

Review of Manuscript

'Characterizing and reducing equifinality by constraining a distributed catchment model with regional signatures, local observations, and process understanding'

by C. Kelleher et al.

Dear Editor, dear Authors,

I have reviewed the aforementioned work (version 2 of the manuscript). My conclusions and comments are as follows:

1. Scope

The article is within the scope of HESS.

2. Summary

The authors present and apply a framework for calibration of distributed hydrological models by applying a hierarchical set of parameter constraints and error metrics to accept or reject randomly generated parameter sets. The constraints and error metrics are distinguished by i) range of applicability (from regional to local), ii) 'softness' (from local observations to heuristically formulated local expert knowledge) and iii) the evaluated characteristic (evaluation of non-dynamical to dynamical aspects).

The framework is presented at the example of the physically based, distributed DHSVM model applied to the 5.5 km² Stringer Creek catchment, whose hydrological behavior is dominated by seasonal snow accumulation and snow melt.

Using 10,000 randomly drawn model parameter sets, 10 signatures and error metrics are applied individually, in groups and in hierarchical combination to identify behavioral parameter sets from the initial set. The resulting subsets are then discussed with respect to the equifinality reduced (i.e. by how much the initial parameter ranges were narrowed).

The authors show that by jointly applying all criteria considerably narrows the behavioral parameter sets (here: to nine). However, these still show large differences, specifically with respect to catchment groundwater table (values and spatial patterns).

From the analysis, the authors conclude that i) a multi-criteria approach to identification of behavioral parameter sets is superior to single-criteria approaches, ii) dynamic constraints to be more effective than non-dynamical ones, and iii) that despite substantial narrowing of the parameter space still large differences among the surviving parameter sets remain, especially with respect to spatial patterns of hydrologic states.

3. Overall ranking

The work is ranked 'Major revision'.

4. Evaluation

Major points

I like the work presented in this paper, especially the strong argument towards using multiple, hard and soft sources of information to identify behavioral model parameter sets, and the thorough literature review. However, despite the fact that the authors' focus in this paper is to present the concept, with the choice of the catchment and time series used, they have clearly missed some very good opportunities to make their results more general and interpretable. More specific:

- Judging from the presented time series of discharge and snow water equivalent, the catchments' hydrological function is very simple (one major discharge event during snowmelt, rainfall-runoff events are hardly playing a role). Arguably, a very simple conceptual hydrological model could reproduce this behavior at least as well as the applied model with respect to all discharge-related signatures and metrics, but with much less parameters.
 - So why choose this very simple-behaving catchment if the goal is demonstrate the usefulness of a targeted constraining approach for a distributed, physically based model? This way, the model stays well below its potential, and this also means a lot of opportunities for more targeted model parameter evaluation are missed.
 - Furthermore, the many degrees of freedom in your model inevitably lead to problems of equifinality, which would not exist in a simpler model appropriate for the simple catchment. So why not choose a more complex catchment, which requires a distributed model?
- Along the same lines: Why was a catchment selected with such little available observations as 'hard truth'? Why not choose one with a network of observed groundwater tables, ET, nested discharge observations, spatially distributed information on soil type and soil depth etc.? In fact, the model elements are all set to the same soil depth and soil type, which makes it much less distributed as it could be. Placing the study in a better equipped catchment would offer the opportunity to fill the very nice framework with many more signatures and metrics, especially those evaluating spatial patterns. Furthermore, this would have opened the opportunity to compare the value of constraints formulated as aggregated/heuristic expert knowledge to 'hard' constraints based on observations. This is a clear miss.
- Along the same lines: All evaluations are done for a single year, and for the calibration period. This way it is impossible to judge
 - to which degree the remaining behavioral parameter sets are dependent on the chosen calibration period, and
 - whether the behavioral parameter sets found in calibration are still behavioral during a different, validation time. This is a clear miss.
- The authors advocate a multi-criteria approach to identify behavioral parameter sets. However, looking at the criteria in Table 2, two questions arise
 - Parameterization of ET plays an important role in the model (parameters 22 to 53 in Table 1). However, ET is only used as a very weak constraint (300-650 mm/a). Why not also evaluate the model with respect to ET error, ET timing, ET peaks etc.? From the text, my understanding is that ET estimates from local observations exist. This would be another important and independent criterion.
 - I do not understand the usage of AI: From my understanding of the text, both PE and P are from observations, so AI is independent of the model. So how can this criterion be used as a signature for model evaluation?

Minor points

- From my experience, the main control on parameter equifinality is model structural choice. The authors discuss this important issue briefly in the conclusions. I encourage them to discuss this aspect in more detail, although I am well aware that model structural choice is not the topic of the paper. However, it can offer an avenue of progress to reduce the still-high equifinality of the final behavioral set of parameter sets.
- P8/I29: RR instead of PET?
- P12/L17: Where is Appendix D?
- P13/L3: Why not also compare simulations to well observations? (see also my major comments)
- P13/L7: Appendix B3 instead of B only?
- P14/L22-23: ... suggest that evaluating certain types of internal behavior by point observations...?

- P21/L14: Figure B3 instead of Fig 9?
- Fig 1:
 - A) The location of the label 'Tenderfoot creek' is misleading
 - A) is the scale really [km]?
 - C) add a legend (which gauge is which)
- Fig 4, caption: remaining instead of removed?
- Fig 6, caption: Where are the black dotted lines? Where is (a) and (b)?
- Fig 7: Please add year indicators (2007, 2008), a legend (which gauge is which) and for clarity add in the caption that these are plots for the final subset of 9 parameter sets
- Fig 8, caption: For clarity please add in the caption that these are plots for the final subset of 9 parameter sets.

Yours sincerely,

Uwe Ehret