Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2017-121-AC1, 2017

Hydrology and Earth System Sciences
Discussions

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

Karthik Kumarasamy and Patrick Belmont

karthik.k@aggiemail.usu.edu

Received and published: 19 May 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 relevant 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.

Response: Thank you for the suggestion. We agree that we should more clearly discuss the purpose of our models and will add content to address this missing information. Each of the tools in the HydroME toolbox were developed to address particular issues, for which we needed more targeted evaluation metrics. Highlighting our own impetus for developing the tools will help ground them in reality for readers. In the revised manuscript we will provide 2-3 concise paragraphs explaining the purpose of the models and link them directly to the tools developed. The Le Sueur River Basin (LSRB) model was developed to evaluate the impact of land use change from predominantly prairie wetlands to tiled row crop agriculture primarily to evaluate changes to flow and its impact on turbidity. The model is also used, in two publications in preparation, to evaluate the effectiveness of different sediment reduction strategies. In the SWAT model, sediment is routed using a modified version of the Bagnold stream power equation which is a function of peak flow (Neitsch et al., 2011). As several studies recommend, we calibrate for sediment after the model is calibrated for streamflow (e.g., (Santhi et al., 2001;Engel et al., 1993;Arnold et al., 2012)). Errors that are not resolved or minimized during streamflow calibration will carry over to sediment calibration. This can misrepresent the sediment processes as they will be adjusted to compensate for flow errors. Hence, we used wavelet analysis to evaluate if modeled flows in the form of shape of the hydrograph matched measured flows on any particular day. The Root River Basin (RRB) model was developed to inform a sediment budget and evaluate the effect of water retention structures (wrs) in the landscape and their role in the changing the flow regime. Flow duration curves (FDCs) show how the flow regime of modeled flows compare with that of the measured flows. FDC served as a tool to provide confidence in the modeled scenarios. In other words, it allowed for a relative comparison

of flows while evaluating scenarios. For example, let's take the case of assessing a particular wrs scenario on flow changes. If the scenario predicts that there is a change corresponding to certain flows, comparison can be made with the fdcs of measured and calibrated model flow. We used this in a relative sense to explain the change in any particular flow type. Although, the use of fdc in the context of calibration is not particularly new (e.g., (Westerberg et al., 2011), its use in calibration is very limited. Similarly, by narrowing the NSE or R2 to evaluate specific seasons using box plots, we can assess the effects of wrs structures on flow in a more targeted manner. In addition, the box plot representation of NSE and R2 also allows us to see a summary of how each of the seasons (or time periods of interest) performed. As both these metrics are biased towards larger flows, they are sensitive to the large flows within in each time period of evaluation. Additionally, analysis in the frequency domain allows us to see if there are any systematic discrepancies in model simulated flow. For example, mismatch corresponding to flows related to particular periods (vertical axis in the wavelet coherence plot). When we view the flow differences in the time domain (Euclidean metric), we only see the magnitude, which cannot be compared to evaluate systematic discrepancies. This is because magnitude differences cannot capture the local shape of the curve. Further, by combining wavelet analysis with the calibration parameter candidate selection classification approach, we can identify parameters that are sensitive to certain periods and not others. This can further help limit the number of parameters that should be calibrated.

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.

Response: Yes, posing the calibration problem in a Bayesian construct can certainly be one approach. We wanted to approach calibration parameter candidate uncertainty in a simple way by means of separating the parameters to include/exclude for calibration in

sequential order. Additionally, with the number of parameters continuously increasing in distributed hydrologic models to describe new processes and posing the calibration problem in the context of pdf would mean considerable additional information, much of which may be unavailable. We wanted to illustrate an approach that does not need additional information and can be an alternative to other approaches.

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?

Response: As we are not evaluating uncertainty using a pseudo-Bayesian approach as implemented within GLUE (Mantovan and Todini, 2006), we did not specifically mention GLUE. Another fundamental difference between GLUE method and the approach presented in this paper is not solely rely on sensitivity for parameter selection (as GLUE method includes parameters as calibration candidates based on a sensitivity analysis (Montanari, 2005;Beven and Binley, 1992)). However, we agree with the reviewer that there are useful parallels between the approaches, so we have briefly discussed GLUE in the revised manuscript.

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 multicomponent +metric calibration (or a multi-metric calibration)? For example, TSPROC (Westenbroek and others, 2012) code allows users to programmatically calculate timeseries statistics. Also, is there a reason not put the HydroME toolbox on github so that the community can extend and improve the package?

Response: HydroME is a general purpose post processing tool with the ability to process and compare time series. It is model independent and currently cannot be used in automated fashion in batch sense. That may be possible in the future, but was not feasible given our focus on our science objectives and lack of funding to generate such

a production grade tool. HydroME source code is now available in a Git repository. https://github.com/Kkumarasamy/HydroME

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

Response: Yes, we agree in the sense that we will learn what combination can result in a model that predicts measured outcome. The process may also provide us with model structures (and the parameters that they represent) that can simulate measured outcomes.

page 2, line 13: I think gathering more data will help with parameter uncertainty (nonuniqueness) 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?

Response: Yes, non-uniqueness problem can potentially be reduced, but not eliminated, with more data. This may even be true for conceptual parameters. What we meant to say in that sentence is: For a given model, model structure flaws and basic problems associated with equifinality cannot be resolved with more data collection.

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?

Response: We have explained in the first comment's response.

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?

Response: Parameter sensitivity is defined as the change in outcome (e.g., a given streamflow metric) in response to a specified change in a parameter value. If there is a large deviation between the pre and post parameter change curves, we consider the parameter to be sensitive.

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

Response: We wanted to describe an approach to include or exclude a parameter as a candidate without regard to sensitivity. This classification is qualitative and hence we are not assigning a prior pdf. Therefore, we are hesitant to define it as a prior.

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.

Response: We understand and agree that there are other ways to approach calibration, some more efficient than others. The approach described in this paper does not require us to define a pdf. The key component of this paper is an analysis in both the time and frequency domain with pseudo accounting of shape of the hydrograph. Though MLE is a powerful tool that has wide application in parameter estimation, it is most useful to assess magnitude errors and cannot evaluate hydrograph shape mismatches. We also think this approach is a qualitative alternative to defining distributions, which has its own challenges. Further defining appropriate likelihood function itself has its own set of challenges (Mantovan and Todini, 2006). So while we agree that there are many other approaches that that can improve model calibration and evaluation, we believe we have provided a useful set of tools and examples that will help move the community forward. We very much appreciate the recommendation of other approaches and will look forward to considering these in future work.

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.

Response: We agree with the reviewer that large posterior uncertainty will influence the forecasts and should warrant further investigation. As this is not the primary scope of this paper, we have not addressed this topic in detail.

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.

Response: We thank the reviewer for bringing up an interesting point. One problem with choosing calibration parameters based on sensitivity is that it may lead to substantial distortion of the system. But the correlation issue is related. We have added the following content to highlight dependency between model inputs in mathematical models in a sensitivity analysis with the following content "most sensitivity analysis assumes that input parameters are independent and are not correlated, however that assumption may not always be appropriate (Song et al., 2015)." With regards to this manuscript, we wanted to describe an approach as to how a modeler can consider if a parameter is a candidate for calibration or not. It is a procedure, where the modeler can sequentially add parameters that are included for calibration if calibration is not achieved using the previous set. This is in contrast to how modelers commonly select highly sensitive parameters as candidates for calibration which is informed from a sensitivity analysis (Doherty and Skahill, 2006). Here we define the most influential parameters in terms of model outcome such as streamflow as sensitive parameters. In that context, we wanted to highlight that sensitivity alone should not be the reason for including a parameter as a candidate for calibration. For example, in the context of SWAT, curve number is commonly used as a candidate for calibration (Arnold et al., 2012) as its effect can be metaphorically compared to the sledge hammer. Meaning it is a very sensitive parameter, but might have not have a physical basis for its inclusion as a calibration candidate.

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

Response: We wanted to illustrate that by changing the values of the parameters and contrasting with other variables we can physically explain range-based sensitivity. Here, we show that the TIMP parameter is sensitive only when temperatures are below zero. For example, if the study area does not experience subzero temperatures, we don't need to include this parameter for calibration. In essence, we wanted to show that process based thinking still can be quite powerful and should be employed as one of many tools to limit physical process distortion represented by the model structure.

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?

Response: Calibration can be defined as an ill-posed, non-linear inverse problem that can lead to non-unique solutions. The outcome to parameter mapping is essentially non-unique. The constrained minimization problem Tikhonov regularization although quite attractive to solve inverse problems can have issues with numerical stability (Doherty and Skahill, 2006). With that said, similar such approaches with modifications could definitely be used in a calibration application. Here we present a rationale to determine parameters that can serve as suitable candidates for calibration based on how

much information we have about a particular parameter. Our reasoning for why only to include subset instead of all 1000+ parameters is based on whether the parameters were measured, derived or solely derived through a calibration exercise.

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.

Response: We agree with the reviewer that more and diverse types of observations is the primary and key weapon to constrain or reduce equifinality. However, in most cases data is still limited to streamflow and available only at few gages with short time periods. Therefore, we need other ways to address the challenge. Here we describe an approach where we are saying by not including certain parameters from calibration, the number of combinations we have is reduced (i.e., constraining equifinality). The reason why we are excluding certain parameters is because some parameters have scientific basis in their initialized values. For example, curve number is commonly adjusted to achieve calibration. Its specification has scientific basis. When papers are adjusting $\pm$10% (e.g., (Williams et al., 2012), it does not have that reasoning other than we need a parameter that can result in a calibrated model. With that said, if there are specific justifiable reasons available for adjusting curve number one should do it.

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".

Response: Thank you and we agree. We have included the following content in the

manuscript "Models are developed for a variety of applications and the metrics used to evaluate their performance should target the application of interest." We have included this content in the abstract, introduction, and conclusion to reiterate the importance of the point.

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?

Response: In that sentence, summer base flows was listed as an example of the flows that a modeler could be interested. For our scenario applications we were interested in several components of match between model simulated flow and observed flow and includes: pseudo shape and timing of flow evaluated in the frequency domain using wavelet approach, general model performance without regard for specific events using magnitude squared coherence approach in the frequency domain, seasonal flows using traditional metrics such as NSE visualized as box plot, magnitude differences identified by time for all flows in the time series and flow duration curves and empirical Q-Q plot as other means of showing the general performance of all model simulated flows. Even in the rare case that one wants to calibrate a model exclusively for summer base flows we would recommend a specific metric focusing on summer base flows in addition to several other metrics to help ensure that other components of the hydrologic cycle, which may affect summer base flows in hypothetical modeled scenarios that don't exactly follow patterns of the calibration and validation dataset, are properly represented.

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

References Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White,

M. J., Srinivasan, R., Santhi, C., Harmel, R. D., Griensven, A. v., Liew, M. W. V., Kannan, N., and Jha, M. K.: SWAT: Model Use, Calibration, and Validation, 55, 10.13031/2013.42256, 2012. Beven, K., and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrological Processes, 6, 279-298, 10.1002/hyp.3360060305, 1992. Doherty, J., and Skahill, B. E.: An advanced regularization methodology for use in watershed model calibration, Journal of Hydrology, 327, 564-577, https://doi.org/10.1016/j.jhydrol.2005.11.058, 2006. Engel, B. A., Srinivasan, R., Arnold, J., Rewerts, C., and Brown, S. J.: Nonpoint Source (NPS) Pollution Modeling Using Models Integrated with Geographic Information Systems (GIS), Water Science and Technology, 28, 685-690, 1993. Mantovan, P., and Todini, E.: Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, Journal of Hydrology, 330, 368-381, https://doi.org/10.1016/j.jhydrol.2006.04.046, 2006. Montanari, A.: Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, Water Resources Research, 41, n/a-n/a, 10.1029/2004WR003826, 2005. Neitsch, S. L., Arnold, J. G., Kiniry, J. R., and Williams, J. R.: Soil and water assessment tool theoretical documentation version 2009, Texas Water Resources Institute, 2011. Santhi, C., Arnold, J. G., Williams, J. R., Dugas, W. A., Srinivasan, R., and Hauck, L. M.: Validation of the SWAT model on a large rier basin with point and nonpoint sources., JAWRA Journal of the American Water Resources Association, 37, 1169-1188, 10.1111/j.1752-1688.2001.tb03630.x, 2001. Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M., and Xu, C.: Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications, Journal of Hydrology, 523, 739-757, https://doi.org/10.1016/j.jhydrol.2015.02.013, 2015. Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C. Y.: Calibration of hydrological models using flow-duration curves, Hydrol. Earth Syst. Sci., 15, 2205-2227, 10.5194/hess-15-2205-2011, 2011. Williams, J. R., Kannan, N., Wang, X., Santhi, C., and Arnold, J. G.: Evolution of the SCS Runoff Curve Number Method and Its Application to Continuous Runoff Simulation, Journal of Hydrologic Engineering, 17,

1221-1229, doi:10.1061/(ASCE)HE.1943-5584.0000529, 2012.

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