Comments by Keith Beven

Thanks for handling our manuscript. We are especially thankful for getting these great colleagues as reviewers.

1. *Line 450 and Section4.4 - you give no references to use of HBV with uncertainty estimation*

Added

2. *L500 Wide use does not imply that parameter variations between regions or catchments are small, and seems immediately in conflict with Line505.*

We agree and reformulated this section.

Comments by Jens-Christian Refsgaard

Thanks for these valuable comments on our manuscript. The reviewer is right that we, in this manuscript for the special issue on History in Hydrology, took a somewhat inward look by focusing on the history of the HBV model. A full review of the history of catchment modelling would be beyond the scope of this manuscript, but we will try to be a bit more self-reflecting in the revised version.

Beyond we shortly discuss we addressed the points raised in the reviewers list:

1. Yes, the term hydrological modeling is broader than we use it here. Your suggestion was to change this to rainfall-runoff modelling. As snow is an important part in many HBV studies, this should be precipitation-runoff modelling, or perhaps even precipitation-temperature-runoff modelling – but then there is also potential evapotranspiration as input. Therefore, we changed this to 'catchment modelling' and clarified at several places that the focus of models such as HBV is catchment runoff.
2. Yes, we agree that for other types of hydrological models other issues than those discussed here from the aspect of (conceptual) catchment models will be important. We clarifed the more narrow focus of the discussion in our manuscript (see also point 1).
3. Good point, we did not want repeat the full (interesting) discussion between these papers, but now refer to this discussion.
4. Yes, we realize that our use of the term hydrological modelling might cause confusion as we are focusing on catchment (runoff) models here (and not all the other types of hydrological models), we clarified this throughout the manuscript
5. Thanks
6. Agreed, we changed the heading to 'Early catchment models'
7. We here wanted to express that the differences between a simple model and a model based on equations such as Darcy might not be so large if the latter is applied using elements or parameterizations that do not capture the real-world heterogeneities. We clarified this section.
8. Good point; we implemented these thoughts in the revised version.
9. We agree and clarified these important details of Breuer et al. (2009) in the revised version.
10. We agree that many hydrologists working with modelling would say that all kinds of catchment models need some form of calibration. However, we also still meet the notion that more complex models with 'measurable' parameters would not require calibration or at least not as much calibration (thus, using other terms such as tuning) by colleagues working with these, in their eyes superior, models. We appreciate the nuanced discussion in

Hrachowitz & Clark (2017) and also the reviewer statement that the view on the need for calibration has changed over the years. We included the H&C reference and also added a pers.com. reference to the reviewer to include the changing-ideas-on-calibration point.

11. The (international) success of the HBV model is in our opinion much related to the 3 P-s (parsimony, performance, persistence) as mentioned in the introduction and above in section 3. The parsimony made the model code easy to understand (and re-program), the HBV model often performed well especially in comparison studies, and it certainly helped that the model was intensely used by SMHI and the hydrological services of Norway and Finland, and became a standard tool for the Nordic hydropower industry (opposite to a single researcher or research group as was the case for other models). We added these thoughts in the beginning of section 3.
12. We agree that new sources of data (including spatial patterns derived from remote sensing) might open new opportunities, but as we write there is a challenging balance between more available data and more model complexity to make these data directly usable for model evaluation. We argue that this applies to any type of catchment model but also see some promising studies such as the one by Stisen et al.. We added text in section 4.4 to clarify that we see multi-criteria calibration as a suitable way forward although this approach is challenging in practice.

Review Axel Bronstert

We appreciate this detailed review of our manuscript. We are glad that our somewhat unusual style is seen as positive. The reviewer lists three more general remarks and a number of detailed comments. Below we discuss how we will address these in the revised version of the manuscript (reviewer comments in blue italic font)

"Physically based" vs. "conceptual" ?

This is an interesting question. We agree that there is no clear distinction which makes some models physically-based and others not. However, as discussed in the review of Jens Christian Refsgaard, there is a difference in the degree to which models are physically-based. We fully agree that in the end, the crucial question is, which situations/tasks a model is suitable for. We clarified this in the text.

Number of parameters and desire for optimization

There is a difference between how many parameters are in a model and how many are actually allowed to take different values. As discussed in the review, this (also) depends on how much spatial variability is considered. We added some text on the discussion of when to use which type of model.

Need for a classification of hydrological models

As also noted in the review of Jens Christian Refsgaard, we used the term hydrological models in a rather narrow sense and clarified that we in this manuscript are focusing on catchment (runoff) models. We agree that also within these models there are different 'families' and extended the text on model classification.

- Line 115: "advocation for realism in model development"? What is 'realism' in this regard? And Is this still valid for today?

We changed the word 'realism' to 'best practices'

- Line 130: Instead "represent all hydrological processes", better term it "represent all relevant hydrological processes",

Agreed, changed

- Line 285: what do you mean with "re-organising observed flood generating factors'?

We changed 're-organising observed flood generating factors' to 'combining observed flood generating factors such as heavy rainfall, extreme snowmelt situations and wet antecedent conditions'

- Line 300: "risk for overparameterisation"? You may discuss it in the relation with complexity of model and nature? Also nature may have several means to come to one state? What can one do in this regards?

We added 'i.e., the inability to determine one single best parameter set' for clarification

- Line 315: "groundwater dynamics"? You may elaborate a bit how far the catchments wetness in the HBV model can be related to observed gw-dynamics.

Groundwater dynamics were adressed in some early applications of a modified version of the HBV-model as described by Bergström and Sandberg (1983) and we have used groundwater dynamics as additional criteria in previous studies (e.g., Seibert, 2000, HESS). We added text in sections 3.1 and 4.4 and discuss opportunities and challenges of multi-criteria model evaluations.

- Lines 465-480: about the application for ungauged catchments: You mention the option of ensemble applications (many sets of parameters). You may elaborate a bit more on this. For my understanding this is a very valuable approach.

An early indication of the value of using ensembles for HBV model simulations can be found in Seibert and Beven (2009, HESS) where we found that the ensemble mean performed generally better than the best single parameter set (see Figure 2). This finding was more recently confirmed in studies such as the two we refer to here (Seibert et al., 2018; Pool et al., 2021). We added a few words and references on this approach.

- Figure 6: The value of a model depends on its purpose!! If the purpose is to look like the reality, the left figure is much "better".

We agree, but the question we posed is 'Which model will fly?' – and trying to let the left plane fly will not end well ☺

Comment Stein Beldring

We fully agree that many HBV concepts can also be used in distributed approaches. We appreciate the reminder us about these papers which we added as a good example on a distributed version of the HBV concept.

Comments by András Bárdossy

We appreciate the positive assessment of our manuscript and the helpful list of comments. Especially we like the statement, "This is a nicely written story about the HBV model in the context of hydrological model developments in the last 50 years. The style of the paper is somewhat unusual - casual but certainly acceptable." Here we reply to the seven points raised by the reviewer:

1) Good point. We are afraid to make the manuscript harder to read if we insert the model equations. However, we added a reference to Seibert&Vis (2012) and could also add an appendix similar to the one in Seibert (1999).
2) We agree on this potential for physically-based models like SHE, although we might be less optimistic that the additional constraints (always) compensate for the increased degrees of freedom. But of course, there are applications where we need a better representation of internal fluxes and states. We clarified this in the revised version.
3) Thanks for making us aware that we missed discussing evapotranspiration in any detail. We added some information on evapotranspiration, especially how to derive Epot for the input, in section 3.1.2.
4) The often limited sensitivity to random errors is indeed an advantage of HBV, but also other water balance accounting errors. We agree that this is another reason why such models often perform well despite not-so-good input data. We see this related to one of the three P-s and added this valuable thought to the discussion of model performance.
5) Thanks for making us aware of this interesting study. We added the use of HBV for land-use change studies as suggested by Hundecha and Bárdossy (2004) in section 4.1
6) With “insignificant subroutines” we meant subroutines that are not affecting the (runoff) simulation substantially. Following the comment, we changed this to “subroutines representing insignificant process details”
7) We agree that it is an interesting question how models like HBV can make use out of different types of data and tried to make this clearer in the revision.