Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2017-121-RC1, 2017

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Multiple domain evaluation of watershed hydrology models" *by* Karthik Kumarasamy and Patrick Belmont

J. White (Referee)

jwhite@usgs.gov

Received and published: 29 April 2017

This manuscript deals with investigating the ability of a hydrologic model to reproduce observed streamflow time-series. Several new metrics are proposed to distil/process the simulated vs observed hydrographs, including Q-Q plots, and wavelet coherence, among others. In general, I believe the authors present an interesting and needed case for using purpose-driven calibration metrics. Below are a few comments and questions that I think are relavent to this manuscript.

You state "Models can be developed for a variety of applications and the metrics used to evaluate them should be targeted to the application of interest". To me, this is the primary reason for using different calibration/conditioning metrics: improve the model's ability to simulate its purpose. To that end, I think it would be useful to state the purpose

of the two example models and then (at least) allude to how using the newly proposed metrics to judge the model's fitness are related to these purposes. Additional and new ways of comparing model outputs to observations have been shown to be an effective way to reduce bias and uncertainty in quantities of interest, but only if the new metrics are "similar" to the quantities of interest.

Why not place the calibration problem you describe into the context of a Bayesian inference problem? Many of the issues you describe related to parameter types and uncertainty and equifinality (posterior parameter distribution) could be stated in the context of Bayesian concepts.

You are focused on forming multi-component, subjective metric(s) to gauge a model's ability to simulate the past. This sounds similar to the purpose of GLUE. Why no specific mention of GLUE?

How can you use the analysis of parameters in the time and freq domain in an automated sense? That is, can a practitioner use the HydroME toolbox in a scripted (batch) workflow so that time-based and freq-based metrics are output for use in a multi-component metric calibration (or a multi-metric calibration)? For example, TSPROC (Westenbroek and others, 2012) code allows users to programmatically calculate time-series statistics. Also, is there a reason not put the HydroME toolbox on github so that the community can extend and improve the package?

page 2, line 3: I think of calibration as the process of reducing model parameter uncertainty by assimilating information in observations.

page 2, line 13: I think gathering more data will help with parameter uncertainty (non-uniqueness) and therefore forecast uncertainty by informing more forecast-sensitive parameters and also breaking up correlation(s) between forecast-sensitive parameters. This is true even for conceptual (non-physical) parameters - do you agree?

page 2, line 27: I agree that different and model-purpose-specific metrics should be

employed for conditioning model parameters. How do you know that these new metrics are aligned with a given model purpose?

page 5, line 17: how are you defining parameter "sensitivity"? Parameters sensitive to the observations or parameter sensitive to the quantities of interest (forecasts) or both?

page 5, line 25: you are describing the process of defining the Prior in a Bayesian sense, why not call it that?

page 5, line 9: this is true of maximum likelihood estimation. However, if prior information is enforced (through parameter ranges, preferred parameter relations, Tikhonov regularization, etc), then parameters should be algorithmically adjusted in accordance with expert knowledge.

page 6, line 15: your discussion of the problem of equifinality is key here. You show that given the available data, many combinations of parameters reproduce the past equally well. The burning question in my mind is how does this large posterior uncertainty influences the forecasts (quantities of interest) - the model's purpose? If the quantities of interest are not sensitive to these parameters, then I would wager that this equifinality does matter.

page 7, line 4: I believe the problem you are alluding to with using sensitivity alone is related to the inability of sensitivities to capture and convey correlation information, right? Might be worth mentioning that specifically.

Figure 4 - what are the blue boxes meant to draw out?

page 9, line 18: I like the idea of allowing uncertain parameters to be adjusted more than parameters which defined to be less uncertain. I believe this is what the Prior parameter distribution and/or Tikhonov regularization attempts to achieve. However, adjusting only subset of parameters while arbitrarily holding other parameters fixed may cause other problems related to parameter compensation. That is, while adjusting the subset of parameters, this subset of parameters are adjusted to fit the observations

as well as possible, even if the information in the observations should in fact be used to inform one or more of the fixed parameters. Then these parameters are "fixed" and the next subset is adjusted to reduce the remaining residuals, and so on. I suppose your multi-component metric should combat this issue to some extent, but why not build your expert knowledge about the different categories into the Prior and let a well-designed algorithm find an objective solution to the inverse problem that respects the specified parameter uncertainty and parameter relationships?

page 9, line 23: the problem of equifinality is a result of information deficits related to model input uncertainty. That is, given what a modeler "knows" about model inputs (parameters) and what observations are available for conditioning/calibration, any QOI-sensitive parameter correlation/uncertainty that exists after calibration is going to give rise to equifinality. I don't see how categorizing parameter and then adjusting parameter categories sequentially can affect equifinality. It seems like the primary weapon to combat equifinality is more and diverse types of observations and ways to extract/distill different components of the observations, which is one of the topics of this manuscript.

page 11, line 11: I really like this statement - it needs to mentioned elsewhere in the manuscript. The reason for using other metrics to judgment a model's fitness for use is because we need metrics that are more aligned with the "application of interest".

page 12: line 16: If the model purpose of the model is to simulate "summer base flows" then shouldn't the calibration metric be directly related to "summer base flows" and not anything else?

Westenbroek, S.M., Doherty, J., Walker, J.F., Kelson, V.A., Hunt, R.J., and Cera, T.B., 2012, Approaches in Highly Parameterized Inversion: TSPROC, a general time-series processor to assist in model calibration and result summarization: U.S. Geological Survey Techniques and Methods, book 7, chap. C7, 79 p., three appendixes

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2017-121, 2017.