The manuscript proposes machine learning (with extensive observational data as input) to identify spatially and temporally varying values of parameters in an extremely simple forward simulation model of hydrology. A small number of parameters represents how water fluxes are partitioned over various storages (however they vary in space and time!). Results are evaluated 1) by comparing prediction error with existing global hydrological simulation models, and 2) by exploring spatio-temporal patterns in parameters values and providing possible explanations of these patterns in terms of processes.

The work presented is innovative. The authors have recently published an article on the same model and data sets (Kraft et al., 2020 – referenced in the manuscript), but the current manuscript extends on this by providing a more extensive evaluation of the results (and it seems minor adjustments in the methodology). Because of the innovation I am in the opinion this is a promising manuscript. However, it needs considerable revision in particular regarding the presentation of the material: figures are often very unclear (it is sometimes even unclear what attributes are shown), the text is rather long and could be condensed providing at the same time more focus. Regarding the latter: this seems to be a proof-of-concept paper. It is thus not essential to provide a complete evaluation of the model. Instead, I believe it is more important to properly explain the methodology and key outcomes. In revising the paper I suggest the authors to possibly leave some of the results out (e.g. less figures, simpler figures) – it wouldn’t harm the paper but may make it more accessible.

Response:

Thank you very much for appreciating the value of the manuscript and providing very relevant comments that will help us to improve the manuscript.

A couple of issues were also raised by the other reviewers, especially the clarity and complexity of the figures. To address that, we will revise the focus of Figure 5 and 6, and show the global means of terrestrial water storage and snow water equivalent (Figure B and C in the supplement to this response). The original figures for the regional variations will be moved to the appendix so that the interested readers can still access them. Additionally, we will only include the critical parts of the performance evaluation in the main text, while providing the details in the appendix. We also agree that the proof-of-concept here is beyond the performance evaluation, and also brings in the hydrological
responses within a data-driven prediction, which is rarely a focus of machine-learning based predictions.

My main comments are:

Figures (and tables)

Most of the figures are hard to understand. Legends are often missing, panels are included without any explanation of what is shown, variables are plotted for which it has not been explained how they are calculated. This needs considerable improvement. Please enlarge the figures as well. In my detailed comments I pin-point a few things that are not clear, please consider this as examples (not a complete list).

Response: As pointed out in the other reviews, the figures were, unfortunately, corrupted on upload. We uploaded the corrected figures, and thus hope that you were able to access them before the review. In the new figures, all the missing texts and legends were corrected, but still, some axis labels were missing. We will also try to make the legends more prominent wherever possible. Regarding the figure size, the HESS style guide proposes two sizes, one for 1-column, and one for full text width, which is still not really using the full vertical space. We try to improve this aspect as well and will coordinate with the HESS editorial team to provide the best possible quality of figures, when and if the manuscript reaches the final publication phase.

Objective function

The objective function contains four different observational data (terrestrial water storage, evapotranspiration, runoff, snow water equivalent). It is completely unclear how each of these are weighted in the objective function. Do you ‘calibrate’ against standardized values of these attributes? If so, how are these standardized? Please explain. Note that it is to be expected that this weighting has a strong influence on the results, e.g. if more ‘weight’ is assigned in the objective function to runoff, the model will perform better in runoff prediction. I do not expect you to explore different objective functions but at least the objective function needs to be given and it needs to be explained that this is quite arbitrarily chosen. Note that for instance in Bayesian data assimilation ‘weights’ of observations will depend on the uncertainty associated with the observations (high uncertainty -> low weight). This is not the case in your approach.

Response: To avoid too much overlap with the first publication (Kraft et al. 2020), we did not elaborate on the multi-objective approach. In L 219/220, we state: “The five loss terms were dynamically weighted using a self-paced task weighting approach proposed by Kendall et al. (2018)—we refer to Kraft et al. (2020) for more details.” The task weights (found by automatic task weighting) are shown in Appendix B.

Training, validation, testing data sets

In machine learning, one uses training and validation sets in the procedural step of model building. Model evaluation then is done using a data set not used for model building (this is often referred to as the test data set in the machine learning world). It seems you are not separating validation and testing data sets. This is an important issue – evaluation of the model (all the performance metrics provided, almost all plots provided with model outcomes) needs to be done on data that are not at all involved in the model building phase. If you are evaluating on the same data as used for building the models, this should be clearly indicated in the manuscript and implications of this extensively discussed.

Response: This aspect is discussed in Section 2.3 Model training (starting at L 200): We split the data in both temporal and spatial domains into training, validation, and test set.
For the spatial splitting, we asserted a minimum distance between the cells of one pixel between the sets (110km at the equator), which is a tradeoff between data limitations and cross-validation requirements. In the time domain, we split the data into two sets (2002 to 2008 / 2009 to 2014) and use the first one for training, and the other for validation and testing. We know that this is not ideal, but the data coverage is limited and we tried to find a good compromise. As stated in L 209, more details on the cross-validation are provided in Kraft et al. (2020).

For performance evaluation, we used the test set only (e.g., Table 3), while for the more qualitative evaluation, we decided to use the full time range covered by the observations. Figure D in the supplement to this response (not contained in the manuscript, we will add it to the appendix) shows the performance (RMSE) of the training, the validation, and the test set. From the figure, we see that there is no systematic difference in RMSE across the sets. We assume that this is a result of the regularization (e.g., early stopping, weight decay) and the physical constraints, which are designed to avoid overfitting.

However, to address the issue, we decided to only use the test years (2009 to 2012) for Figure 3 as this makes the comparison fairer and to explain why we use the full time range (2003 to 2012) for the qualitative evaluation/comparison.

Context, aim of modelling

I could argue that in your comparison of your approach to modelling (hybrid modelling) and existing global hydrological models you are comparing apples and oranges. The main aim of hybrid modelling (as defined in your manuscript) is prediction (in the statistical sense): estimating variables at time steps for which observational data are not available. Existing global hydrological models however aim at scenario analysis (and oftentimes prediction as well). Scenario analysis may involve evaluating effects of climate change, effects of future changes in water allocation (e.g., irrigation, domestic water use), effects of future changes in land use, etc. Hybrid modelling is not suitable at all for scenario analysis as it almost completely relies on observational data on the system. If the system changes, it won’t work anymore. I may exaggerate somewhat (to make my point clear), and you may disagree in which case I challenge you to convince me otherwise. In any case I suggest you 1) mention this difference in the introduction of the manuscript and 2) discuss this issue in the Discussion section. I consider this important in particular because this is a 'proof-of-concept' paper and it is thus important to position this work in the broader context.

Response: Thanks for an interesting comment. Perhaps, we may not have been clear enough in the manuscript, but the comparison of H2M with GHMs was actually intended as a validation of the H2M itself rather than as an effort to provide an alternative GHM. The holy grail of hybrid modeling would be to bring tightly together with the best of machine learning methods and the physical models. For example, the mass balance and hydrological responses in machine learning methods, while still having a robust data adaptability that is not always the case in physical models with process simplification and abstraction and rigidity of physical/empirical parameters.

Future predictions are a challenge in general even for physical models, as seen through numerous generations of model intercomparison projects. Even the validity of the “physical” parameters (that are often learned through iterations of model development) can be an interesting puzzle to solve. But, given that for GHMs, we still cannot state that machine learning methods are better. In fact, the data-space that the machine learning methods are trained with may provide a range of climate-spectrum that may be within the projected changes in climate. A particular aspect of the “interpolation” vs “extrapolation” by machine learning methods is discussed well in Jung et al. (2020, https://doi.org/10.5194/bg-17-1343-2020). Lastly, one could also argue that the data-
adaptivity of the machine learning methods gives them larger flexibility in dealing with non-linear responses (as touched upon in the hydrological responses section of the manuscript) of the land surface to changes in climate. We, however, confide that more research on the applicability of the hybrid models on scenario-type analysis is necessary. Only such research would be able to make a convincing argument rather than a speculative one we can make now based on the state-of-the-art. We hope that our proof-of-concept can contribute towards that.

As suggested, we will add relevant aspects of this comment and response in the revision.

**Detailed comments (please note my comments on the figures are not complete)**

p. 3, end of introduction

*The introduction gives a good overview of past work in the domain. However, on line 68-72) it remains somewhat unclear what the contribution of this paper is. I suggest stating this more explicitly and also to state more explicitly what this paper adds compared to your previous publication (Kraft et al., 2020).*

**Response:** We will improve this aspect in the revision.

**Code sharing**

*I strongly suggest sharing the code of your model on a public repository (e.g., GitHub). It will make your work more credible and will enable other researchers to build on your research.*

**Response:** The code and simulations are available online (see Code and data availability statement, L 653) at https://github.com/bask0/h2m.

p. 5, line 113

*Clearly state what Q refers to. It is the amount of runoff generated in a pixel (or area of land), not the discharge from the pixel (streamflow). The latter can only be calculated in a spatial model that does channel routing.*

**Response:** The runoff from the GRUN dataset is the streamflow measured at the outlet divided by the catchment area. The authors (Ghiggi et al., 2019) use catchments with a similar size as the forcing variables for model training and state that on a monthly scale, runoff equals the streamflow. We will add a short explanation in the revised version.

From the original paper (Ghiggi et al., 2019):

"Runoff is defined here as all the water draining from a small land area. Runoff cannot be observed directly, but at a monthly timescale the average catchment runoff can be assumed to equal the monthly streamflow measured at the outlet divided by the catchment area, provided storage of river water (e.g. in dams, reservoirs) and/or river water losses (e.g. river channel and lake evaporation, irrigation) are minimal. Thus, runoff rates (millimetres per month) are obtained by dividing the GSIM river discharge (cubic metres per month) with the station's upstream catchment area (km2). We then select catchments with an area comparable to the grid-cell size of the atmospheric forcing data in order to derive observational estimates of the runoff rate response to changes in atmospheric forcing."

p. 5, line 123
Adjust the numbering, is this a nested numbered list?

**Response:** The first mention of 1-3 lists the three criteria, the second time, the numbers refer back to the criteria. We agree that this is not needed, and we will improve this.

**p. 6, model description**

*Is this a spatial model, i.e. does it include spatial interactions in the time transition functions? I don’t think so, it is a local (point model). Please state this clearly.*

**Response:** There are no spatial interactions. This is mentioned here and there in the discussion, but we will make this clear in the model description.

**p. 6, figure 1**

*The caption is too long. Reduce it and explain concepts in the main text.*

**Response:** We will shorten the caption.

**p. 7, neural network**

*I am unsure you are building the machine learning model for all cells at once (single model) or for each cell separately (number of models equals number of cells). If the former, the method you propose is, I believe, fully non-spatial (point model for hydrology, identified separately for each cell with observations for that cell). Please explain this clearly.*

**Response:** We train one global model without between-cell interactions, but we do not explain this explicitly in the text. We will, thus, improve this aspect.

**p. 9, line 196**

*Why DELTA T instead of T?*

**Response:** Our notation is a bit imprecise. We actually model the T dynamics, but not the mean, we could therefore say that we model the delta T. But what we do is: remove the mean from T and S + G + (-C) independently, such that only the dynamics are modeled (because T is not absolute terrestrial water storage but just the variations). We will improve the notation. Note that calculating delta T is needed to compare with the GRACE observation, as it only provides the variation.

**p. 9, model training**

*I am wondering how you train the model. Machine learning models typically do not run forward in time. However, in this application, they are fed by temporally changing data, in a forward timestep approach. How is this done? Please explain. Providing the code would help as well.*

**Response:** The code is available here: https://github.com/bask0/h2m

We do a “many-to-many” prediction with a spin-up (random years from training features) and a warm-up (one year of input features in correct order, just the year before the actual model run) to equilibrate the model states. To understand the sequential processing of the model, we can ignore the “hybrid” part and think of a traditional hydrological model or a plain recurrent neural network (RNN). In both cases, a state (e.g., soil moisture, or the RNN state) is updated at each time step in interaction with the current inputs (e.g.,
meteorological variables), which yields an updated state (and a couple of output variables). In practice this is just a for loop (https://github.com/bask0/h2m/blob/04f0e15b60fe90d0c897aca2d801289bec158ebe/src/models/hybridmodel_loop.py#L171). The model does not predict into the future, as it does not simulate meteorological forcings, i.e., it needs inputs at time t to simulate response at time t. Figure E in the supplement to this response nicely illustrates the concept. On the left side, we see the “for loop” version, on the right side the unfolded version of the “for loop”. The outputs o are only present up to the last input element x.

We try to improve the description of the sequential processing in the revision.

**Figure 2: Too small.**

**Response:** We will enlarge the Figure in the revised version.

**Section 2.5: Which runs of the global hydrological models were used?**

**Response:** The ones originally published by Schellekens et al. (2017) that are available at https://wci.earth2observe.eu/

**Caption Table 3: What is ‘median-cell level’? Please explain.**

**Response:** It is the median across all grid cells, i.e., metrics are calculated per grid-cell first, and the median is reported. The spatially averaged signal is the global signal (i.e., mean across all time steps, then we calculate the metrics for the global time series). We will improve the terminology.

**Figure 3: Why are you not including results for ET and Q as well?**

**Response:** ET and Q are not direct observations, and are upscaled from local measurements, and can be classified as observation-based products. We assume that the performance on these signals is more prone to be affected by upscaling prediction uncertainties than those with direct remote sensing. ET and Q are both affected by substantial biases and uncertainties. We do provide ET and Q metrics in Table 3, and show the local biases in Appendix A, Figure A2. And, after all, as you have suggested, the manuscript is already quite long.

**Figure 4: Explain root phase and variance error in Methods. I am in the opinion however this plot could be left out.**

**Response:** We will consider your suggestion and either explain the metrics or remove the plot in the revision.

**Figure 5: What are the small panels on the left and right side? What is plotted on the x-axis? The figure is hard to understand. It needs to be simplified (and possibly include more detailed information in a digital supplement or appendix).**

**Response:** The panels (“inset axes”) show the NSE per model (color corresponds to model). If you refer to the outer panels, this is the mean seasonal cycle. We will only show the global signal in the main text and move the current figures (with improved axis labels and without the inset axis) to the appendix in the revision.

**Figure 7: What is represented by each line in the figure? A single location? A single year? ‘quantile of the spatio-temporal distribution’ is not easy to understand.**

**Response:** We “mix” all time steps and grid cells and calculate the quantiles for bins of
CWD. For example, we filter all the values (space and time) for CWD values from 0 to 10 mm. From these values, we calculate the quantiles (independently per cross validation run). Then we take CWD from 10 to 20 mm, etc. We will improve the caption in revision.

*Figure 8: What is represented by the colours?*

**Response:** The legend may be hard to spot (top left panel), and, so, we will move it outside. We also mention the color scheme in the caption.

*Figure 10: Too small. Consider selecting runs and plotting these.*

**Response:** We will improve the figure size, but we do not understand what “runs” is referring to here.

*Figure 11: Too small. What is along the x-axis? Months? It runs up to 13.*

**Response:** We will improve axis labels and make the figure larger.

**Discussion section:** The discussion is interesting but it is somewhat long. Consider reducing it somewhat focusing on the main things (that are relevant to the research objective and questions).

**Response:** We will consider shortening the discussion.

Please also note the supplement to this comment:  
https://hess.copernicus.org/preprints/hess-2021-211/hess-2021-211-AC5-supplement.pdf