha itself. We have revised the wording here to avoid the apparent contradiction.

11. P6L25-30: In general, this part of the text is quite difficult to "digest". Please improve clarity.
-- RESPONSE -- We agree and thanks: we have removed the middle sentence, which we hope has improved the clarity of the paragraph.

12. P6L31: What do you mean by "internal model uncertainty"?
-- RESPONSE -- We have added in the explanation that this is "a measure of the diversity of the calculation methods used to derive *X* between models".

13. P7L3-4: "...are more sensitive to precipitation extremes in wet environments". Be careful here, you should state "...more sensitive to precipitation uncertainty".
-- RESPONSE -- Corrected with thanks.

14. P7L15-16: I believe that there is a confusion here between model uncertainty and uncertainty propagation. This is a very important aspect and the authors should clarify it in their discussion. For example, even with zero model uncertainty, transformation of precipitation uncertainty to runoff uncertainty could potentially amplify as a result of the nonlinear transformation of rainfall-to-runoff.
-- RESPONSE -- We have specifically defined separate quantities for model uncertainty (beta) and uncertainty propagation (epsilon) and we believe that we have not confused the two issues in this paper: in fact, drawing attention to the difference is one of the overall points of the paper.

If runoff is generally 1000-1500 mm/yr with 7 peaks/yr when precipitation inputs are 500-1000 mm/yr with 3 peaks/year, then output uncertainty differs not only in terms of absolute value (which can be a linear effect) but also in terms of distribution (a nonlinear effect). By focusing our study on extreme event occurrence, linear effects should be cancelled out (as long as the extremes are calculated in terms of an appropriate baseline for each quantity, which we have done), however of course there will be nonlinear effects that can give nontrivial values to epsilon (and alpha) even in the case of beta=0 because the number of peaks may still change. At no point in the paper have we assumed that this will not happen: in fact, we have accounted for this in all analyses.

At P7L15-16 we have simply stated that model uncertainty is usually greater than data uncertainty. We believe that the reviewer here does not like the implication that when model uncertainty is small then data uncertainty must be even smaller, and it was certainly not our intention to imply that. We have modified the text to say "when a set of models is under consideration, model uncertainty is usually greater than data uncertainty". To avoid the same implication we have added "in a simulation ensemble" to the start of section 4.2 as well.

15. The same point as in 14(above) should be considered in the discussion of section 4.2 (e.g. L26-27).
-- RESPONSE -- Please see last point #14.

16. P9L10: ". . .to improve prediction of water cycle quantities". Ok I agree but the analysis presented has not done anything on the quantitative aspect. Perhaps revise to "improve prediction of water cycle extremes"?
-- RESPONSE -- Thank you for the suggestion: changed to "extremes"

17. Section 4.3. (L15-22). The text here is relevant to work that is evaluating uncer- tainty and compares against observations. However, this is not the scope of your work. You isolate (correctly) the forcing and model uncertainty by considering as reference a model/forcing combination.
-- RESPONSE -- These comments are made in a section entitled "4.3 Sources of unquantified uncertainty" and we state clearly in the preceding sentence that we could not analyse these kind of situations in our particular study given the data available. However, we find these issues to be entirely within scope of this study and, in fact, we would have been remiss not to have mentioned them. We hope very much in a follow-up study to find some way to tackle these sorts of issues and we believe it is entirely appropriate to have a brief mention of them here.

18. Fig2 [now Fig. 3]: What is (a) and what is (b). Also, some of the explanation on the calculation of results could be added to text in manuscript as well.
-- RESPONSE -- The legend stated "a. Uncertainty in precipitation extreme highs and b. Uncertainty in precipitation extreme lows", which perhaps was not clear because we did not use parentheses on the (a) and (b), so parentheses have been added in. We agree that the explanatory text in the legend was perhaps too long and was mostly superfluous because the calculation is already described in the main text (section 2.1) so we have now omitted it.

19. Fig3 [now Fig. 4]. Similar comment on the explanation.
-- RESPONSE -- Unlike for Fig. 2 [now Fig. 3], the description of the calculation for Fig. 3 [now Fig. 4] is not repeated in the main text, but after reconsidering the legend we would like to argue that the amount of detail here is appropriate: the explanation here simply describes what the rows and columns of this multi-panel plot display and we do not see any way to abbreviate this without forcing the reader to hunt through the text for this description. Therefore we have left this text as it is and we hope the reviewer will either reonsider this comment or specify more precisely what change he/she would like us to make, please?.

20. Figure 4 [now Fig. 5]. I find this map very useful. It would be nice to provide for the other cases analyzed.
-- RESPONSE -- The other 3*3=9 maps from Figs. 4, 5 and 6 [now Figs. 5-7] were excluded before simply from space considerations. They have now been added.

21. Figure 7 [now Fig. 8]: "erros bars show SE". Do you mean standard error? And how the error is defined. Perhaps you refer to standard deviation instead?
-- RESPONSE -- We have left this text as it is: what was calculated here was standard error, which differs from standard deviation because you divide by the square root of sample size (the abbreviation SE is standard). The "averaged over 50°S to 50°N" earlier in the legend makes it clear that this is calculated across gridcells rather than time (i.e. sample size is the number of gridcells in this case).

A global scale evaluation of extreme event uncertainty in the *eartH2Observe* project

Toby R. Marthews[1], Eleanor M. Blyth[1], Alberto Martínez-de la Torre[1], Ted I. E. Veldkamp[2]

[1]Centre for Ecology & Hydrology, Maclean Building, Wallingford OX10 8BB, U.K.
[2]Institute for Environmental Studies, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands

Correspondence to: Toby R. Marthews (tobmar@ceh.ac.uk)

Abstract. Knowledge of how uncertainty propagates through a hydrological land surface modelling sequence is of crucial importance in the identification and characterisation of system weaknesses in the prediction of droughts and floods at global scale. We evaluated the performance of five state-of-the-art global hydrological and land surface models in the context of modelling extreme conditions (drought and flood). Uncertainty was apportioned between model used (model skill) and also the satellite-based precipitation products used to drive the simulations (forcing data variability) for extreme values of precipitation, surface runoff and evaporation. We found in general that model simulations acted to augment uncertainty rather than reduce it. In percentage terms, the increase in uncertainty was most often less than the magnitude of the input data uncertainty, but of comparable magnitude in many environments. Uncertainty in predictions of evapotranspiration lows (drought) in dry environments was especially high, indicating that these circumstances are a weak point in current modelling system approaches. We also found that high data and model uncertainty points for both ET lows and runoff lows were disproportionately concentrated in the equatorial and southern tropics. Our results are important for highlighting the relative robustness of satellite products in the context of land surface simulations of extreme events and identifying areas where improvements may be made in the consistency of simulation models.

1 Introduction

Producing robust predictions about the future dynamics of the water cycle at local, regional and global scales is critically important because it is the only way to avoid or mitigate the effects of water cycle extremes (e.g. flood, drought) (IPCC, 2012) and, in the longer term, to improve our use of resources and achieve long-term adaptation to climate change (Bierkens, 2015). Over the 21st century, climate and hydrological regimes are predicted to undergo significant shifts in baseline variables such as temperature, precipitation and runoff, leading to changes in the frequency of extremes of precipitation, evaporation and overland flow, and ultimately to changes in the frequency and intensity of both floods and droughts (Bierkens, 2015; Dadson et al., 2017; Marthews et al., 2019; Prudhomme et al., 2014). Understanding and predicting these shifts in the global dynamical system, both at atmospheric and land surface level is therefore of crucial importance (Santanello *et al.* 2018).

All model predictions have uncertainties, and linked modelling sequences have identifiable uncertainties at each step in the sequence (uncertainty propagation). In the case of a hydrological land surface modelling sequence, where climate data inputs are used to drive a simulator of the surface water cycle and land surface interactions, there are two main sources of uncertainty: *data uncertainty* (differences between forcing data used) and *model uncertainty* (differences between the simulation models). Data and model uncertainty differ greatly not just between themselves at particular locations, but also between coastal and floodplain areas of the world, and remote regions with heterogeneous terrain (Bhuiyan et al., 2018a; Riley et al., 2017) and between extreme high flows (floods) (Mehran and AghaKouchak, 2014; Nikolopoulos et al., 2016) and extreme water scarcity (droughts) (Veldkamp and Ward, 2015).

We focus on the relative dominance of model uncertainty (we take this as a broadly defined measure, including uncertainty from hydrology models that simulate water dynamics, vegetation models that focus on carbon dynamics and land surface models that attempt to integrate all biogeochemical cycles) and uncertainty in the precipitation product used to drive those models. In situations where model uncertainty is significant, the range of predictions possible from standard model simulations is of great importance to stakeholders and other users. If precipitation data uncertainty dominates, however, then greater attention should arguably be focused on selecting the most appropriate product to use, and perhaps additionally on interrogating the potentially sparse data base of precipitation measuring stations used by the precipitation products.

1.1 Uncertainties in land surface model simulations

Model uncertainty, i.e. prediction variation as a result of differing process representations within a model (e.g. Li and Wu (2006)), is commonly the dominant uncertainty in complex systems used in risk-informed decision-making (Oberkampf and Roy, 2010). Although historically often overlooked (Li and Wu, 2006), model uncertainty has recently come under increasing scrutiny in the context of land surface models (Huntingford et al., 2013; Long et al., 2014; Schewe et al., 2014; Ukkola et al., 2016). A lack of adequate representation of flood-generation processes (both from surface and subsurface runoff) and permafrost or snow dynamics can lead to an imprecise simulation of runoff peaks in many large river basins, and a lack of proper representation of wetland evaporation and human effects such as water consumption and inter-basin transfers can lead to over- or under-estimated discharge in many basins, especially those with large semiarid regions (Bierkens, 2015; Veldkamp et al., 2018). Additionally, even though regional-scale precipitation is predominantly caused by the atmospheric moisture convergence associated with large-scale and mesoscale circulations, processes operating on smaller length scales significantly modify even regional-scale dynamics, so it is to be expected that uncertainty in land surface models will depend on local topography, the presence or absence of vegetation or water bodies and, importantly, which type of precipitation is dominant at a particular point and time (cyclonic, orographic or convective, Table 1).

1.2 Uncertainties in precipitation products

Precipitation is a necessary forcing input for land surface and hydrological models that is extremely challenging to estimate independently (Beck et al., 2017b; Bhuiyan et al., 2018a; Bhuiyan et al., 2018b; Levizzani et al., 2018). The accuracy and

precision of precipitation measurements fundamentally influences predictions of land surface and hydrological models (Hirpa et al., 2016), however many widely-used precipitation products have high uncertainties over the tropics and/or areas of high relief (Bierkens, 2015; Derin et al., 2016; Kimani et al., 2017; Yin et al., 2015).

High precipitation extremes are not always well-characterised: Mehran and AghaKouchak (2014) reviewed the capabilities of satellite precipitation datasets to estimate heavy precipitation rates at different temporal accumulations. For example, the precipitation radar on board TRMM (Table 2) is capable of capturing moderate to heavy precipitation but does not detect light rain or drizzle (Huffman et al., 2007; Luo et al., 2017).

Low precipitation extremes are also not always well-characterised: Veldkamp and Ward (2015) reviewed the advantages of different drought indices and highlighted many issues at the global scale. This relates to a more general point about remote sensing rainfall intensity: a precipitation product is more likely to record correctly that it is raining at a particular location than to record correctly the amount, which is unfortunate because it is usually precipitation amount that is most important for predictive modelling of drought or flood intensity.

Accuracy of meteorological data including precipitation will be expected to be lower (and uncertainty higher) for 'real-time' precipitation products because they have not been 'blended' with raingauge or reanalysis data (Table 2) (Munier et al., 2018). If a near-real time estimate of drought or flood is needed, therefore, then a cost-benefit balance arises with the end user having to make a choice between up-to-date information versus lowest uncertainty (Munier et al., 2018).

1.3 The *eartH2Observe* project

During 2014-2018, the *eartH2Observe* project http://www.eartH2Observe.eu/ brought together a multinational team of modelling and Earth Observation (EO) researchers to improve the assessment of global water resources through the integration of new datasets and modelling techniques. The uncertainties described above for different parts of the forcing data - land surface model system have been the starting point for this investigation, and *eartH2Observe* has quantified these uncertainties using an ensemble of forcing data and modelling systems. The project aimed to provide an overall understanding of the uncertainty in the EO products and EO-driven water resources models. This understanding is needed for optimal data-model integration and for water resources reanalysis, and their use for basin scale and end-user applications (e.g. floods, droughts, basin water budgets, stream flow simulations) (Nikolopoulos et al., 2016). As part of *eartH2Observe*, and in order to make progress towards this aim, in this study we asked the following two research questions:

(1) Under what circumstances can uncertainty in the prediction of water cycle quantities be attributed clearly to the model in use (model uncertainty) and/or to the precipitation product used to drive the model (data uncertainty)?

(2) When uncertainty is attributable to both model and data sources, is data uncertainty generally the greater (i.e. the model contributes less than 50% of total~~acts to reduce or 'stabilise'~~ uncertainty) or the lesser ~~(i.e. the model effectively augments variation)~~?

2 Data and methods

Uncertainty in extreme event representation varies both between models used (model uncertainty) and also between satellite-based precipitation products used to drive the simulations (data uncertainty). Five of the most widely-used and well-supported precipitation data products were used in this study (Table 2) and five state-of-the-art land surface models and hydrological models were run using each of those forcing data products (Table 3). This produced an ensemble of 25 estimates for each output variable ~~(mirroring the *method of competing models* approach advocated for complex systems by Oberkampf and Roy (2010))~~.

Only the precipitation forcing data for each model were allowed to vary between simulations: the remaining non-precipitation drivers (temperature, wind speed, radiation, etc.) were held constant across all simulations and taken from global Water Resources Reanalysis 2 baseline forcing data used in other *eartH2Observe* projects (WRR2) (Arduini et al., 2017). The combination of WRR2 non-precipitation drivers and the selected precipitation drivers (Table 2) is called WRR-ENSEMBLE (Arduini et al., 2017). All simulations used a global spatial resolution of 0.25° and covered the period 2000-2013. Because of source data limitations (Table 2), we restricted our analysis to latitudinal zones between 50°S and 50°N (Fig. 1).

2.1 Focus on extremes

Performance was assessed in terms of the variability of evapotranspiration (ET) and surface runoff under extreme rainfall conditions (both high extremes and low extremes). We quantified the relative magnitudes of these uncertainties under (i) varying simulation model (model uncertainty) and (ii) varying choice of precipitation product (data uncertainty). We quantified uncertainty in terms of the number of extreme events per month, with the *extreme event* defined as the occurrence of an extreme value for the monthly average of a given variable, and *extreme* defined as a value in the top/bottom 10% of the baseline distribution of values for that variable (following IPCC (2014)). Extreme event probability was calculated within each pixel for each month of the year, summed over the year and then the standard deviation (SD) taken across either the model outputs or precipitation products ~~(see below)~~ in units of (occurrence of extreme events per year). In order to avoid spurious extremes occurring in deserts and other areas with very low variability in water cycle values, gridcells with less than 20 mm annual precipitation (multi-year mean) or <0.1 SD in their monthly precipitation across the year were excluded.

Extremes for any particular variable may only be assessed in relation to an estimate of 'normal' conditions, and for this we took a baseline distribution of values calculated at each gridcell (i.e. not globally, regionally or per biome) from an average of the five simulations involving the 2000-2013 MSWEP forcing data (Beck et al., 2017a). We took MSWEP to be our baseline ~~'gold standard'~~ product ~~in this sense~~ because of its high reliability and multi-source nature (satellite observations blended with reanalysis and gauge data, Beck et al. (2017a), Munier et al. (2018)) in comparison to other available products (Table 2). Carrying out the analysis on a month-by-month (e.g. comparing to a baseline calculated from all the Februaries in the MSWEP dataset) excludes spurious matching in any gridcell of e.g. winter months to summer months.

2.2 Uncertainty propagation

We defined three indices of uncertainty propagation α, β and ε (Fig. 2). These indices ~~to~~ quantify the extent to which a given simulation~~land surface or hydrological~~ model increases or augments the uncertainty introduced to its simulations via the precipitation driver inputs ~~(we consider water cycle variables only in this analysis so it is reasonable to assume that uncertainty in our variables is not independent of uncertainty in precipitation)~~. The α measure quantifies the increase or decrease in uncertainty attributable to the precipitation drivers, β measures the equivalent for uncertainty attributable to the simulator model itself and ε quantifies the overall change in uncertainty over the course of the simulation (Fig. 2). Note that the quantification of absolute uncertainty in predicted quantities (Li and Wu, 2006) is not our focus: we are instead concerned with the relative contributions of data and model uncertainty in a combination setting (Oberkampf and Roy, 2010).

The defining equations are (calculated on a gridcell by gridcell basis):

$$\text{Scaled data uncertainty } \alpha_{X,j} = DOU \div DIU \quad (1)$$

$$\text{Scaled model uncertainty } \beta_{X,j} = MU \div DIU \quad (2)$$

$$\text{Scaled total uncertainty } \varepsilon_{X,j} = \alpha_{X,j} + \beta_{X,j} = (DOU + MU) \div DIU \quad (3)$$

where DIU = Mean uncertainty across products in precipitation extreme occurrence (input forcing data uncertainty)

DOU = Mean uncertainty across products in variable X extreme occurrence (output model uncertainty attributable to forcing data input)

MU = Mean uncertainty across models in variable X extreme occurrence (output model uncertainty attributable to model differences)

All mean uncertainties are in units of (extreme event occurrence frequency per year: EE/yr hereafter) and j can be either *high* or *low* depending on whether high or low extremes are being considered. The uncertainty propagation involves input uncertainty from the precipitation driver (DIU), which under the simulation is modified into the uncertainty of X when averaged across the different results obtained from using different precipitation products (DOU), but, unlike the forcing data, the simulation results have uncertainty as a consequence of the differences between simulator model used (MU) which means that total uncertainty at output level is (DOU+MU) (Fig. 2).

In summary, $\varepsilon_{X,j}$ may be understood as a measure of how much input precipitation product data uncertainty (DIU) is amplified into output uncertainty (DOU+MU) during an ensemble of simulations. ~~$\alpha_{X,j}$ may be understood as the special case of $\varepsilon_{X,j}$ where the ensemble consists of one model only, and $\beta_{X,j}$ as the special case of $\varepsilon_{X,j}$ where all ensemble members use one precipitation product only. Values of $\varepsilon_{X,j}$ > 1.0 indicate that the model simulation acts effectively to increase (amplify) the uncertainty in the forcing precipitation data.~~ Note that it is possible for (DOU+MU) to be less than DIU (i.e. to have~~Values~~ $0.0 < \varepsilon_{X,j} < 1.0$~~) indicate that the model simulation acts to decrease (stabilise) the uncertainty in the forcing precipitation data (i.e.~~

~~the model acts effectively to 'stabilise' the input uncertainty to ($\varepsilon_{X,j}$*100)% of the input data uncertainty)~~, which will occur if we have models that are broadly similar in output (i.e. similar columns in the table of Fig. 2) and also little variability in the responses of those models to different levels of precipitation and/or precipitation correlates (i.e. similar rows). This may be interpreted as the ensemble models 'stabilising' the input uncertainty *DIU* to a lower amount of uncertainty in the outputs (*DOU*+*MU*) and reinforces the interpretation of ε as a measure of the 'augmentation' of input uncertainty as a result of model calculations.. This augmentation comes from two sources: firstly, a model ensemble can produce outputs with higher sensitivity to input precipitation e.g. through a significant nonlinear relationship between *X* and precipitation in the majority of ensemble models (α), but it must not be forgotten that higher uncertainty in the outputs may also come from the differences in non-precipitation dependencies inside these models, which may also be larger in magnitude than *DIU* (β). Division by zero in the case *DIU*=0.0 will~~should~~ not occur because of the masking to avoid spurious ~~'~~extremes~~'~~ in arid areas (above).

3 Results

Comparison of precipitation extreme event occurrences across the forcing precipitation products shows immediate differences both spatially (~~Fig. 2~~Fig. 3) and between the products themselves (~~Fig. 3~~Fig. 4). Notably, the precipitation products differ in their extreme event occurrence rates, with especially TRMM-RT presenting increased rates of extreme high precipitation events across the globe and particularly GSMaP presenting increased rates of extreme low events (for uncertainty maps, see Fig. S1, Fig. S2, Fig. S3 and Fig. S4). Calculating these absolute uncertainty values is a necessary step towards assessing the relative magnitudes of data and model uncertainty for different extreme events.

3.1 Scaled uncertainty

Considering firstly $\alpha_{X,j}$, the uncertainty that is directly attributable to the precipitation data products, we found that in terms of global average $\alpha_{X,j}$ was mostly <1 (i.e. $\log_{10}(\alpha_{X,j})<0$) for ET highs (58.1% *vs.* 41.9%) and decreased as precipitation increased in all latitudinal zones except the northern tropics, but for runoff highs, $\alpha_{X,j}$ increased with precipitation in all latitudinal zones except the equatorial tropics (~~Fig. 4~~Fig. 5). Points where data uncertainty greatly increased on propagation through models ($\alpha_{X,j}>1$) occurred mostly during the prediction of low extremes (ET or runoff) and were restricted to areas with rainfall <2000 mm/yr (~~Fig. 4~~Fig. 5). Points where data uncertainty greatly decreased on propagation through models ($\alpha_{X,j}<0.1$, $\log_{10}(\alpha_{X,j})<-1$) occurred mostly during the prediction of runoff extremes (mostly low extremes, but also high) and were restricted to areas with rainfall <1000 mm/yr (~~Fig. 4~~Fig. 5). Points with high precipitation uncertainty occurred in both dry and wet environments.

Considering $\beta_{X,j}$, the increase in model uncertainty relative to input data uncertainty, we found that $\beta_{X,j}$ was dominantly <1 (i.e. $\log_{10}(\beta_{X,j})<0$) for ET highs (80.1% *vs.* 19.8%) and decreased as precipitation increased in all latitudinal zones; for runoff highs, $\beta_{X,j}$ was also mostly <1 (55.6% *vs.* 44.4%) but increased with precipitation in all latitudinal zones except the equatorial tropics (~~Fig. 5~~Fig. 6).

The scaled increase in total (data + model) uncertainty is measured by $\varepsilon_{X,j}$. In all latitude zones except the northern tropics, we found that uncertainty in ET highs increased over the course of the simulation ($\varepsilon_{X,j}$ was dominantly >1 - i.e. $\log_{10}(\varepsilon_{X,j})>0$) at the great majority of locations (80.5% *vs.* 19.5%), though the magnitude of the increase reduced in wetter environments (~~Fig. 6~~Fig. 7). In all latitude zones except the equatorial tropics, we also found that uncertainty in runoff highs increased over the course of the simulation at the great majority of locations (76.2% *vs.* 23.8%), but for runoff the magnitude increased with precipitation (~~Fig. 6~~Fig. 7). This implies that the causes of higher model uncertainty operate differentially in wet and dry environments, with dry environments being perhaps generally less well-modelled than wetter environments.

3.2 Global uncertainty

The global mean value of α is a measure of the amount a given quantity is affected as precipitation changes relative to the input precipitation data uncertainty (Eq. 1). For quantities that 'track precipitation' ~~(i.e. are sensitive to precipitation extremes)~~, we would expect this to be close to 1 (e.g. runoff values, ~~Fig. 7~~Fig. 8a), but especially in drier climates small variations in precipitation can drive much higher variation in output variables through threshold effects, so we might expect higher values in such regions (e.g. ET values, ~~Fig. 7~~Fig. 8b).

The global mean value of β_X is a measure of the internal model uncertainty in quantity X, relative to the input precipitation data uncertainty (Eq. 2), i.e. a measure of the diversity of the calculation methods used to derive X between models. If quantity X is equally sensitive to precipitation extremes across models, we should expect low model uncertainty and therefore low values of β_X (e.g. under conditions where evapotranspiration and soil storage are minimal we would expect runoff highs and lows to be closely similar to precipitation highs and lows with the model introducing little modification of the input data). Our results show that evapotranspiration extremes are more sensitive to precipitation uncertainty in wet environments than dry environments (~~Fig. 7~~Fig. 8c).

Globally, model uncertainty was generally less than data uncertainty (~~Fig. 5~~Fig. 6, ~~Fig. 7~~Fig. 8). In the equatorial tropics, ET prediction uncertainty was more attributable to data uncertainty, but runoff uncertainty was more attributable to model uncertainty, either indicating a wider variety of model representations of runoff generation processes within the tested models, or a greater dependence of ET estimates on precipitation inputs (~~Fig. 5~~Fig. 6).

Munier et al. (2018) found that the occurrence of flood (high runoff values) is generally more sensitive to high precipitation extremes than the occurrence of high evapotranspiration values, but that the reverse is true for low extremes. We do find this in our results as a rule of thumb across all environments (e.g. ($\varepsilon_{ET,high}<\varepsilon_{runoff,high}$) and ($\varepsilon_{ET,low}>\varepsilon_{runoff,low}$) and the same for α and β in ~~Fig. 7~~Fig. 8a), but we also note that in very dry and very wet environments this pattern does not persist (~~Fig. 7~~Fig. 8) and it also does not persist in all latitudinal zones when taken separately.

The total change in uncertainty over the course of the simulation of variable X is measured by $\varepsilon_{X,j}$ (Eq. 3) and our values for $\varepsilon_{X,j}$ were universally >1.0, indicating that the model simulation does act effectively to increase (amplify) the uncertainty in the forcing precipitation data. This also implies that when a set of models is under consideration, model uncertainty is usually greater than data uncertainty. Finally, high uncertainty points for ET lows and runoff lows were

disproportionately concentrated in the equatorial and southern tropics not only for $\varepsilon_{X,j}$ but also for both components $\alpha_{X,j}$ and $\beta_{X,j}$ (~~Fig. 4~~Fig. 5, ~~Fig. 5~~Fig. 6 and ~~Fig. 6~~Fig. 7; cf. ~~Fig. 2~~Fig. 3).

4 Discussion

Model output uncertainty is always a mixture of input data uncertainty and uncertainty accumulated during the simulation (Li and Wu, 2006; Oberkampf and Roy, 2010; Van Loon, 2015). However, these uncertainties are ~~unfortunately~~ not orthogonal in general because the models encode nonlinear relationships and therefore cannot be assumed to react consistently to different levels of precipitation input (e.g. (Bhuiyan et al., 2018a; Munier et al., 2018; Ukkola et al., 2016)). In this study we have had unprecedented access through the *eartH2Observe* project to an ensemble of simulations that has combined a selection of widely-used and validated precipitation data products with a spread of cutting edge land surface and hydrology simulation models.

4.1 Clear attribution of uncertainty to data and/or model sources

Under what circumstances can uncertainty in the prediction of water cycle quantities be attributed clearly to the model in use (model uncertainty) and/or to the precipitation product used to drive the model (data uncertainty)? Ukkola et al. (2016) found that land surface models diverged in evapotranspiration prediction during the dry season, and the results of our study strongly support this conclusion, with our calculated envelope of uncertainty widening in drier climates across the globe for all our uncertainty measures.

We found that high data and model uncertainty points for both ET lows and runoff lows were disproportionately concentrated in the equatorial and southern tropics. These zones are dominantly covered by tropical rainforests and savanna grasslands, so one possibility is that low fluxes in xeric environments are better characterised - both in data products and model characterisation - than low fluxes in these mesic and hydric environments. Data products are known to be more accurate away from areas with consistent cloud cover and a high occurrence of convective rainfall (Table 1) (Derin et al., 2016; Levizzani et al., 2018), which might explain this for data uncertainty, but having model uncertainty follow the same geographic distribution indicates that we must also consider uncertainties in the calculations of runoff and evapotranspiration. It seems also to be the case that the simple water balance approach taken by land surface and hydrology models becomes approximate in latitudinal zones where low flows are generally combined with higher temperatures and more episodic rainfall events (McGregor and Nieuwolt, 1998). This could indicate that using generalised approaches for all environments (e.g. the Priestley-Taylor or Penman-Monteith equations) is no longer sufficient for simulations at these spatio-temporal scales (Long et al., 2014; Wartenburger et al., 2018) or perhaps because we still lack crucial processes in these models, e.g. soil crusting or sealing, which only occur in semi-arid or arid areas (Marshall et al., 1996). However, we must also be careful to draw strong conclusions from these zones because another possibility is that this result simply confirms that these regions are where our available sources data are of lower quality (q.v. ~~Fig. 2~~Fig. 3a).

Uncertainty in predictions of evapotranspiration lows (drought) in dry environments is especially high, indicating that these circumstances are a weak point in current modelling approaches. Importantly, our results quantify this effect and show that even though uncertainty in the precipitation inputs is highest in these environments, the uncertainty in model representation of the processes involved is also significant and should not be ignored. A practical application of this is that when robust predictions of drought are required in very dry environments, not only should a spread of precipitation products be applied, but also more than one simulator model, and the model outputs should be validated as closely as possible against local data sources in order to ensure that conclusions drawn from these analyses are suitable for decision-making.

4.2 Relative importance of data and model uncertainty

When uncertainty is attributable to both model and data sources in a simulation ensemble, is data uncertainty generally the greater ~~(i.e. the model acts to reduce or 'stabilise' uncertainty)~~ or the lesser ~~(i.e. the model effectively augments variation)~~? In a report for the Intergovernmental Panel on Climate Change (IPCC), Bates et al. (2008) drew attention to the high uncertainty there was in climate models in precipitation data (= *data uncertainty*), and also suggested that for aspects of the hydrological cycle such as changes in evaporation, soil moisture and runoff, the relative spread in projections (= *total uncertainty*) was similar to, or larger than, the changes in precipitation (points echoed later by Schewe et al. (2014) and others). Precipitation observations are known to have high uncertainty (Beck et al., 2017a; Bierkens, 2015; Kimani et al., 2017; Levizzani et al., 2018; Yin et al., 2015), but responses to precipitation low extremes (drought) should not be expected to be proportional to responses from the same model to precipitation high extremes (flood) (Veldkamp et al., 2018).

We found in general that the model simulations we analysed acted to augment uncertainty rather than reduce it. In percentage terms, the increase in uncertainty was most often less than the magnitude of the input data uncertainty, but uncertainty did not decrease through the model for any variable so the simulation models did not in any case act to 'stabilise' or decrease the uncertainty supplied to them through the precipitation data products used to drive them. We do agree with Wartenburger et al. (2018)'s finding that the forcing (data uncertainty) generally dominates the variance in ET extremes, but we found model uncertainty to be important in all cases analysed and very nearly the magnitude of the forcing uncertainty in both very dry and very wet environments. This is a very significant result because it implies that a focus on the reduction of both data and model uncertainty will be necessary in order to improve the prediction of water cycle extremes.

4.3 Sources of unquantified uncertainty

It is important to bear in mind that some sources of uncertainty exist in these water cycle quantities that are as yet unmeasured in any existing data products, and therefore cannot be analysed in this study. There is a very strong current emphasis in climate science on identifying global areas of high precipitation uncertainty, for example (Bierkens, 2015; He et al., 2017; Levizzani et al., 2018), from which we can highlight two uncertainty sources: Firstly, most precipitation products record observations of amount, not the type of precipitation (Table 2), however it is very likely that precipitation type strongly influences our precipitation data uncertainty: for example, convective processes are dominant in the precipitation generating processes in

dryland ecosystems (Table 1), and different precipitation types occur at different spatial scales as well (Table 1). Secondly, our equatorial tropical zone (Fig. 1) includes the tropical rain belt (also known as the Inter-Tropical Convergence Zone, ~~(~~ITCZ) of low pressure, characterised by convective activity generating many storms. It is well-known that because of the transitory nature of the cloud dynamics in the rain belt~~ITCZ~~, precipitation products necessarily have higher uncertainty and, simultaneously, these conditions are of too short duration to be captured reliably in our analysis (Marthews et al., 2019).

For evapotranspiration in particular, Lopez et al. (2017) drew attention to the global lack of high quality *in situ* site data and the "inevitable scale mismatch" when using such data to calibrate Earth Observation datasets. Regional estimates of evapotranspiration rely on scaling-up methods to take account of regional advection effects and, additionally, the use of estimated values for evaporation rates from unmeasured land use types. Each step in these calculations potentially introduces significant uncertainty with the result that there is currently wide variation between the values suggested by various~~even the best~~ global evapotranspiration products (Martens et al., 2017).

Finally, runoff: Surface runoff estimates are linked to precipitation and evapotranspiration estimates via the water cycle balance equation (Beck et al., 2017b; Bierkens, 2015; Veldkamp et al., 2018). Because soil storage terms are usually taken as constant, underestimation of evapotranspiration often means overestimation of runoff and streamflow data (and *vice versa*). In this way, uncertainty in surface runoff is related to uncertainty in evapotranspiration estimates. However, because of the wide availability and high quality of global streamflow datasets (e.g. the Global Runoff Database, GRDC), and a much lower requirement for approximation and gap-filling in comparison to evapotranspiration data, runoff data is usually considered to be of the highest quality in water balance studies.

4.4 Conclusions

Water resources management has become one of the most important challenges facing hydrologists and decision-makers at state and national levels, motivated by increasing water scarcity in some global regions and a higher frequency of extreme flood events in others (Bierkens, 2015; Dadson et al., 2017; Schewe et al., 2014). At the same time, precipitation extremes are predicted to increase in frequency and impact under committed climate change (Ali and Mishra, 2017). Therefore, reliance on robust model predictions has never been greater (Kundzewicz and Stakhiv, 2010; Riley et al., 2017). In this study we have used an ensemble of simulation results from the *eartH2Observe* project derived from cutting-edge model simulators driven by a wide variety of~~the best available published (and validated)~~ precipitation observations, but the sources of uncertainty are nevertheless many and varied.

We found that models ~~always~~ augmented uncertainty relative to the magnitude of forcing data uncertainty at the great majority of spatial points, and therefore always did so in terms of global average uncertainty. Although, for predicting the extremes of evapotranspiration and runoff, the uncertainties inherent in the current generation of precipitation observation products are generally larger than the uncertainty introduced into the calculation by the land surface and hydrology models used, model uncertainty cannot be ignored and in many environments is comparable in magnitude to forcing data uncertainty. Therefore, in order to reduce prediction uncertainty we need very much to make progress on two fronts: (1) we need

precipitation data product uncertainty to be reduced (improved satellites are always welcome, of course, but we believe that much progress can also be made through moving towards blended products that are sensitive to more types of precipitation) and (2) we need to improve the mechanistic equations used in these models to derive water cycle quantities (including a better consideration of scale issues and domains of validity for existing equations).

It is important to resolve both data and model uncertainty much more clearly and identify exactly at which points in our linked modelling systems these uncertainties become the most significant. Our current model representation of land surface hydrological and biogeochemical processes remains approximate especially in very dry and very wet environments and there is a clear need for a better characterisation of these environmental extremes in order for us to move forward to the next generation of climate and land surface prediction models.

Acknowledgements

We gratefully acknowledge funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 603608, Global Earth Observation for integrated water resource assessment: *eartH2Observe*.

References

Ali, H. and Mishra, V.: Contrasting response of rainfall extremes to increase in surface air and dewpoint temperatures at urban locations in India, Sci Rep-Uk, 7, https://doi.org/10.1038/S41598-017-01306-1, 2017.

Arduini, G., Boussetta, S., Dutra, E., and Martínez de la Torre, A.: Report on the Ensemble Water Resources Reanalysis, 2017.

Bates, B., Kundzewicz, Z. W., Wu, S., and Palutikof, J.: Climate Change and Water. In: IPCC Technical Paper 6, Geneva, Switzerland, 2008.

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data, Hydrology and Earth System Sciences, 21, 589-615, https://doi.org/10.5194/hess-21-589-2017, 2017a.

Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A. I. J. M., Weedon, G. P., Brocca, L., Pappenberger, F., Huffman, G. J., and Wood, E. F.: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling, Hydrology and Earth System Sciences, 21, 6201-6217, https://doi.org/10.5194/hess-21-6201-2017, 2017b.

Bhuiyan, M. A. E., Nikolopoulos, E. I., Anagnostou, E. N., Albergel, C., Dutra, E., Fink, G., Martínez de la Torre, A., Munier, S., and Polcher, J.: Assessment of Precipitation Error Propagation in Multi-model Global Water Resources Reanalysis, Hydrology and Earth System Sciences Discussions, doi: 10.5194/hess-2018-434, 2018a. https://doi.org/10.5194/hess-2018-434, 2018a.

Bhuiyan, M. A. E., Nikolopoulos, E. I., Anagnostou, E. N., Quintana-Segui, P., and Barella-Ortiz, A.: A nonparametric statistical technique for combining global precipitation datasets: development and hydrological evaluation over the Iberian Peninsula, Hydrology and Earth System Sciences, 22, 1371-1389, https://doi.org/10.5194/hess-22-1371-2018, 2018b.

Bierkens, M. F. P.: Global hydrology 2015: State, trends, and directions, Water Resources Research, 51, 4923-4947, https://doi.org/10.1002/2015WR017173, 2015.

Dadson, S. J., Hall, J. W., Murgatroyd, A., Acreman, M., Bates, P., Beven, K., Heathwaite, L., Holden, J., Holman, I. P., Lane, S. N., O'Connell, E., Penning-Rowsell, E., Reynard, N., Sear, D., Thorne, C., and Wilby, R.: A restatement of the natural science evidence concerning catchment-based 'natural' flood management in the UK, P Roy Soc a-Math Phy, 473, https://doi.org/10.1098/Rspa.2016.0706, 2017.

Derin, Y., Anagnostou, E., Berne, A., Borga, M., Boudevillain, B., Buytaert, W., Chang, C. H., Delrieu, G., Hong, Y., Hsu, Y. C., Lavado-Casimiro, W., Manz, B., Moges, S., Nikolopoulos, E. I., Sahlu, D., Salerno, F., Rodriguez-Sanchez, J. P., Vergara, H. J., and Yilmaz, K. K.: Multiregional Satellite Precipitation Products Evaluation over Complex Terrain, Journal of Hydrometeorology, 17, 1817-1836, 2016.

He, J., Deser, C., and Soden, B. J.: Atmospheric and Oceanic Origins of Tropical Precipitation Variability, Journal of Climate, 30, 3197-3217, https://doi.org/10.1175/Jcli-D-16-0714.1, 2017.

Hirpa, F. A., Salamon, P., Alfieri, L., Thielen-del Pozo, J., Zsoter, E., and Pappenberger, F.: The Effect of Reference Climatology on Global Flood Forecasting, Journal of Hydrometeorology, 17, 1131-1145, https://doi.org/10.1175/Jhm-D-15-0044.1, 2016.

Huffman, G. J., Adler, R. F., Bolvin, D. T., Gu, G. J., Nelkin, E. J., Bowman, K. P., Hong, Y., Stocker, E. F., and Wolff, D. B.: The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, Journal of Hydrometeorology, 8, 38-55, https://doi.org/10.1175/JHM560.1, 2007.

Huntingford, C., Zelazowski, P., Galbraith, D., Mercado, L. M., Sitch, S., Fisher, R., Lomas, M., Walker, A. P., Jones, C. D., Booth, B. B. B., Malhi, Y., Hemming, D., Kay, G., Good, P., Lewis, S. L., Phillips, O. L., Atkin, O. K., Lloyd, J., Gloor, E., Zaragoza-Castells, J., Meir, P., Betts, R., Harris, P. P., Nobre, C., Marengo, J., and Cox, P. M.: Simulated resilience of tropical rainforests to CO2-induced climate change, Nature Geoscience, 6, 268-273, https://doi.org/10.1038/NGEO1741, 2013.

IPCC: Climate Change 2014: The Physical Science Basis, Intergovernmental Panel on Climate Change (IPCC), U.K., 2014.

IPCC: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, Cambridge, UK, 2012.

Kimani, M. W., Hoedjes, J. C. B., and Su, Z. B.: An Assessment of Satellite-Derived Rainfall Products Relative to Ground Observations over East Africa, Remote Sens-Basel, 9, https://doi.org/10.3390/Rs9050430, 2017.

Kundzewicz, Z. W. and Stakhiv, E. Z.: Are climate models "ready for prime time" in water resources management applications, or is more research needed?, Hydrolog Sci J, 55, 1085-1089, https://doi.org/10.1080/02626667.2010.513211, 2010.
Levizzani, V., Kidd, C., Aonashi, K., Bennartz, R., Ferraro, R. R., Huffman, G. J., Roca, R., Turk, F. J., and Wang, N. Y.: The activities of the International Precipitation Working Group, Quarterly Journal of the Royal Meteorological Society, 144, 3-15, https://doi.org/10.1002/qj.3214, 2018.
Li, H. B. and Wu, J. G.: Uncertainty analysis in ecological studies: An overview. In: Scaling and Uncertainty Analysis in Ecology: Methods and Applications, Springer, Netherlands, 2006.
Long, D., Longuevergne, L., and Scanlon, B. R.: Uncertainty in evapotranspiration from land surface modeling, remote sensing, and GRACE satellites, Water Resources Research, 50, 1131-1151, https://doi.org/10.1002/2013wr014581, 2014.
Lopez, O., Houborg, R., and McCabe, M. F.: Evaluating the hydrological consistency of evaporation products using satellite-based gravity and rainfall data, Hydrology and Earth System Sciences, 21, 323-343, https://doi.org/10.5194/hess-21-323-2017, 2017.
Luo, Z. J., Anderson, R. C., Rossow, W. B., and Takahashi, H.: Tropical cloud and precipitation regimes as seen from near-simultaneous TRMM, CloudSat, and CALIPSO observations and comparison with ISCCP, Journal of Geophysical Research: Atmospheres, 122, 5988-6003, 2017.
Marshall, T. J., Holmes, J. W., and Rose, C. W.: Soil Physics, Cambridge University Press, Cambridge, U.K., 1996.
Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geoscientific Model Development, 10, 1903-1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.
Marthews, T. R., Jones, R. G., Dadson, S. J., Otto, F. E. L., Mitchell, D., Guillod, B. P., and Allen, M. R.: The Impact of Human-Induced Climate Change on Regional Drought in the Horn of Africa, J Geophys Res-Atmos, 124, 4549-4566, https://doi.org/10.1029/2018JD030085, 2019.
McGregor, G. R. and Nieuwolt, S.: Tropical Climatology, Wiley, Chichester, UK, 1998.
Mehran, A. and AghaKouchak, A.: Capabilities of satellite precipitation datasets to estimate heavy precipitation rates at different temporal accumulations, Hydrological Processes, 28, 2262-2270, https://doi.org/10.1002/hyp.9779, 2014.
Munier, S., Minvielle, M., Decharme, B., Calvet, J., Blyth, E., Veldkamp, T. I. E., and Nikolopolous, T.: Report on uncertainty characterization of the WP5 WRR tier 2 products, 2018.
Nikolopoulos, E., Anagnostou, M., Albergel, C., Dutra, E., Fink, G., Martínez-de la Torre, A., Munier, S., Polcher, J., and Quintana-Segui, P.: Report on precipitation error modeling and ensemble error propagation using LSM and GHM models from tier 1 reanalysis, 2016.
Oberkampf, W. L. and Roy, C. J.: Verification and Validation in Scientific Computing, Cambridge University Press, 2010.
Prudhomme, C., Giuntoli, I., Robinson, E. L., Clark, D. B., Arnell, N. W., Dankers, R., Fekete, B. M., Franssen, W., Gerten, D., Gosling, S. N., Hagemann, S., Hannah, D. M., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Wada, Y., and Wisser, D.: Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment, P Natl Acad Sci USA, 111, 3262-3267, https://doi.org/10.1073/pnas.1222473110, 2014.
Riley, K., Thompson, M., Webley, P., and Hyde, K. D.: Uncertainty in Natural Hazards, Modeling and Decision Support. In: Natural Hazard Uncertainty Assessment: Modeling and-Decision Support, Riley, K., Webley, P., and Thompson, M. (Eds.), Wiley, 2017.
Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., Dankers, R., Eisner, S., Fekete, B. M., Colon-Gonzalez, F. J., Gosling, S. N., Kim, H., Liu, X. C., Masaki, Y., Portmann, F. T., Satoh, Y., Stacke, T., Tang, Q. H., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., and Kabat, P.: Multimodel assessment of water scarcity under climate change, P Natl Acad Sci USA, 111, 3245-3250, https://doi.org/10.1073/pnas.1222460110, 2014.
Ukkola, A. M., De Kauwe, M. G., Pitman, A. J., Best, M. J., Abramowitz, G., Haverd, V., Decker, M., and Haughton, N.: Land surface models systematically overestimate the intensity, duration and magnitude of seasonal-scale evaporative droughts, Environmental Research Letters, 11, 104012, https://doi.org/10.1088/1748-9326/11/10/104012, 2016.
Van Loon, A. F.: Hydrological drought explained, Wires Water, 2, 359-392, https://doi.org/10.1002/wat2.1085, 2015.
Veldkamp, T. I. E. and Ward, P.: Report on global scale assessment of physical and social water scarcity, 2015.
Veldkamp, T. I. E., Zhao, F., Ward, P. J., de Moel, H., Aerts, J. C. J. H., Schmied, H. M., Portmann, F. T., Masaki, Y., Pokhrel, Y., Liu, X., Satoh, Y., Gerten, D., Gosling, S. N., Zaherpour, J., and Wada, Y.: Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study, Environmental Research Letters, 13, 2018.
Wartenburger, R., Seneviratne, S. I., Hirschi, M., Chang, J. F., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Gosling, S. N., Gudmundsson, L., Henrot, A. J., Hickler, T., Ito, A., Khabarov, N., Kim, H., Leng, G. Y., Liu, J. G., Liu, X. C., Masaki, Y., Morfopoulos, C., Muller, C., Schmied, H. M., Nishina, K., Orth, R., Pokhrel, Y., Pugh, T. A. M., Satoh, Y., Schaphoff, S., Schmid, E., Sheffield, J., Stacke, T., Steinkamp, J., Tang, Q. H., Thiery, W., Wada, Y., Wang, X. H., Weedon, G. P., Yang, H., and Zhou, T.: Evapotranspiration simulations in ISIMIP2a-Evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets, Environmental Research Letters, 13, 2018.

Yin, H., Donat, M. G., Alexander, L. V., and Sun, Y.: Multi-dataset comparison of gridded observed temperature and precipitation extremes over China, International Journal of Climatology, 35, 2809-2827, https://doi.org/10.1002/joc.4174, 2015.

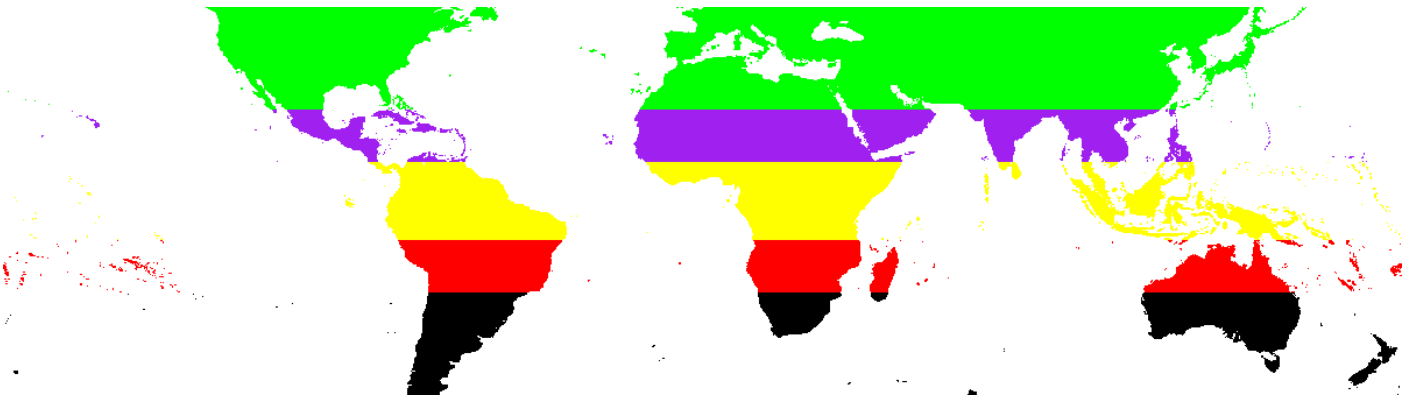
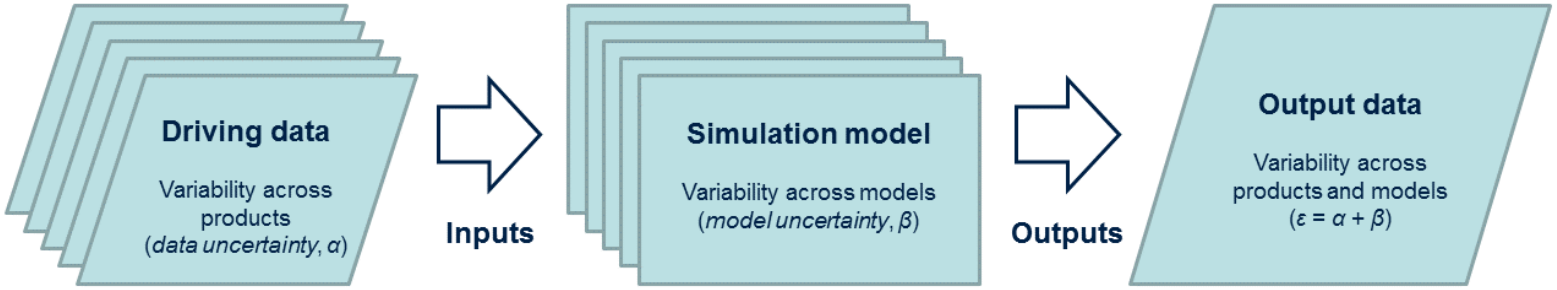


Fig. 1: Latitudinal zones used in this study: black = southern temperate 23.5°S to 50.0°S, red = southern tropical 10.0°S to 23.5°S, yellow = equatorial tropical 10.0°N to 10.0°S, purple = northern tropical 23.5°N to 10.0°N and green = northern temperate 50.0°N to 23.5°S. Analyses are restricted to the area 50.0°N to 50.0°S because of the bounds of data validity in the TRMM and TRMM-RT precipitation data products (Table 2).



The relationship between α, β and ε is most clearly explained by example (*P*=precipitation):

1. Say at this point and time we have 3 *P* estimates from different data products: 5 mm, 8 mm and 10 mm. We can calculate the standard deviation *DIU* =SD(5,8,10) =2.52 mm

2. Assume also that we have 3 models for predicting *X*=runoff:
   - *Model 1* assumes runoff is equal to 2 mm/day plus an exponential contribution from *P* if it exceeds 4 mm.
   - *Model 2* is a very basic model, assuming constant runoff at this location based on the historical average, say 8.2 mm.
   - *Model 3* assumes runoff is 50% of *P* plus a contribution from groundwater return flow that ranges from 0.1 mm to 100.0 mm depending on the state of belowground aquifers.

Driving our models with those *P* numbers to produce an estimate of *X*, we might get a table like this:

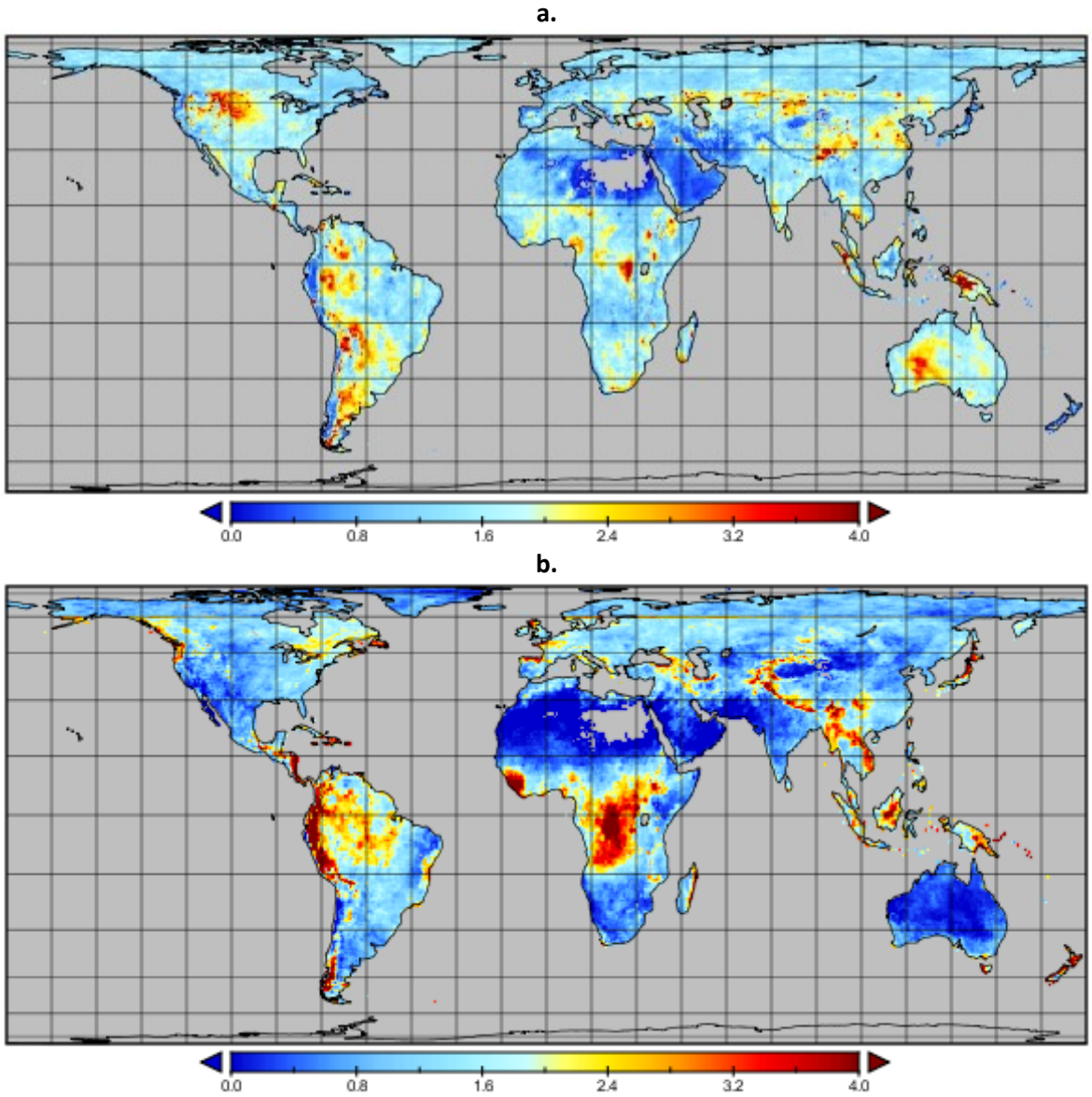
***P* estimate**	**Runoff (mm/day) from**			**SD across models (mm/day)**
	Model 1	***Model 2***	***Model 3***	
5 mm	2.0+exp(5-4.0) = 4.7	8.2	(0.50*5)+0.1 = 2.6	2.8
8 mm	2.0+exp(8-4.0) = 56.6	8.2	(0.50*8)+10.0 = 14.0	26.4
10 mm	2.0+exp(10-4.0) = 405.4	8.2	(0.50*10)+100.0 = 105.0	207.1
SD across products (mm/day):	217.9	0.0	56.1	Mean from the left = 91.3 mm/day Mean from above = 78.8 mm/day

3. Note that *DOU* = mean(SDs across products) = 91.3 mm/day, which is not equal to *MU* = mean(SDs across models) = 78.8 mm/day (there is no constraint for these to be equal in general). We are interested in when these values are greater or less than *DIU*, so we consider the scaled uncertainties α=(*DOU*÷*DIU*) and β=(*MU*÷*DIU*).

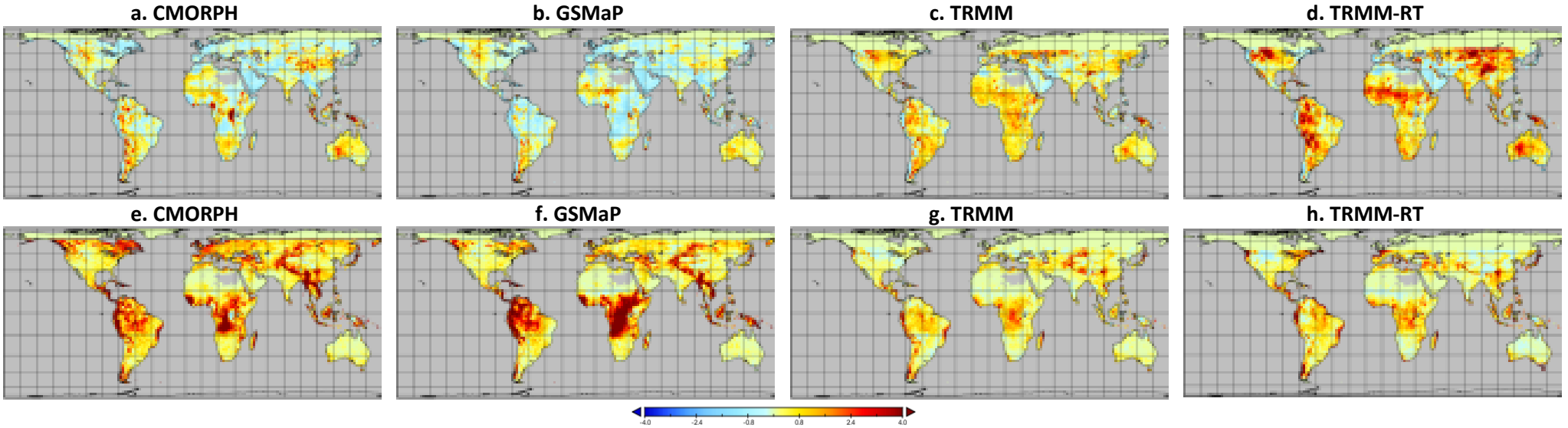
4. Note the key difference between α, which is calculated from the outputs of the model, and *DIU*, which is calculated from the inputs: why not just consider *DIU*? Because our focus is on *X* and therefore we need to quantify the *uncertainty introduced into X by the precipitation* (α), which is not the same as the uncertainty in the precipitation (*DIU*) (this is an attribution study, therefore we focus on α rather than *DIU*).

5. In this analysis, we considered SDs of extreme event occurrence (EE/yr) rather than SDs of straight *X* values, which we have done for two reasons: (i) this allows us to consider and compare consistently the uncertainties of different response variables with different units (e.g. *X*=runoff *vs.* *X*=evapotranspiration) and (ii) in a global analysis it is necessary to compare across biomes (e.g. a desert point with a rainforest point) and using event occurrence statistics avoids the bias towards wet or dry regions (because of their greater absolute values of e.g. runoff) that must be corrected for in studies that work with the absolute values of *X*. Using occurrence statistics doesn't change the calculations of α, β and ε above, but does involve the additional assumption of a baseline distribution against which we may measure how extreme conditions are (see §2.1).

Fig. 2: Uncertainty measures quantifying how much a simulation model (land surface or hydrological model) alters the uncertainty introduced to its simulations via the precipitation driver inputs, following the *method of competing models* approach advocated for complex systems by Oberkampf and Roy (2010).



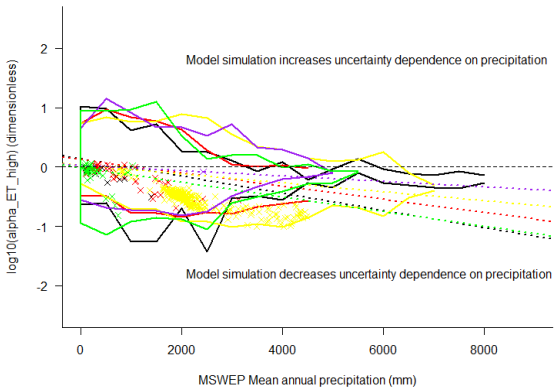
~~Fig. 2~~Fig. 3: Uncertainty in the precipitation inputs to the *eartH2Observe* ensemble models: (a) Uncertainty in precipitation extreme highs and (b) Uncertainty in precipitation extreme lows (standard deviation (SD) taken across the precipitation products) in units of (occurrence of extreme events per year). Areas of consistently very low precipitation are masked in grey. Note that only isolated global areas exceeded 4 events/yr, so the scale is restricted to 0-4 events/yr.



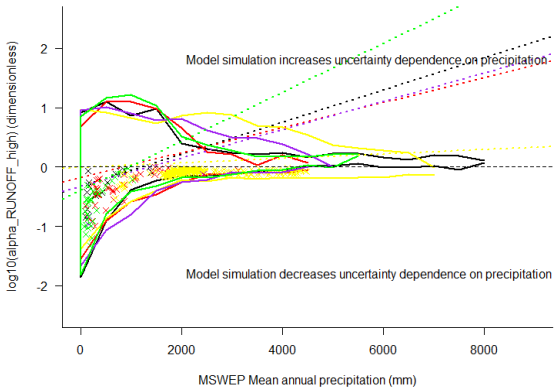
~~Fig. 3~~Fig. 4: Increase in extreme precipitation event occurrence in relation to MSWEP. Subtracting extreme high event occurrence rates in the MSWEP precipitation input from the rates in the CMORPH precipitation input gives map (a), and (b) to (d) are the same calculation using GSMaP, TRMM and TRMM-RT instead of CMORPH. (e) to (h) is the same calculation, but for extreme low event occurrence (i.e. the averages of the upper and lower rows are effectively the maps ~~Fig. 2~~Fig. 3a and ~~Fig. 2~~Fig. 3b, respectively). The clear lines at 50°N (TRMM, TRMM-RT) and 60°N (CMORPH, GSMaP) show the bounds of data validity for these products (Table 2). Note that only isolated global areas exceeded 4 events/yr, so the scale is restricted to -4 to +4 events/yr.

33
34

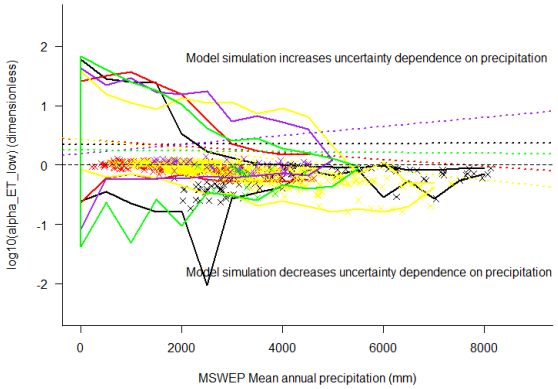
a.



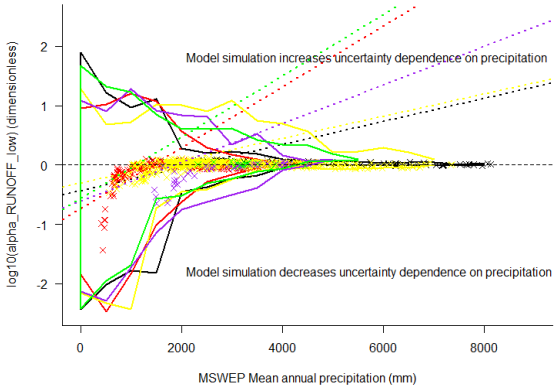
b.



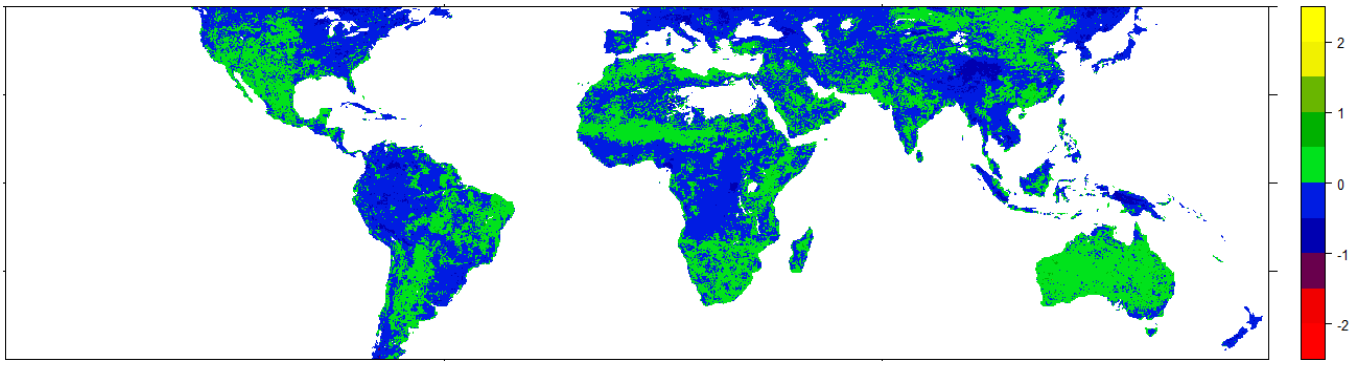
c.



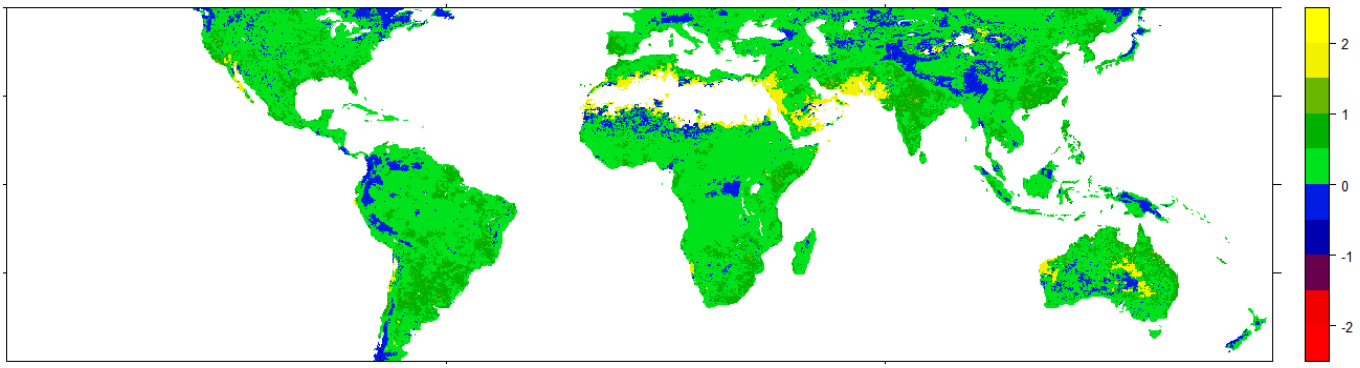
d.



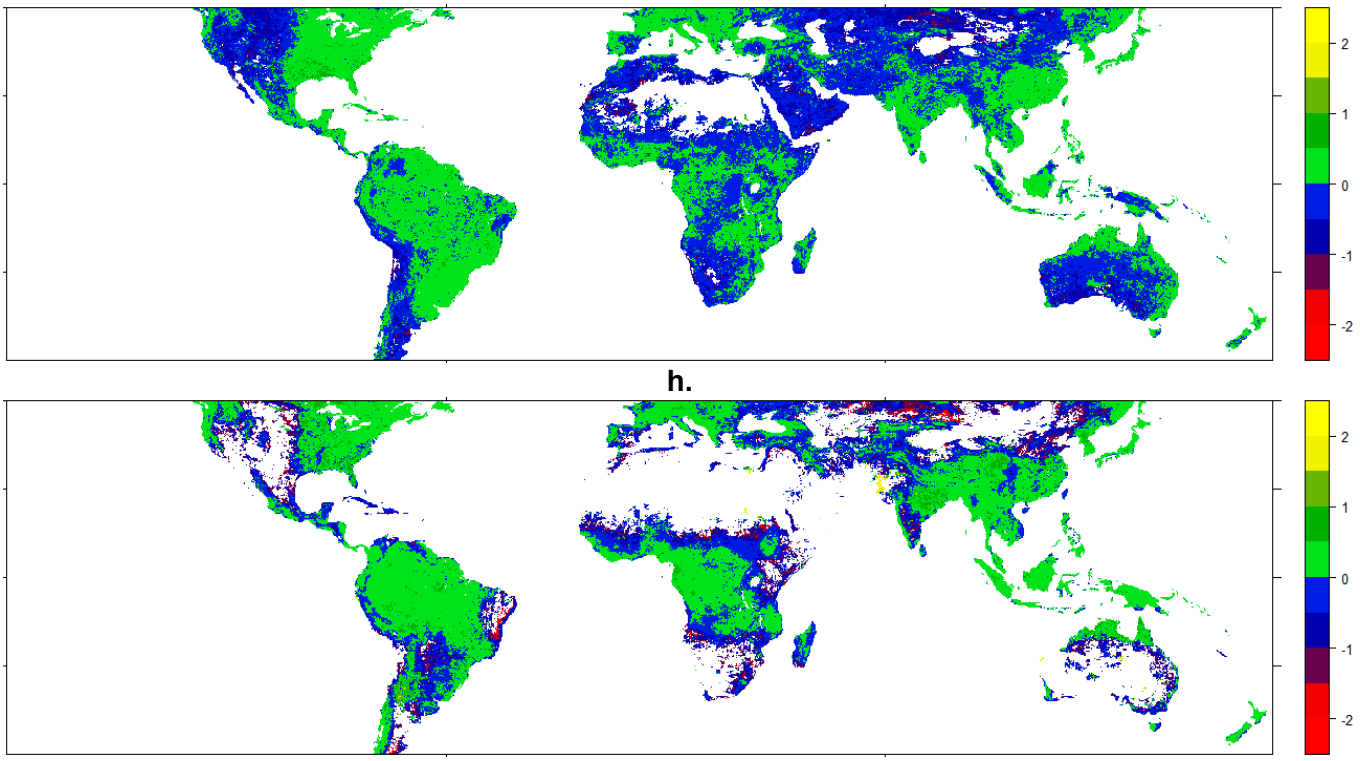
e.



f.



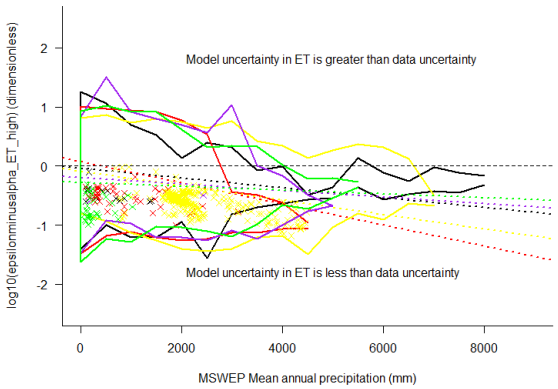
g.



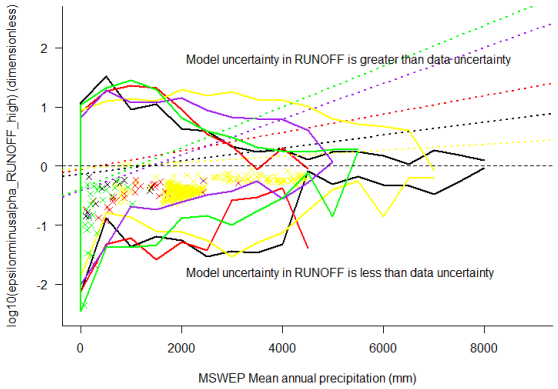
~~Fig. 4~~Fig. 5: Values of $\log_{10}(\alpha_{X,j})$, where $\alpha_{X,j}$ is the scaled data uncertainty in variable X (eqn 1). ($\log_{10}(\alpha_{X,j})<0$ indicates uncertainty in the predicted variable X attributable to the data is less than the variability in the input precipitation forcing data; >0 indicates uncertainty in the predicted variable X is greater), where X is evapotranspiration (a, c, e, f) or runoff (b, d, g, h) and j refers to either high extremes (a, b, e, g) or low extremes (c, d, f, h). Points on the scatter plots are coloured according to latitudinal zones (Fig. 1). Because of the density of overlapping points, only the envelope of points for each latitudinal zone is shown and the points with the highest uncertainty (uncertainty $DIU \geq (2/3)$*(global maximum of DIU)). Linear regression lines for each latitudinal zone indicate the trend as precipitation increases within each zone (all regressions were significant at the 1% level), although n.b. we do not contend in any way that the distribution of points shown is linear: these lines simply indicate a trend that is not clear to the eye from the envelopes displayed (which do not show the complete point cloud). Maps (e-h) show the corresponding spatial distributions of $\log_{10}(\alpha_{X,j})$ values for each variable, with the colour scales corresponding to the vertical axis on scatter plot (a).

49
50

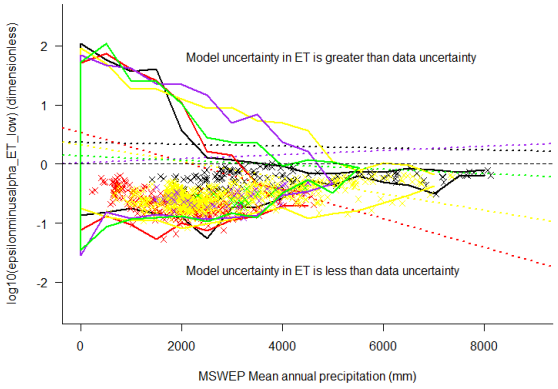
a.



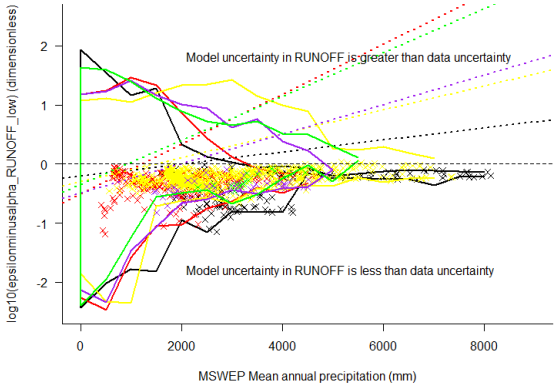
b.



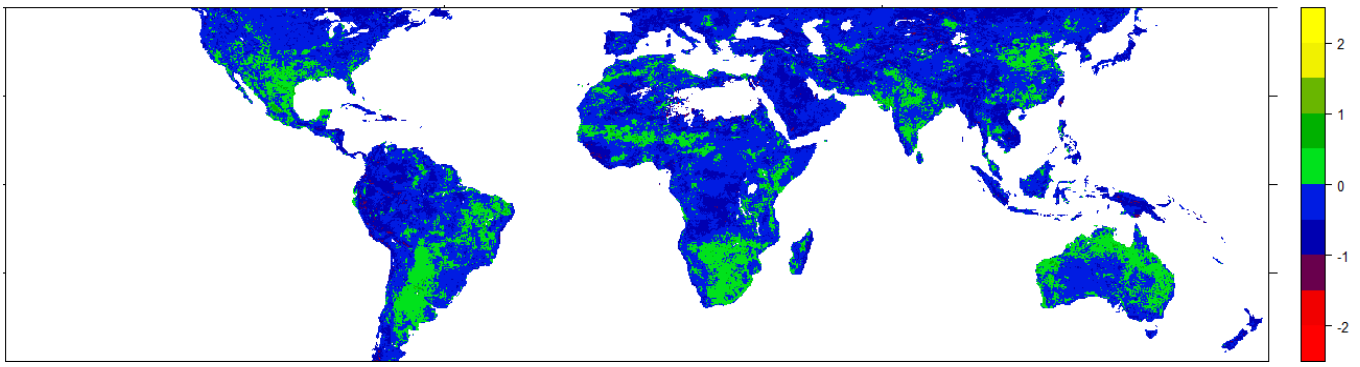
c.



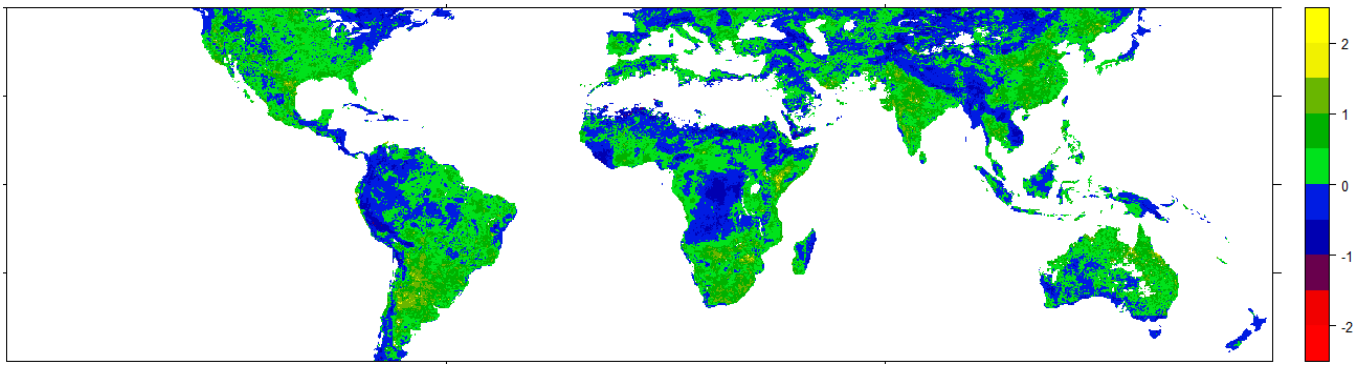
d.



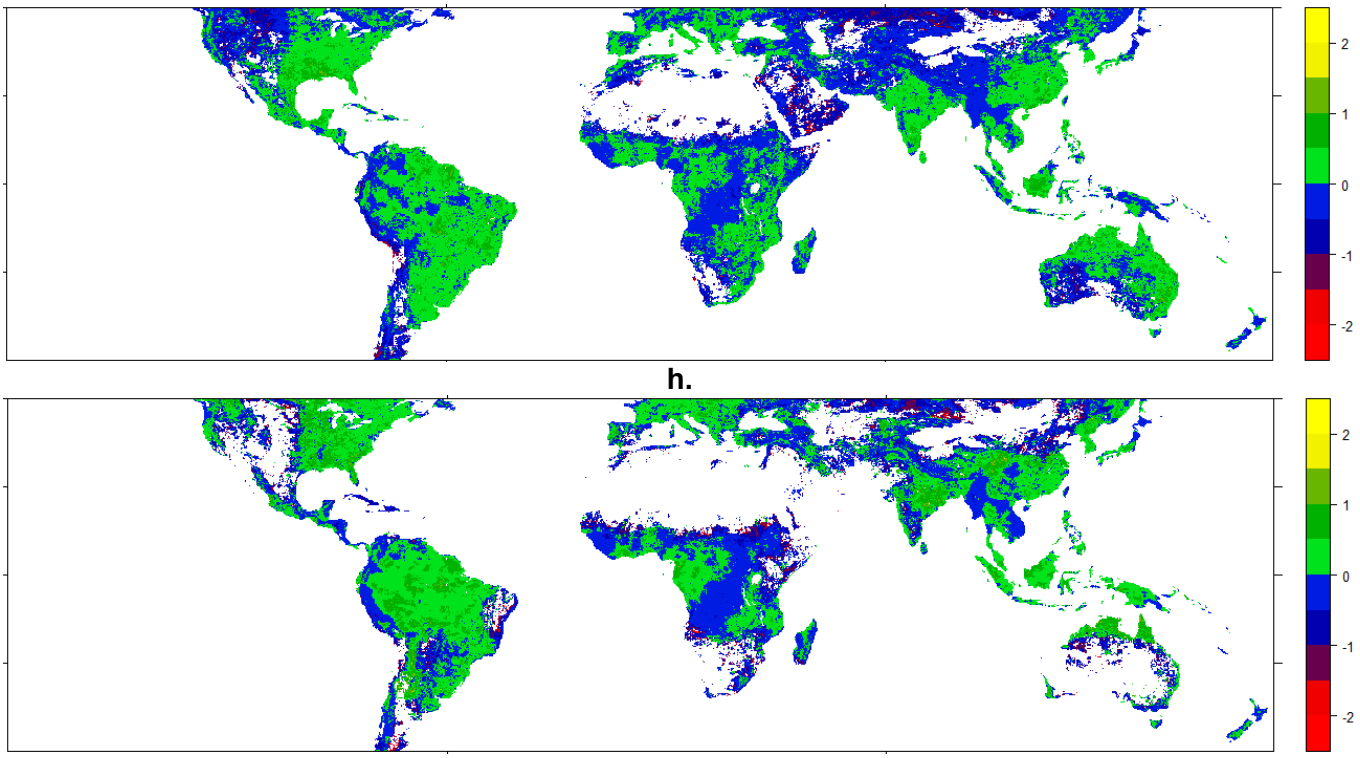
e.



f.



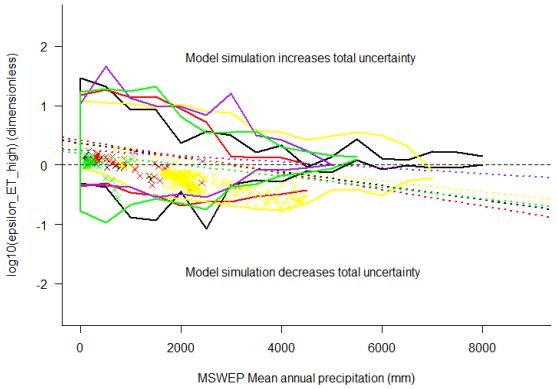
g.



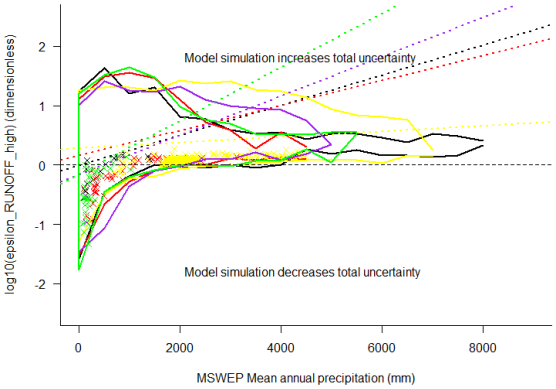
~~Fig. 5~~Fig. 6: Values of $\log_{10}(\beta_{X,j})$, where $\beta_{X,j}$ is the scaled model uncertainty in variable X (eqn 2). ($\log_{10}(\beta_{X,j})<0$ indicates model uncertainty in the predicted variable X is less than the variability in the input precipitation forcing data; >0 indicates model uncertainty in the predicted variable X is greater), where X is evapotranspiration (a, c, e, f) or runoff (b, d, g, h) and j refers to either high extremes (a, b, e, g) or low extremes (c, d, f, h). Points on the scatter plots are coloured according to latitudinal zones (Fig. 1). Because of the density of overlapping points, only the envelope of points for each latitudinal zone is shown and the points with the highest uncertainty (uncertainty $DIU \geq (2/3)$*(global maximum of DIU)). Linear regression lines for each latitudinal zone indicate the trend as precipitation increases within each zone (all regressions were significant at the 1% level), although n.b. we do not contend in any way that the distribution of points shown is linear: these lines simply indicate a trend that is not clear to the eye from the envelopes displayed (which do not show the complete point cloud). Maps (e-h) show the corresponding spatial distributions of $\log_{10}(\beta_{X,j})$ values for each variable, with the colour scales corresponding to the vertical axis on scatter plot (a).

65
66
67

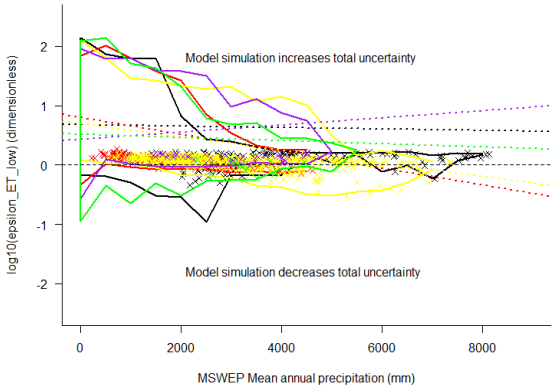
a.



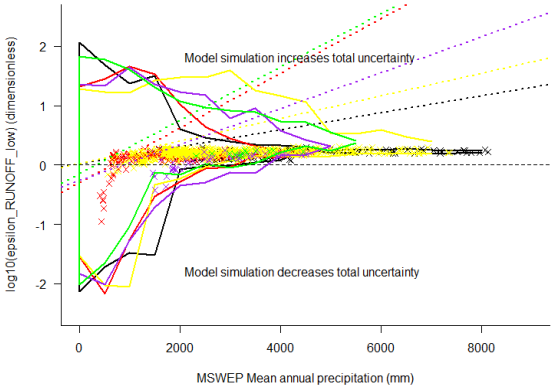
b.



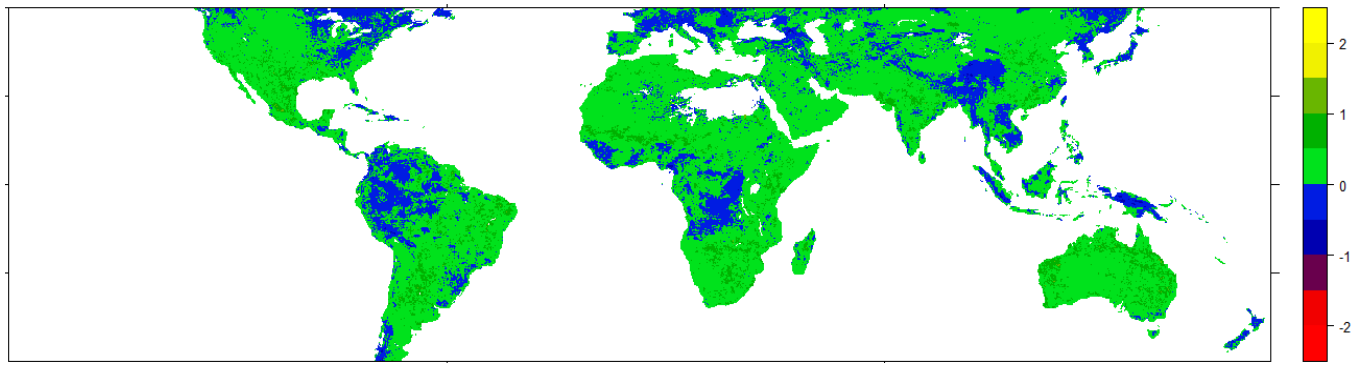
c.



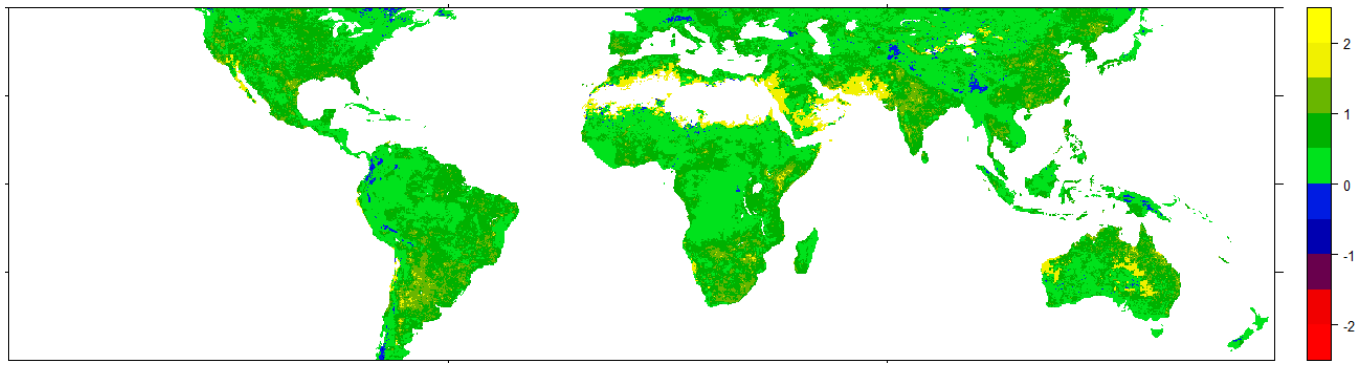
d.



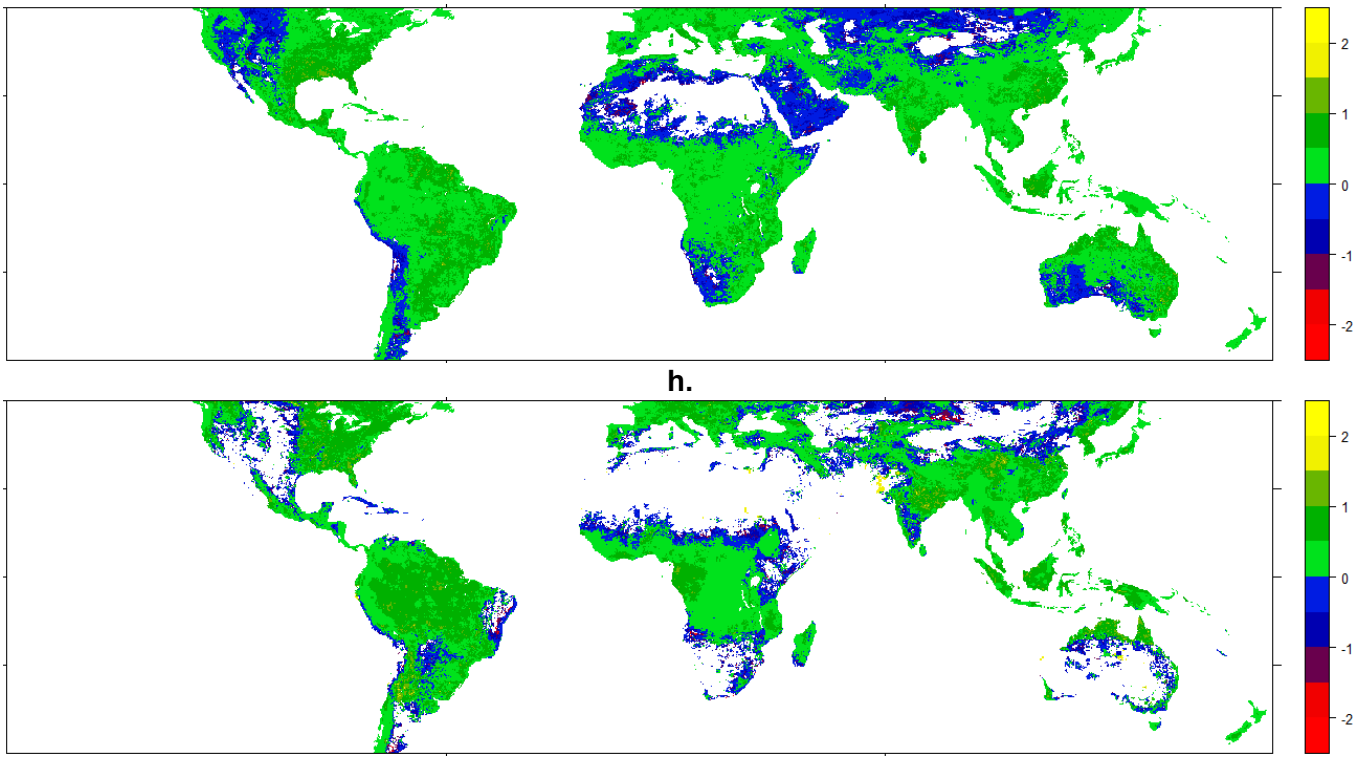
e.



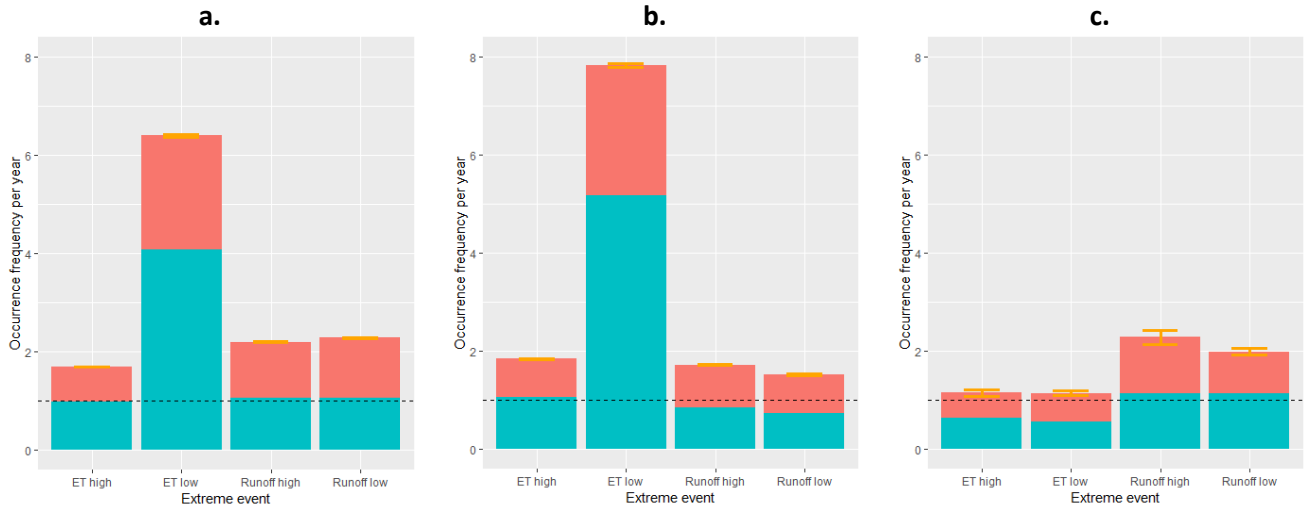
f.



g.



~~Fig. 6~~Fig. 7: Values of $\log_{10}(\varepsilon_{X,j})$, where $\varepsilon_{X,j}$ is the total uncertainty in variable X (eqn 3), where X is evapotranspiration (a, c, e, f) or runoff (b, d, g, h) and j refers to either high extremes (a, b, e, g) or low extremes (c, d, f, h). Points on the scatter plots are coloured according to latitudinal zones (Fig. 1). Because of the density of overlapping points, only the envelope of points for each latitudinal zone is shown and the points with the highest uncertainty (uncertainty $DIU \geq (2/3)$*(global maximum of DIU)). Linear regression lines for each latitudinal zone indicate the trend as precipitation increases within each zone (all regressions were significant at the 1% level), although n.b. we do not contend in any way that the distribution of points shown is linear: these lines simply indicate a trend that is not clear to the eye from the envelopes displayed (which do not show the complete point cloud). Maps (e-h) show the corresponding spatial distributions of $\log_{10}(\varepsilon_{X,j})$ values for each variable, with the colour scales corresponding to the vertical axis on scatter plot (a).



~~**Fig. 7**~~**Fig. 8**: Global mean values (averaged over 50°S to 50°N) from scatter plots in ~~Fig. 4~~Fig. 5, ~~Fig. 5~~Fig. 6 and ~~Fig. 6~~Fig. 7. Plots show (a) all values, (b) values from dry environments with mean annual precipitation <1000 mm/yr only and (c) values from wet environments ≥6000 mm/yr only. Bar heights are ε values (scaled total uncertainty), with ■ showing α values (scaled data uncertainty) and ■ β (scaled model uncertainty); error bars show SE.

Oberkampf, W. L., and C. J. Roy (2010). *Verification and Validation in Scientific Computing*, Cambridge University Press.

Table 1: Types of precipitation and their main controlling factors (McGregor and Nieuwolt, 1998).

Precipitation type	Spatial scale	Characteristics	Challenges
Cyclonic (=frontal)	Synoptic, regional	The leading edge of a warm and moist air mass (warm front) meets a cool, dry air mass (cold front). The warmer air mass rises over the cooler air, with precipitation occurring along the front. If the air begins to circulate, a cyclonic storm can occur.	• It is widely accepted that global warming will lead to a higher water-holding capacity for the atmosphere as well as increased rates of evaporation, and therefore increased extreme weather (Trenberth et al., 2015; Yi et al., 2015). However, the mechanisms through which the location and magnitude of these extreme events may be predicted (e.g. tipping points, thresholds) remain inadequately understood (Marthews et al., 2012).
Orographic	Intermediate	Warm, moist air entering a mountain range is forced to rise, and then cools and precipitation ensues (= *orographic lift*).	• Scale is an important issue: mountains can modify large-scale circulation, causing changes in local moisture convergence, but local condensation and microphysical processes also influence flow stability upstream (Marthews et al., 2012).
Convective	Local (often sub-grid)	A warm soil or vegetation surface warms the air above it, which then rises vertically and cools, with precipitation occurring on cooling. 'Convection-permitting' model runs usually require a sub-daily timestep and <10 km spatial resolution, and in the absence of these a convection parameterisation scheme (CPS) is necessary (i.e. assumptions about subgrid and subdaily dynamics) (Prein *et al.* 2015).	• *Stratiform precipitation* is when the rise is diagonal rather than vertical (i.e. similar to orographic, but not as a result of landform) • Sub-grid displacement of cloud occurrence from driver (Taylor et al., 2012) • Land surface exchange (e.g. evapotranspiration) has a significant effect, but often not modelled explicitly. • Resolution of snow versus rainfall in mountain regions is critical for water resources management, but not well-characterised in models. • CPSs generally overestimate light rain (drizzle) because they overestimate the number of precipitation days (by equating clouds with rain) and / or underestimate precipitation intensity (Marthews et al., 2012; Prein et al., 2015). Conversely, it is a known limitation of some satellites that they are not sensitive to, and therefore underestimate, light rain (e.g. Luo et al. (2017)). This introduces a 'calibration gap': calibration of large-scale models against satellite-based precipitation observations must not only factor out the overestimation of CPSs, but also the underestimation of the observations.

Table 2: Global precipitation products used to drive the models selected from Dorigo *et al.* (2014). Data files used are available through the Water Cycle Integrator https://wci.eartH2Observe.eu/ at 25 km resolution for the period 2000-2013. Algorithm type is as given by the International Precipitation Working Group (IPWG) *.

Product	Algorithm	Notes
Multi-Source Weighted-Ensemble Precipitation (MSWEP)		Global reanalysis data (Beck et al., 2017)
Climate Prediction Center MORPHing Technique (CMORPH)	Blended microwave-infrared	Restricted to 60°S to 60°N A passive microwave-based product advected in time using geosynchronous infrared data (Joyce et al., 2004). When microwave observations are not available, infrared observations are used to advect the last microwave scan over time. In addition to advecting precipitation forward in time, the algorithm propagates precipitation backward once the next microwave observation becomes available (Mehran and AghaKouchak, 2014).
Global Satellite Mapping of Precipitation (GSMaP)	Blended microwave-infrared	Restricted to 60°S to 60°N (Tian et al., 2010)
Tropical Rainfall Measuring Mission (TRMM)	Satellite-based	Restricted to 50°S to 50°N
TRMM Real Time (TRMM-RT)	Satellite-based	Restricted to 50°S to 50°N Mainly based on microwave data aboard Low Earth Orbit satellites (Huffman et al., 2007). The TRMM-RT algorithm is primarily based on microwave observations from low orbiter satellites. Gaps in microwave observations are filled with infrared data (Mehran and AghaKouchak, 2014).

* *Real-time* usually = there is at most a 1-2 hour delay before observation data is made available raw (i.e. with no gap-filling or other modification).
Near-real-time = there is at most a 1-2 day delay before delivery, allowing some initial data checks to be carried out.
Reanalysis data = data assimilation techniques have been used to fill gaps in the observation data (e.g. missing variables).
Blended = observation data have been combined with either or both of raingauge and reanalysis data to create a more robust and quality-controlled product.

Table 3: Modelling systems details (Dutra et al., 2015; Nikolopoulos et al., 2016). Each model was driven using as close as possible to the same configuration: Global Water Resources Reanalysis 2 (WRR2, Arduini et al. (2017) and http://jules.jchmr.org/content/research-community-configurations). Simulation results are available on the THREDDS data server (https://wci.eartH2Observe.eu/thredds/catalog.html, see Schellekens et al. (2017)).

Model	Institution	Simulations
Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land model (**H-TESSEL**) (Balsamo et al., 2009)	ECMWF	A 10-year spin-up was carried out: an initial run from 1 January 1979 to 1 January 1989, while the land surface state of January 1989 was used to initialize the main simulation.
JULES is the Joint UK Land Environment Simulator model (**JULES**) (Best et al., 2011; Clark et al., 2011)	MetO/CEH	A 10-year spin-up was carried out: an initial run from 1 January 1979 to 1 January 1989, while the land surface state of January 1989 was used to initialize the main simulation.
ORganizing Carbon and Hydrology In Dynamic EcosystEms model (**ORCHIDEE**) (d'Orgeval et al., 2008; Krinner et al., 2005)	CNRS/~~LSCE~~IPSL	The model was spun up with a simulation from 1 January 1979 to 31 December 1990. This simulation started with an average soil moisture and empty aquifers. After the 12 years of spin-up, river discharges have reached equilibrium.
SURFace EXternalisée model (**SURFEX**) (Decharme et al., 2011; Decharme et al., 2013)	Météo-France	A 20-year spin-up was carried out using the 1979–1988 period twice.
Water – Global Assessment and Prognosis-3 (**WaterGAP3**) (Schneider et al., 2011; Verzano et al., 2012). A grid-based, integrative global fresh water resource assessment tool.	University of Kassel	Storage compartments were initialized by re-running the model with the first year of available meteorological forcing 10 times. WaterGAP includes a water use model (domestic and industrial water use are parameterised as a function of average income per country (GDP/capita), allowing global water use calculations.

References

Arduini, G., S. Boussetta, E. Dutra, and A. Martínez de la Torre (2017). Report on the Ensemble Water Resources Reanalysis*Rep.*

Balsamo, G., P. Viterbo, A. Beljaars, B. van den Hurk, M. Hirschi, A. K. Betts, and K. Scipal (2009). A Revised Hydrology for the ECMWF Model: Verification from Field Site to Terrestrial Water Storage and Impact in the Integrated Forecast System. *Journal of Hydrometeorology*, *10*(3), 623-643, doi:10.1175/2008JHM1068.1.

Beck, H. E., A. I. J. M. van Dijk, V. Levizzani, J. Schellekens, D. G. Miralles, B. Martens, and A. de Roo (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, *21*(1), 589-615, doi:10.5194/hess-21-589-2017.

Best, M. J., et al., (2011). The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes. *Geoscientific Model Development*, *4*(3), 677-699, doi:10.5194/gmd-4-677-2011.

Clark, D. B., et al., (2011). The Joint UK Land Environment Simulator (JULES), model description – Part 2: Carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, *4*(3), 701-722, doi:10.5194/gmd-4-701-2011.

d'Orgeval, T., J. Polcher, and P. de Rosnay (2008). Sensitivity of the West African hydrological cycle in ORCHIDEE to infiltration processes. *Hydrology and Earth System Sciences*, *12*(6), 1387-1401, doi:DOI 10.5194/hess-12-1387-2008.

Decharme, B., A. Boone, C. Delire, and J. Noilhan (2011). Local evaluation of the Interaction between Soil Biosphere Atmosphere soil multilayer diffusion scheme using four pedotransfer functions. *J Geophys Res-Atmos*, *116*, doi:10.1029/2011jd016002.

Decharme, B., E. Martin, and S. Faroux (2013). Reconciling soil thermal and hydrological lower boundary conditions in land surface models. *J Geophys Res-Atmos*, *118*(14), 7819-7834, doi:10.1002/jgrd.50631.

Dutra, E., et al., (2015). Report on the current state-of-the-art Water Resources Reanalysis*Rep.*

Huffman, G. J., R. F. Adler, D. T. Bolvin, G. J. Gu, E. J. Nelkin, K. P. Bowman, Y. Hong, E. F. Stocker, and D. B. Wolff (2007). The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of Hydrometeorology*, *8*(1), 38-55, doi:10.1175/JHM560.1.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. P. Xie (2004). CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *Journal of Hydrometeorology*, *5*(3), 487-503, doi:Doi 10.1175/1525-7541(2004)005<0487:Camtpg>2.0.Co;2.

Krinner, G., N. Viovy, N. de Noblet-Ducoudré, J. Ogée, J. Polcher, P. Friedlingstein, P. Ciais, S. Sitch, and I. C. Prentice (2005). A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system. *Global Biogeochemical Cycles*, *19*(1), doi:10.1029/2003gb002199.

Luo, Z. J., R. C. Anderson, W. B. Rossow, and H. Takahashi (2017). Tropical cloud and precipitation regimes as seen from near-simultaneous TRMM, CloudSat, and CALIPSO observations and comparison with ISCCP. *Journal of Geophysical Research: Atmospheres*, *122*, 5988-6003.

Marthews, T. R., et al., (2012). Simulating forest productivity along a neotropical elevational transect: temperature variation and carbon use efficiency. *Glob Chang Biol*, *18*(9), 2882-2898, doi:10.1111/j.1365-2486.2012.02728.x.

McGregor, G. R., and S. Nieuwolt (1998). *Tropical Climatology*. 2nd ed., Wiley, Chichester, UK.

Mehran, A., and A. AghaKouchak (2014). Capabilities of satellite precipitation datasets to estimate heavy precipitation rates at different temporal accumulations. *Hydrological Processes*, *28*(4), 2262-2270, doi:10.1002/hyp.9779.

Nikolopoulos, E., M. Anagnostou, C. Albergel, E. Dutra, G. Fink, A. Martínez-de la Torre, S. Munier, J. Polcher, and P. Quintana-Segui (2016). Report on precipitation error modeling and ensemble error propagation using LSM and GHM models from tier 1 reanalysis*Rep.*

Prein, A. F., et al., (2015). A review on regional convection-permitting climate modeling: Demonstrations, prospects, and challenges. *Reviews of Geophysics*, *53*(2), 323-361, doi:10.1002/2014RG000475.

Schellekens, J., et al., (2017). A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset. *Earth Syst Sci Data*, *9*(2), 389-413, doi:10.5194/essd-9-389-2017.

Schneider, C., M. Florke, S. Eisner, and F. Voss (2011). Large scale modelling of bankfull flow: An example for Europe. *Journal of Hydrology*, *408*(3-4), 235-245, doi:10.1016/j.jhydrol.2011.08.004.

Taylor, C. M., R. A. de Jeu, F. Guichard, P. P. Harris, and W. A. Dorigo (2012). Afternoon rain more likely over drier soils. *Nature*, *489*(7416), 423-426, doi:10.1038/nature11377.

Tian, Y. D., C. D. Peters-Lidard, R. F. Adler, T. Kubota, and T. Ushio (2010). Evaluation of GSMaP Precipitation Estimates over the Contiguous United States. *Journal of Hydrometeorology*, *11*(2), 566-574, doi:10.1175/2009JHM1190.1.

Trenberth, K. E., J. T. Fasullo, and T. G. Shepherd (2015). Attribution of climate extreme events. *Nature Climate Change*, *5*(8), 725-730, doi:10.1038/nclimate2657.

Verzano, K., I. Barlund, M. Florke, B. Lehner, E. Kynast, F. Voss, and J. Alcamo (2012). Modeling variable river flow velocity on continental scale: Current situation and climate change impacts in Europe. *Journal of Hydrology*, *424*, 238-251, doi:10.1016/j.jhydrol.2012.01.005.

Yi, C., E. Pendall, and P. Ciais (2015). Focus on extreme events and the carbon cycle. *Environmental Research Letters*, *10*(7), 070201, doi:10.1088/1748-9326/10/7/070201.

Supplementary information

Data and code availability

All simulation data have been made publicly-available through a Water Cycle Integrator portal (WCI) at https://wci.eartH2Observe.eu/. Requests for further data are very welcome and may be addressed to the corresponding author.

Global maps were calculated for sections of the globe using a custom script written in Python v.2.7.5 and then knitted together using NetCDF Operators (NCO) Tools (Zender, 2008) called from a custom script written in R v.3.5.1 (R Core Team, 2018) (scripts are available on request from the corresponding author). Visualisations were created using Panoply v.4.4.1 and R v.3.5.1 (R Core Team, 2018).

Uncertainty maps

Absolute uncertainty numbers may not be comparable between this study and other simulations, but our results give a first estimate of the relative uncertainties of predictions from particular models and precipitation products of evapotranspiration highs (Fig. S1), evapotranspiration lows (Fig. S2), runoff highs (Fig. S3) and runoff lows (Fig. S4).

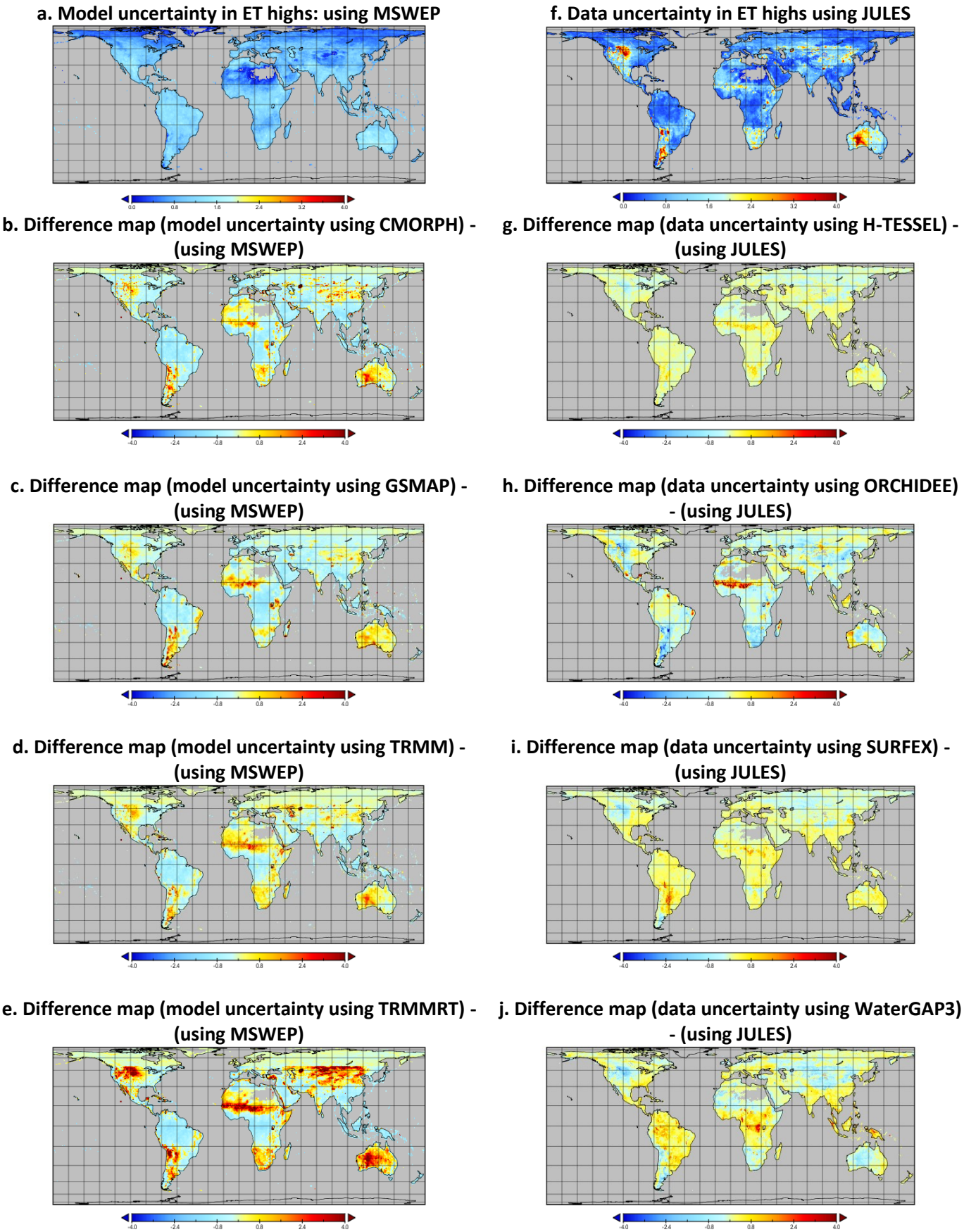


Fig. S1: Evapotranspiration (ET) highs. Note the differing scales: plots in top row scale ranges 0.0-4.0 extreme events per year (EE/yr) while the remaining rows ranging -4.0 to 4.0 EE/yr.

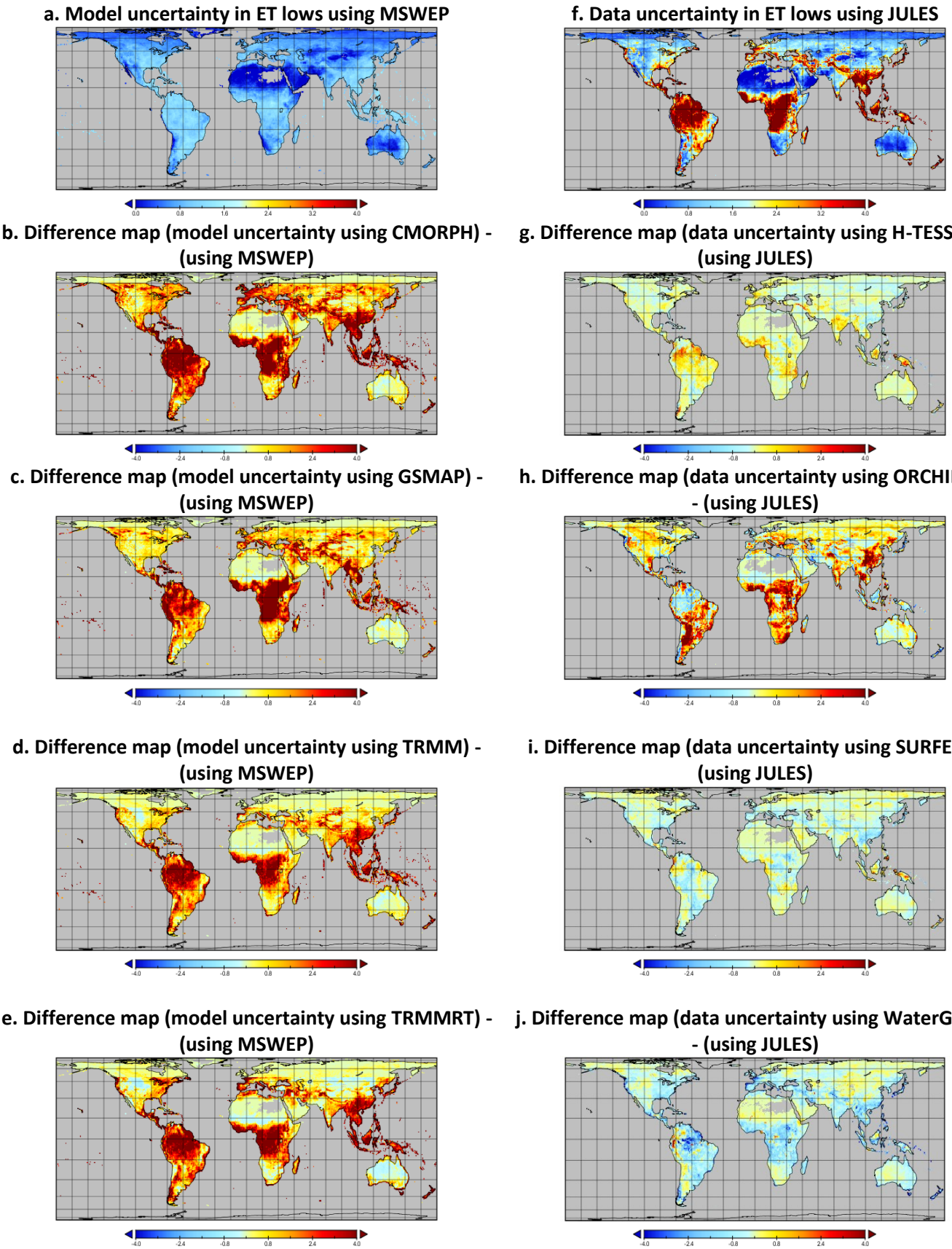
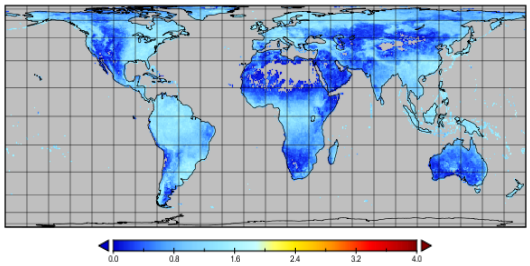
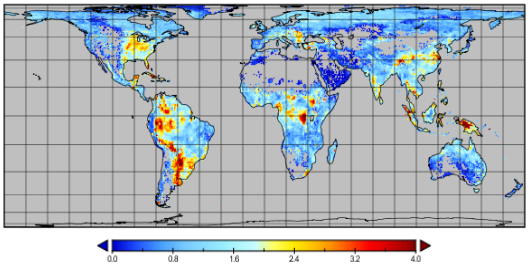


Fig. S2: Evapotranspiration (ET) lows. Note the differing scales: plots in top row scale ranges 0.0-4.0 extreme events per year (EE/yr) while the remaining rows ranging -4.0 to 4.0 EE/yr.

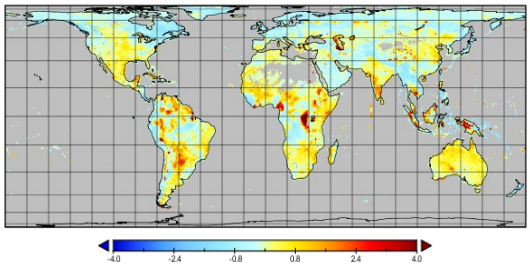
a. Model uncertainty in RUNOFF highs: using MSWEP



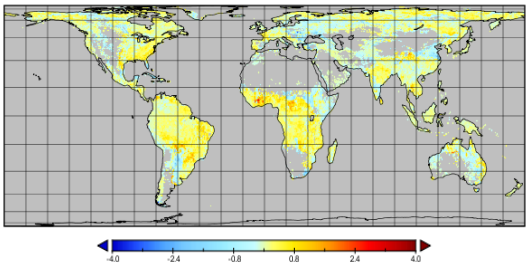
f. Data uncertainty in RUNOFF highs using JULES



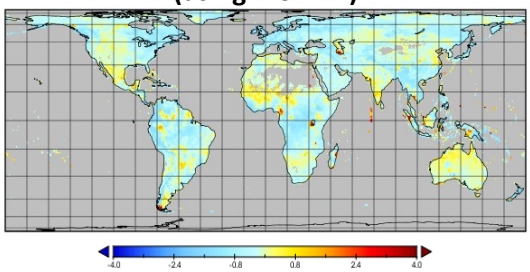
b. Difference map (model uncertainty using CMORPH) - (using MSWEP)



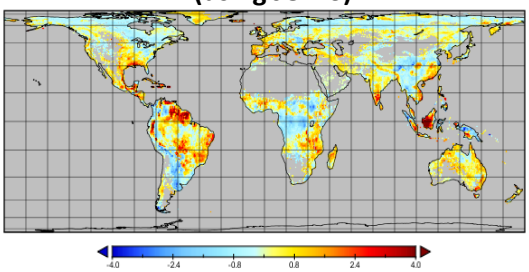
g. Difference map (data uncertainty using H-TESSEL) - (using JULES)



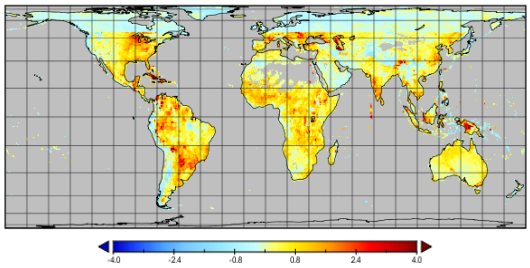
c. Difference map (model uncertainty using GSMAP) - (using MSWEP)



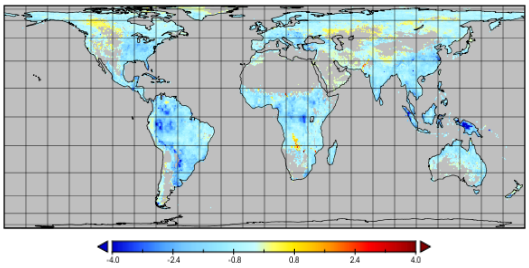
h. Difference map (data uncertainty using ORCHIDEE) - (using JULES)



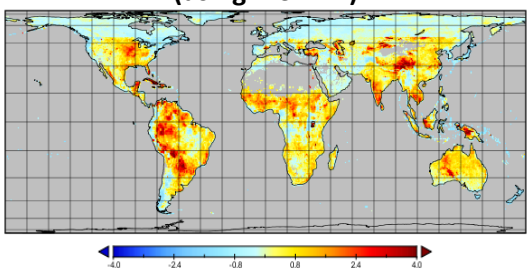
d. Difference map (model uncertainty using TRMM) - (using MSWEP)



i. Difference map (data uncertainty using SURFEX) - (using JULES)



e. Difference map (model uncertainty using TRMMRT) - (using MSWEP)



j. Difference map (data uncertainty using WaterGAP3) - (using JULES)

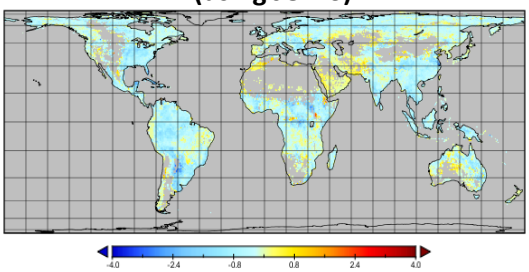


Fig. S3: Runoff highs. Note the differing scales: plots in top row scale ranges 0.0-4.0 extreme events per year (EE/yr) while the remaining rows ranging -4.0 to 4.0 EE/yr.

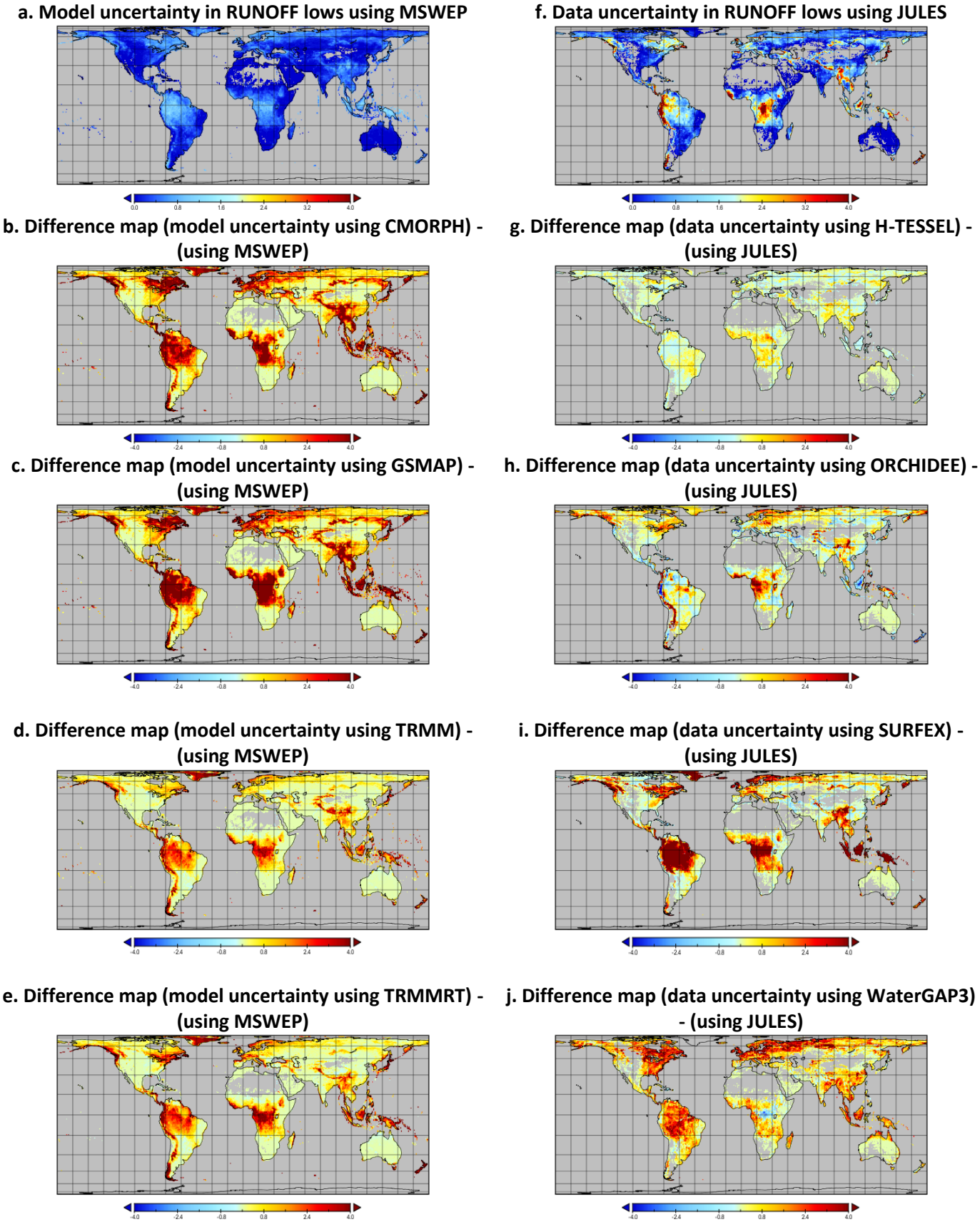


Fig. S4: Runoff lows. Note the differing scales: plots in top row scale ranges 0.0-4.0 extreme events per year (EE/yr) while the remaining rows ranging -4.0 to 4.0 EE/yr.

References

R Core Team (2018). R: A language and environment for statistical computing, (ed.), R Foundation for Statistical Computing, Vienna, Austria.

Zender, C. S. (2008). Analysis of Self-describing Gridded Geoscience Data with netCDF Operators (NCO). ***Environmental Modelling & Software*****, *23*(10), 4.**