Minor revisions for the submission
hess-2021-211
"Towards hybrid modeling of the global hydrological cycle"

Kraft et al.

January 25, 2022

Dear Editor, dear Reviewers,

Your comments and suggestions helped us tremendously to improve the manuscript. We highly appreciate your time and efforts. Also, we are glad that you acknowledge our efforts to improve the study.

In this second review, we received comments from three referees:

- Anonymous #1: Suggested *accepted as is* with no comments.
- Derek Karssenberg: Response in Section 1.
- Anonymous #2: *Comments via email.*: Response in Section 2.

Please find our response to the comments below. In addition, several typos were fixed and minor changes were applied to improve text flow.

Kind regards,
Basil Kraft (on behalf of co-authors)

1 Response #1

General comments

(1.1) General comment *The manuscript has been thoroughly revised and the authors have dealt with most of my comments (except one, see below) in a satisfactory manner. In my opinion this is an important contribution as it provides a proof of concept for the approach and a roadmap for future research. Please find below my comments related to the revised manuscript:*

Reply: We are very grateful for your comments. Your previous and current suggestions helped us a lot to improve the manuscript.

(1.2) Context, aim of modelling *Related to my comment (and your response) on the original version of the manuscript, item 3.5 'Context, aim of modelling' I agree with your response to my comment. However, in my opinion a short (!) paragraph on this could be added to the discussion, in particular on how informative or valuable 'hybrid models' are expected to be compared to 'process-based models' (GHM). You could discuss the use case when models are used for prediction (or reconstruction) alone (state estimation), when models are used for improving our understanding of mechanisms, and when models are used for scenario analysis. Regarding the latter, I would like to note that scenario analysis does not only involve climate change scenarios (as referred to in your rebuttal), it also includes scenarios that may require a change in landcover/land use (e.g. due to biofuel expansion), a change in allocation of water over industrial, urban, agricultural water use, and possibly other policies or water management scenarios.*

Reply: We added the following paragraph to the discussion:
L. 680ff Global hydrological models are often used for different tasks such as the assessment of the water cycle at past and present, predictions for the future, for evaluating implications of, e.g., land use changes

by scenarios, and to gain process-understanding. In principle and technically, a global hybrid hydrological model can be applied for the same tasks while related simulations need to be interpreted with care. The strongest use case of H2M is the assessment of recent variations of the water cycle since it can act as a physically consistent yet data-adaptive bridge between heterogeneous global data streams and complements traditional data assimilation approaches. Interpreting predictions too far into the past or future can be risky when factors that are not represented physically play a role that had little impact during the training period (e.g., permafrost melting, CO2 fertilization of water use efficiency). Likewise, scenarios of, for example, different land use could make sense to conduct if the conditions represented by the scenarios have been represented during learning in some way while there always remains the danger that learned relationships by the neural network are just statistical associations rather than causal relationships ("shortcut learning", Geirhos et al., 2020). As we could show, gaining process understanding from the hybrid model can be feasible as the spatially and temporally varying coefficients learned by the neural network are plausible and partly very interesting. However, such uncovered patterns may rather represent hypotheses that should be tested with complementary approaches like physical process modeling, direct observations, or experiments.

(1.3) Abstract, line 14 *Runoff (Q) -> runoff generation (Q). In my opinion it is important to be very clear on the fact that streamflow discharge of large rivers is not modelled, neither is it used for constraining model parameters or process representations. I suggest changing this in Table 1 as well.*

Reply: We changed "Runoff (Q)" to "Grid cell runoff (Q)" in the abstract and in Table 1 to emphasize that it does not include discharge. We are hesitant to use the term "runoff generation (Q)" because it refers to a set of processes.

(1.4) Runtime *Runtime is quite a down to earth matter but it in practice it is important. Many GHMs have long runtimes and this is one of the reasons they are not extensively calibrated. What is the typical runtime (and hardware requirements) of a training (and prediction run without training) run for H2M? I suggest adding this somewhere in the manuscript (e.g. in the Results section and possibly in the Discussion section).*

Reply: We agree that this is a relevant fact and we added it to the manuscript:
L. 422ff An optimization run of a single cross-validation iteration takes about 6 hours, a forward run for all grid-cells and the entire period from 2002 to 2014 takes about 15 minutes. Each model was run on a NVIDIA Tesla Volta V100 16 GB GPU with up to 10 CPUs (Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz) for data buffering and background tasks.

(1.5) Figure 1 *In my opinion the figure does not clearly explain the general concepts, because it gives too many details (in the c) panel). Consider leaving out the equations and instead giving only parameter (and key variable) names. The figure should show the links between the hydrological model (in particular its parameters as these are linked to the neural net) and the neural network (and input data), not more. Equations are in the text so they do not need to be included in the Figure, in my opinion. Note that the figure gives α_{snow}, I think this should be α_{smelt} (equation 2).*

Reply: We simplified Figure 1 as suggested and changed α_{snow} to α_{smelt}. To indicate the usage of the coefficients and global parameters in the balance equations, they were added in parantheses. We updated the figure caption accordingly.

(1.6) p. 9, line 189 *'the neural network'. There is more than one. Same on line 205.*

Reply: The "neural network" refers to the entire module, we changed it to "neural network module".

(1.7) p. 12, line 278 *I suggest leaving out the URL here (www.deeplearningbook.org).*

Reply: The URL was removed as you suggested.

(1.8) p. 13, line 314 *The cross-validation setup implies that validation sets are in the spatial domain very close to tuning data sets; offset of 1 pixel. As pixels close to each other will be similar, it is not so surprising (arguably) the validation works well (and thus possibly overfitting); training and validation data sets are not completely independent, they represent similar combinations of states, fluxes and parameter* *values. An alternative would be training for instance in continent A (all pixels) and validating in the other continents. Please consider discussing this if you agree this may have an effect on the results.*

Reply: We agree with your statement. However, the discussion is already quite lengthy and the topic of cross-validation with spatio-temporal datasets is a complex one. Of course, we do not reduce the dependeny completey, but we tried to chose a “fair” setup for the comparison with the GHMs. From different experiments, we got the impression that the physical constraints and the multi-task training exert a strong regularization on the model, which reduces the data-adaptivity and thus overfitting. This is, however, just anectodal and needs further investigation. We added the following statement to the manuscript:
L. 327f Note that the spatial splitting reduces the dependency between the cross-validation sets, but does not completely remove it.

(1.9) p. 13, line 322 *Validation runs. You may want to emphasize somewhere that the approach also enables time varying parameters when no tuning is done. This is because the neural network does a state update just like any other forward simulation model. I am sure you find this obvious, but it may not be that obvious for some readers*

Reply: We added the following sentence:
L. 174f For inference (after the optimization of the neural network), the model can be applied to unseen data like any forward simulation model without further model tuning.

(1.10) Figure 6 *This is a very interesting figure of course. It shows how model parameters change with soil water deficit. This is extensively described in the text. However, it also shows considerable variation between pixels (or time steps) with the same soil water deficit (e.g. completely wet soils may have an α_{soil} between 0.2 and 0.8). I am wondering whether this variation entails mainly spatial variation or mainly temporal variation (or both). Please consider adding this to the manuscript (without extending it too much).*

Reply: We added the following statement to the manuscript:
L 489ff The relatively large variation under wet conditions (low CWD) in Fig. 6 can be attributed about equally to temporal and spatial variability. The groundwater recharge fraction α_{gw} shows a slightly larger temporal variability than the other fractions, and the contribution of the temporal component was generally a bit lower in the transitional regions.

(1.11) Figures *In general, the figures are still quite small, please be sure they appear in the right size in the final manuscript.*

Reply: We revisited every figure and improved readability. The latex template wants us to use 8.3cm for one-column and 12cm for full width figures, which is only a fraction of the full width, and we tried our best to follow these suggestions. It will not be an issue in the online version, and we think the figures are still readable in the print version.

(1.12) p. 23, line 506 *PC-GLOBWB → PCR-GLOBWB*

Reply: This typo was fixed.

(1.13) p. 27, line 585 *indicates → indicate*

Reply: This typo was fixed.

(1.14) p. 41 *Consider acknowledging the reviewers.*

Reply: Thanks for pointing this out. Of course, we should acknowledge the efforts taken by the reviewers and the editor.

2 Response #2

(2.1) General comment *The authors have done an excellent job at revising their initial submission. The manuscript has been restructured and the figures have been improved in way that makes it concise and easy to follow (which was the largest shortcoming of the initial submission as noted by all referees). All of my comments on the initial submission have been addressed in the rebuttal letter and in the revised manuscript and I recommend publication in HESS.*

Reply: Thank you very much for taking the time to help us improve the manuscript and for appreciating our efforts.

(2.2) Eq. 21 *L_v should read L_C?*

Reply: The lefthand side of the equation was $\mathcal{L}_{v=C}(f_{\boldsymbol{\phi},\boldsymbol{\beta}}, \boldsymbol{x})$, with $v = C$ in the subscript. To avoid confusion, we changed the equation (now Eq. 20) as you suggested.

(2.3) L. 489 *refers to arid and semiarid climates in Fig. 7, however the classification distinguishes boreal, temperate, transitional and tropical. Please clarify.*

Reply: We agree and changed the sentence. It now reads:
L. 495f In arid (S1-2) and semiarid (N1-5) climates, α_{et} exhibits a large range with steep gradients given low water input ($w_{\text{in}} = 0$ mm), decreasing with larger CWD (dryer soil).

(2.4) L. 526 *Which data biases are meant here?*

Reply: We added a paragraph on data biases and uncertainties:
L. 545ff In the hybrid modeling framework, the quality of the observational constraints is a major source of uncertainty. The data used in this study have well-documented deficiencies: The precipitation product, for example, shows large uncertainties in Africa due to limitations in density and quality of measurement sites (Sylla et al., 2013) and exhibits biases in snowfall estimates in the Northern Hemisphere due to over-correction of snowfall under catch (Behrangi et al., 2016; Panahi and Behrangi, 2019). The GlobSnow SWE saturates above 120 mm and underestimates the interannual variability (Luojus et al., 2010). TWS quality is generally difficult to quantify as an equivalent ground-based measurement does not exist, and its complex preprocessing has known impacts on the data quality (Scanlon et al., 2016). The machine learning–based constraints of Q and ET are not directly observed and thus, they are expected to have considerable global and regional uncertainties and biases (Ghiggi et al., 2019; Jung et al., 2020). This could lead to inconsistencies in the water balance (Trautmann et al., 2021). However, the multi-objective optimization may dampen negative effects of biases, as the model can trade off the different constraints.