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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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

Minor comments and technical corrections

I. 89: Please rephrase "average content [...] of bulk density".
I. 95: Please rephrase "keep".
I. 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 $q_r = a_r \times w_{in}$?
I. 229: Replace with "ratio of modeled and observed standard deviation (SDR)" for clarity.
Tab. 2 (footnote): Should read "interception"?
I. 280-81: "SDE" is not defined in the manuscript, should be "SDR"?
I. 285: Should read "provided"?
I. 307: Should refer to Fig. 5, not Fig. 6?
I. 311: $\text{SWE}_{\text{MSE}}$ should read $\text{SWE}_{\text{MSC}}$ and NES should read NSE?