Kraft et al. present a hybrid global modelling framework combining machine learning with simple water balance equations to simulate the most relevant hydrological states and fluxes at global scale and daily resolution. The model has been simultaneously trained with observational data-sets of total water storage, snow water equivalent, evapotranspiration and runoff. Results are evaluated against the same data-sets (split sample) as well as compared to simulations provided by four global hydrological models. The model intercomparison focuses on the attribution of water storage changes to variations in snow pack, soil moisture and groundwater storage. The authors prove that their hybrid modelling framework is able simulate relevant water storage components and fluxes with similar or better performance than state-of-the-art GHMs. Systematic differences between the models and the implications of their findings are extensively discussed.

The contribution is novel and fits well within the scope of HESS, however requires major revision before potential publication. This especially concerns the results section which contains a large number of (complex) figures that are often difficult to read and/or grasp. The reader is required to go back and forth between the main text and the caption in order to understand what is displayed. These comments are already based on the revised set of figures uploaded by the authors. I would recommend to thoroughly revise the results section, potentially drop a few figures and revise the remaining ones, in order to arrive at a more concise and digestable presentation.

The discussion touches many important aspects, however appears lengthy and repetitive at times. I would recommend revision to make it more concise. Further, the significance of the parameter estimates for process-based modeling is overstated in my opinion. The authors acknowledge that, due to the simple model structure and small number of parameters, parameter estimates will tend to compensate for insufficient or lacking process representations and uncertainty in the input data, which undermines their 'physical meaningfulness' and ability to describe specific processes.

**Response:**

We would like to thank you for appreciating the novelty of the study and for providing valuable comments.

We received similar comments from the other reviewers regarding the complexity of the
manuscript and the figures. We plan to simplify some of the figures and also move some parts of model performance to the appendix. This involves reducing the information density (e.g., Figures 6 and 7, see simplified versions in the supplement to this response, Figures B and C) or removing (e.g., Figure 4) figures and streamlining the discussion section, and unifying the terminology. We will also provide additional information on the machine learning model and training procedure, as several questions have been raised by you and other reviewers.

Regarding the relevance of parameters in process-based models, we wanted to highlight that the model parameters are a critical source of uncertainty as well (e.g., Liu and Gupta, 2007, doi: 10.1029/2006WR005756). Several related specific instances are mentioned throughout the text for covering different aspects of “parameter uncertainty,” such as equifinality issues, scale mismatch between the estimation of parameters (such as hydraulic conductivity in the lab vs the same at half-degree resolution), discrete vs continuous in space (e.g., vegetation type and a corresponding lookup table based on observations at plant/ecosystem level), and temporally fixed vs varying. In the revision, we will tone down the text wherever possible, while still maintaining the message that there are ample rooms for improvement regarding parameter estimations in process-based models.

Below, we provide detailed answers to your comments.

Specific comments

It remains unclear which thresholds and data-sets were used to identify cells with “high anthropogenic impact”. Groundwater abstraction is given as an example, however cells with extensive irrigation (irrespective of source) should be removed due to its effects on soil moisture and evapotranspiration.

Response: We used an ad-hoc solution based on Rodell (2019) [https://www.nature.com/articles/s41586-018-0123-1]. We removed all regions with only groundwater depletion (#7, #12, and #14). We will add this explanation in the revision.

It remains unclear if the parameters beta_s and beta_g were estimated by the neural network or were preset. In the latter case, please clarify how these parameters were determined.

Response: They are estimated as free parameters (in the sense of not being connected to data), i.e., they are not predicted by a neural network but rather by the optimizer. The gradient-descent based optimizer receives all parameters of the neural networks (NN) and the global parameters. In each optimization iteration, the optimizer can update the weights of the neural network, as well as the two parameters. We will elaborate on it in the revision.

It is my understanding that global, area-weighted averages of TWS, SWE, Q, and ET have been used to calculate the performances reported in Tab. 3 and Fig. 3. If correct, I cannot quite see the value in doing so. Both H2M and the GHMs aim to estimate hydrological states and fluxes in a spatially distributed manner, i.e. numbers based on a global average provide little insight into the models' performances. Further, jumping between global performance and cell median makes Sect. 3.1 rather hard to follow. I'd suggest to focus on cell median and to ditch the global numbers.

Response: The global signal is less affected by data uncertainties and reflects how well the model gets the global-scale patterns. We agree that the global signal can be misleading (e.g., Northern and Southern Hemisphere can cancel out parts of the signal due to opposite seasonality), but the ability to reproduce the regional and global signal is
still insightful, and they have been used extensively in other previous studies (Trautmann et al. (2018; doi: 10.5194/hess-22-4061-2018), Jung et al. (2010; doi: 10.1038/nature09396), Schellekens et al. (2017; doi: 10.5194/essd-9-389-2017), Humphrey et al. (2018, doi: 10.1038/s41586-018-0424-4). The best would be to report regional performance together with the cell level performance, but it would add more complexity to the manuscript. So, even some of the regional performance evaluations would actually be moved to the appendix (based on suggestions from other reviewers), and the main text would focus on the global signal.

The comparison of model performances between H2M and the GHMs in Fig. 3 seems little meaningful since the better part of the common time series (2003-2008) was part of the training data-set; particularly since NSE and MSE are closely related. In this regard, it would also be interesting to see a direct comparison of the performances achieved by H2M in the training period and the evaluation period, respectively.

Response: We decided to use the entire time-series because 1) longer time-series yield a more robust performance 2) the model performances of H2M were very similar in the training, validation, and test period. We checked the latter by comparing the model performance for the different sets. 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 avoid overfitting as well.

However, we agree that the performance comparison should be made on the test set. We decided to make a compromise: The splitting of the training, validation and test set in the spatial domain makes a comparison difficult, as we would need to show the GHM performance for the different cells corresponding to the test set separately as well. Thus, we will include all the cells (from training, validation, test), but only use the years 2009 to 2012, which were not used for training.

Fig. 3 is hardly readable, please rescale/revise.

Response: We will increase the figure size.

Fig. 4 shows performance metrics that have not been introduced in the methods section or used in the previous figures and tables which, frankly, is confusing. I’d recommend to stick with the performance metrics used earlier.

Response: We agree that it makes little sense to use new metrics here. The phase and variance error are insightful, but the message of the figure is that our model struggles in the same regions as the GHMs, a point that could also be made with NSE, for example. We also consider moving the figure to the appendix in the revision, or to removing it entirely.

The insets in Figs. 5 and 6 severely compromise readability and I’d suggest removing them. Further, the x-axis labels seem to be cut off. Please revise.

Response: We agree that the figures are hard to read/grasp. We will replace the figures with the global signal and move the original figures (with improved axis labels) to the appendix (see Figures B and C in the supplement to this response).

Fig. 7: Q is an unfortunate abbreviation for quantile here since used for runoff in other parts of the manuscript. Please revise. Which variable/quantity are the quantiles exactly based on? Please clarify.
Response: Regarding the abbreviation, we agree. For the calculation of the quantiles, 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.

Fig. 9: The masking color (black) and the darkest shade of the color scale are hardly distinguishable, please revise. In general, I feel that the figure conveys a similar message as Fig. 8. Given the overall large number of figures in the manuscript, this one could be dropped for conciseness.

Response: True, the masking color is not ideal, and we will change it to white. With regards to the similarity of Fig. 8/9: we will consider removing one of them in the revision.

Fig. 10: The masking color (grey) is also part of the color scale (equal contribution from all three components), please revise.

Response: Agreed, we will change the masking color.

Minor comments and technical corrections

Response: Thanks for the technical corrections, we will follow all your suggestions listed below and clarify where needed.

l. 89: Please rephrase "average content [...] of bulk density".

l. 95: Please rephrase "keep".

l. 127: What percentage of global land area do the remaining 12084 grid cells represent?

Fig. 1 (caption): Is $s_{corr}$ the same as $\beta_s$?

eq. 15: Shouldn’t the equation be $q_f = \alpha_f \times w_{in}$?

l. 229: Replace with "ratio of modeled and observed standard deviation (SDR)" for clarity.

Tab. 2 (footnote): Should read "interception"?

l. 280-81: "SDE" is not defined in the manuscript, should be "SDR"?

l. 285: Should read "provided"?

l. 307: Should refer to Fig. 5, not Fig. 6?

l. 311: $\text{SWE}_{\text{MSE}}$ should read $\text{SWE}_{\text{MSC}}$ and NES should read NSE?

Please also note the supplement to this comment: https://hess.copernicus.org/preprints/hess-2021-211/hess-2021-211-AC6-supplement.pdf