Dear Editor,

We appreciate that the manuscript was well received but needed some improvements. Please, find our responses below to the suggestions and questions by the referees. A new version of the manuscript is now available.

In addition we had some comments by email from Dr. Hylke Beck, which we have also included in the revision of the manuscript.

Response to Editors comments:

*vague terms being used about what is 'successful';

The term 'successful' was used in three places in the text and has now been replaced as follows:

(1) In the abstract to: 'more useful results for water management'
(2) In the Result section 5.1: (removed)
(3) In the Discussion section 6.1 to: 'eventually'

*about the concepts of what is preserved in the model in terms of mass balance losses or gains in terms of fitting signals;

We are fitting the parameters but the mass (water balance) is conserved in the HYPE model concept. We have now specified this better in the model description (Chapter 2). We also included the following sentence with reference to a more detailed description of all algorithms in the code: "Parameter values regulate the fluxes between water storages in the landscape and interaction with boundary condition of the atmosphere and deep ground water aquifers (see detailed model documentation at hypeweb.smhi.se)."

*a discussion of how models can be compared across the GRDC and all the human impacts and the implications that has for understand process variability vs human impacts;

Thanks for this comment; it is truly a very important next step in model evaluation and improvement. We included this aspect as follows:

First, in Chapter 5 (Results) the WWH model results are compared with previous modelling results by Beck et al., 2016, who used the HBV model in selected basins across the globe. The text mentions the difficulties when comparing model results only using the literature as different river gauges and time-periods are being used, and we now also added a text explicitly saying that "…. more concerted actions for model inter-comparison are needed at this scale." (section 5.1)

Second, in Chapter 6 (Discussion), we have now added a paragraph at the end of section 6.1 (potential for improvements) discussing the importance of model inter-comparison and how to design such a study to better understanding process variability vs human impacts. In fact, such an initiative was already taken in a splinter session at EGU 2019 for the continental scale (Europe), but should be extended to also include the global scale.

*understanding and interpreting the KGE values.

We have now extended the text to better explain the interpretation of the KGE components in the model evaluation section 3.4, where the KGE is explained.

*I also feel that there needs to be a better evaluation as to why certain regions behave well or not in terms of increasing a scientific interpretation in the paper to be novel for publication…

First paragraph in discussion section 6.1 is now extended with two paragraphs describing more in detail the problems and potential for improvements in certain regions. We agree that this improve the scientific value of the paper. The spatial analysis is now organised as follows:

The spatial patterns and discrepancy in model performance are described in the Result sections 5.1. (especially the 4rd paragraph) and in 5.3.

In the Discussion section 6.1 we discuss major potentials for improvements to capture processes better in regions where the model fails. This section is now extended with examples of interpretations and references to other studies and on-going work in model evaluation.

A more thorough analysis would benefit from evaluation against independent data of spatial patterns of hydrological variables, for instance from Earth Observations. This is now mentioned in the text. In fact, such an extensive analysis has been performed, but is too extensive to be included here and will be described in another paper, soon to be submitted for scientific publication to continue the diagnosis of the model concept. We chose not to refer to this study yet, as we don't know when (or where) it will be published.

Response to Anonymous Referee #1

The discussion paper represents a very impressive body of research that is well summarized given the amount of information to be documented for the global application of a catchment model. The use of this global model to then gain insights into global water availability is an additional important contribution only achievable through the thoughtful and comprehensive modeling work presented here. I also want to make note of the well-referenced datasets used in this modeling effort that are nicely documented in the form of several tables and will provide an excellent reference for readers that have an interest in catchment modeling at the global domain. Lastly, I commend the authors on their candid discussions regarding the model performance, parameter estimation, data challenges, and limitations, which will serve to identify the critical research paths moving forward to improve catchments modelling at the global domain.

Thank you for this very positive statement and appreciation of our work, which was indeed very extensive and demanding for the team. We hope the results will be useful for further studies and the scientific discussion.

My comments for improvement mainly lie in providing additional detail on the methods and justification for choices in the modeling process. In some places, the language is feels rushed and I hope my comments also address this point.

Line 95-96: Define what is meant by a "multi-catchment approach for a large domain"

Thanks, this is now done when mentioned the first time: "… (i.e. nested catchment units instead of grids, and entire landmass coverage instead of isolated catchments)."

Line 199: Consider emphasizing here the difference and advantages of catchment scale modeling and other types of models so the importance and significance of catchment based models at the global domain is fully understood.

This is now better clarified in the introduction (line 99) by adding a few sentences on how catchments can be described as evolving and living units, not only as aggregation of separated building blocks.

It might be more logical to the flow of the paper to make Section 3.2 into Section 2, meaning to introduce the model first and then discuss the data and methods used to develop and evaluate the model.

The HYPE model is part of the Methods, but we have now changed the orders of Chapters into: 2. the HYPE model, 3. Data and 4. Model set-up. One more chapter is thus included.

Lines 136-143: How did resolve any spatial differences with the areas not covered by the higher resolution dataset?

We have now explicitly added for which region we had high resolution (60S to 80N) and that each of the mentioned datasets were used independently.

Line 148-150: How did you identify these flood risk areas?

This was achieved from UNED/GRID Europe (see Table 1) which is now better highlighted in the text.

Change the title of Section 2.2 to “Climate Data” and Section to 2.3 to “Hydrologic Data”

We have changed Forcing data to ‘Meteorological data’: Climate data would strictly be average for 30 year periods, while we are using daily time-series to force our hydrological model. Hydrological data, however, is a bit too wide as we have only used river flow and we describe the observed time-series in this section. Here we would like to stay with ‘Observed river flow’.

Table 3: Are all time series available at a daily time step?

No, there are also monthly time-series, which have now been added. More detailed information about the time-series can be found in the paper cited (which now has a DOI).

Lines 202-203: Are versions 1.0 and 1.2 published and citable?

No they are not published in the scientific literature and that’s why they are described with some details in this section. Although, they have been presented at EGU conferences.

Lines 210-211: How many catchments fell into the category of needing a longer initialization? Was it 10% (100% minus the 90% mentioned in line 209)? This is not clear. How was the screening done to pick these catchments needing a longer initialization?

We did not categorise catchments by initialisation requirements explicitly and used 15 years for all catchments. To avoid misunderstanding, we have now better explained how the judgement was done in the model set-up section.

Line 221: Is the WHIST also new to this paper or should there be a reference here?

This GIS software is not new, but not scientifically published. We inserted a link to the code in the text.

Line 271: Change “percolating” to “percolated”

Thank you – this is now changed!

Lines 275-276: Change to read: “encompassess a large number of sinks due to climate and topography and there existed a national. . ..”

Thank you – this is now changed!

Lines 284-288: How were these steps 1-3 done? More detail needs to be provided here.

The method has now been explained a bit more in detail by adding some information to each point.

Line 294-298: Can you offer some specific justification for this assignment was done for soil depth?

We have now made a reference to the description of the HYPE model, where this was mentioned first now, according to the new ordering of chapters, but we also added the references to Gao et al. and Troch et al. again, to remind the reader.

Line 318 could read “fluxes well relate to vegetation...conditions rather than soil. . .”

Thank you – this is now changed!

Line 324: Should read “while a gladter routine accounts for”

Thank you – this is now changed!

Line 326: Start new paragraph at the word “There”

Thank you – this is now changed!

Section 3.3 and Figure 3: Could you add more detail about how the process used to group catchments?

We have now entered a text to better describe how catchments were selected. It’s actually taking the representative gauged basin approach.

Line 367: Why was the DEMC preferred over other approaches?

We have now explained this in the text by adding: “ The advantage of DEMC versus plain DE is both the possibility to get a probability based uncertainty estimate of the global optimum and a better convergence towards it.”

Response to Anonymous Referee #2

Overall opinion:

Overall, this is a useful synthesis to all those who want to know the state of development of the HYPE model. The only problem I see is in the overly emphatic statements : I have no problem with publishing modest modelling results, or even complete model failures, provided modest results are called modest, failures are named failures. A mean KGE of 0.4 reflects a poor fit, but it may be the most that such a worldwide approach can bring. At this scale, we lack references. In the “model usefulness” section especially you should try to be more

modest. Start by stating that in many areas HYPE should still be considered as a scientific tool, and that it cannot be useful to managers because of its poor performances. Only for a small percentage (10% or 20 % ?), it is usable by a manager. Don't forget that managers have always high model efficiency expectations.

Thank you for this recommendation of being more modest; we have now moderated the Results section by saying that: "Catchment modellers would normally judge these results as poor, but given that global data…." (section 5.1) and also included the sentence suggested above in the Discussion section, section 6.2 on Model usefulness: "in many areas HYPE should still be considered as a scientific tool, and that it cannot be useful to managers because of its poor performances."

Much of the paper discusses the model errors and failures, so we do not try to hide this. We think that the judgment of model performance should be made in the light of other available source of hydrological data/information and the purpose of the model, i.e. what it should be used for. Managers are picky, but also prefer something rather than nothing. That's why we focus on potential usefulness, which is described in section 6.2.

We also appreciate that the model show similar or better performance than many other global models (we have some references in the text and discuss the need of more model intercomparisons at this scale), although, the model performance at the global scale is of course much poorer than local models using local input data and local calibration…

Nevertheless, we still believe that a global model can have some value and be improved over time, which is highlighted in the discussion. In fact the national Swedish HYPE model showed similar results for the very first model set up, but after ten years of improvements, this model now has an average NSE of 0.8 - so based on this experience we are rather optimistic also for the future of WWH.

Technical suggestions:

1. To be able to judge of the fit quality, we would have needed something like a classical (let's say HBV) lumped hydrological model applied as reference, in calibration mode, or even in ungagged mode with a single parameter set for the entire globe.

Thanks for this suggestion. We have now added one paragraph in the Discussion section 6.1 to promote model intercomparisons at the global scale, trying to identify some key elements in such a study. In fact, we have already tried to initiate this on the continental and global scale at a splinter session at EGU and within the ISIMIP community, but this is another story and will hopefully result in other papers.

However, in this paper we already do compare our results with reported results from using HBV by another research group (Beck et al. 2016) in the Result section 5.1, 3rd paragraph. The best median monthly KGE was then 0.32 for catchment scale calibration of regionalized parameters, using a gridded HBV model globally (Beck, 2016). In the text, we do recognise that it is difficult to compare results when not using the same validation sites or time-period and that more concerted actions for model inter-comparison are needed at this scale.

2. I am not convinced by your introduction much too long, not really informative, a lot of commonplace statements. You cite all the "politically correct" papers of the moment, but you could go straight to the point: you are a recognized group with a first worldwide application of your model. That's all.

True, but this depends on the reader. This paper could have a broad audience of global modellers, who are not familiar with catchment modelling or even hydrological modelling. It's therefore judged as important to set the scene and give the context of the study. This is normally valuable also for young scientists, who are in need of guidance to current state of the art and reference literature.

3. Using KGE is OK, but I would suggest to use a bounded version (between -1 and 1, see Mathevet et al. 2006) because we are not interested by large negative values.

We agree that this new KGE could be useful, but the original KGE is more commonly used and we don't want to confuse the reader - and we want to make it easy to compare our results with results from other studies reported in the literature. Moreover, the negative values are not highlighted in the figures (where the legends focus on classes between 1 and -1). In fact, less than 10% of the catchments have values below -1. All over, we only present and discuss median values for this reason – and they will have minimal impact by using the version of KGE that you recommended. Thus, we prefer to stay with the current KGE values, although we show large negative values at some sites.

Nevertheless, we have now changed the text to better motivate our choice of performance metric in the Method section 4.3 and also included the reference below in the paper. Thanks for this suggestion for improvement!

References

Mathevet, T., Michel, C., Andréassian, V. & Perrin, C., 2006. A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins. IAHS Red Books Series n◦307, pp. 211-219.