Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2017-121-AC2, 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

GENERAL COMMENTS 1. The authors aim providing a model evaluation framework for complex hydrological schemes, comprising a large number of unknown parameters. Taking as example the SWAT model, applied to two river basins in southern Minnesota, USA, they highlight several interesting issues, involving the choice, estimation and classification of parameters and the use of appropriate evaluation metrics, which are contrasted to single-objective calibration practices. In this context, they propose a three-stage procedure for parameter classification, and a more comprehensive model assessment framework, implemented within the so-called HydroME Toolbox. The latter comprises several graphical tools for the visual comparison of simulated vs. observed outputs both in time and frequency domain, which allow the user better evaluating the outcomes of calibrations.

We thank the reviewer for a comprehensive summary of the manuscript.

2. Although the article contains quite interesting ideas, key assumptions of the underlying rationale are somehow trivial. For instance, several times the authors dispute the use of NSE and R2 as lumped distance metrics for model evaluation, thus promoting more integrated approaches for taking into account further aspects of the simulated hydrographs. However, most of these arguments are not new. In contrast, the recent (and not so recent) hydrological literature has widely discussed the shortcomings of the well-known NSE function, and the current approaches in hydrological calibration have been already incorporated most of the issues that the authors reveal (among many others, cf. Gupta et al., 2009; Ritter and Munoz-Carpena, 2013).

Response: Throughout the paper we have explained that lumped metrics such as NSE and R2 are useful, but insufficient metrics. We do not dispute the use of these metrics, or discourage others from using them in the future. However, we believe that in addition to the common lumped metrics, more targeted evaluation metrics, such as those we have presented and built HydroME to evaluate, will provide much more targeted and comprehensive model evaluation. We agree with the reviewer that some of the issues with NSE, R2 and other lumped metrics have been well documented in the literature and we cite many such papers. However, despite the vast literature discussing the limitations of these metrics, NSE and R2 still serve as the most common measures of model evaluation, with very few studies using alternatives. Therefore, we believe there remains work to be done not only in discussing and demonstrating the limitations of these metrics, but also in providing alternatives and toolboxes to allow users to readily quantify such alternate metrics, as we have done.

3. The objectives of the article are not very clear, maybe because its title and abstract are little informative. In particular, the article refers to "multiple domain evaluation" and the abstract highlights the development of HydroME Toolbox. However, significant part of the text is dedicated not on model evaluation but on SWAT parameterization issues. Yet, due to the particular emphasis given to SWAT, the article loses its generality and

may be difficult to be followed by readers that are not very much familiar with SWAT. At least, the authors should explain some essential concepts, as explained in specific comments.

Response: For the revised version, we have made changes to the manuscript to address the reviewer suggestion. In particular, we have added more information to the abstract that reflects calibration parameter candidate selection. We have also rearranged some content to keep the material consistent within the manuscript. In addition, we have modified the title to "Calibration parameter selection and watershed hydrology model evaluation in time and frequency domains." The new title better reflects the content as suggested by the reviewer and removes any ambiguity. We understand the reviewer's concern regarding the focus on SWAT and we considered this issue extensively during the development of the manuscript. In the submitted version we included generalizable points in the introduction and discussion, which apply to virtually any watershed hydrology model. However, we are cognizant of the fact that SWAT is the dominant model used for watershed hydrology modeling, with an enormous user base and frequent model updates. For this reason, we believe our manuscript will have the greatest impact if we keep the focus and details somewhat specific to SWAT, while highlighting general insights where possible. We expect that we can highlight these points better in the revised version and will keep the reviewer's suggestion in mind.

4. Both case studies involve multisite runoff data (i.e. observed time series of river flows at many stations across the two basins). However, it is not clear how did the authors handled this data within calibrations. In the revised article, this issue requires further development, also revealing the value of multiobjective calibration approaches.

Response: This information was initially minimized in an effort to keep the article as concise as possible. However, we agree with the reviewer that it should be covered in more detail. We have added a section with more details about the calibration step and how we leveraged the rich multisite data available in these watersheds. We have explained the process in the Study area section. Briefly, we performed an upstream

to downstream calibration without changing the upstream parameters once they were deemed calibrated during the downstream subbasin calibration. We primarily adjusted the pure calibration parameters and when we did not obtain a satisfactory calibration, we included derived parameters. In addition, we have added information specifically on the usefulness of multi-objective calibration approaches using these two case studies.

5. SWAT is useful for representing the heterogeneity of catchment processes and properties. In your case studies, it is not clear whether the parameters to calibrate, shown in Figure 6, are considered lumped, thus having the same value across all HRUs, or distributed. In particular, the curve number parameter, CN2, is expressed in percentage terms – does this mean that CN2 is allowed to deviate around spatially varying reference values across HRUs?

Response: We have included the following content in the manuscript "Parameters were described in a distributed setting. CN2 is specified for each HRU based on land use, soils type and condition." All the parameters are distributed at the spatial resolution of the HRU and were varied as a percentage of their initial value.

6. To my point-of-view, this article is useful for "promoting" hybrid calibration strategies, as the most appropriate means to handle complex and highly-parameterized models (cf. discussion by Nalbantis et al., 2011). On the other hand, I found rather trivial the author's claims about the classical metrics used so far, and specifically their continuous reference to NSE and R2 limitations. In this context, I propose generalizing the target of the article (and maybe change the title) to include the proposed parameterization practices, and also paying more attention on the treatment of parameter uncertainties, which remains key challenge in hydrology.

Response: Based on reviewer recommendation and in this comment, we have changed the title to better reflect the content in the manuscript. However, the reason the authors have stressed on classical metrics is because all most all if not all of the publications involving SWAT, which is over 428 articles in 2016, just last year have

used classical metrics as measure for evaluating model performance. We believe it is sufficiently important to reiterate some of these limitations, to re-emphasize key points to readers and to set up the solutions and alternatives we provide. At the same time, we take the reviewer's point seriously and will attempt to make these points in a more concise manner in the revised manuscript.

SPECIFIC COMMENTS Page 1, lines 23-24: In contrast to SWAT, which is widely-known among hydrologists, I do not think that WEPP and GSSHA can also be characterized "leading" models.

Response: This varies regionally and is a minor point. Nevertheless, we deleted "leading" and replaced it with "Watershed hydrology"

Page 2, lines 6-7: Uniform Monte Carlo and Latin Hypercube are elementary random search schemes that are apparently not efficient for hydrological calibration purposes, particularly for complex model with many parameters. Why you omit referring to evolutionary optimization schemes and the numerous heuristics that have been employed in the context of parameter estimation procedures?

Response: We will include content to describe other approaches that are used for parameter estimation procedures.

Page 2, line 12: Manual calibration of complex models is not simply "somewhat" inefficient and time-consuming. Please, refer to the recent (e.g., Fatichi et al., 2016) as well as not so recent (e.g., Eckhardt and Arnold, 2001) literature, where you may find several interesting articles dealing with calibration approaches for distributed hydrological models.

Response: We don't question the findings reported in Figure 1 of Fatichi et al. (2016), however, in the Le Sueur River Basin (LSRB) lack of input data and not model structure necessitates calibration from a hydrology stand point. In LSRB, the landscape is extensively tiled to remove excess soil moisture to cultivate crops. The presence and

distribution of tiles are unknown. Therefore, even though SWAT contains the model structure in the form of Hooghoudt and Kirkham tile drain equations to model subsurface tile drainage (Moriasi et al., 2013), the only means of matching predicted flows with observed flows is through calibration. With that said, SWAT also does not explicitly model Karst systems, which is made rather more difficult with unknowns in terms of ground and surface water connectivity that exist in Root River Basin (RRB). To that end, calibration was necessary in these landscapes. Additionally, absence of model structure that accounts for Karst means we have to use hydraulic conductivity and other pseudo representations to characterize Karst behavior. In a general sense we agree with this comment, but when we do not have sufficient information or when model structure does not explicitly account for the process we had to resort to adjusting parameters that were not explicitly introduced to represent a process. This is not an uncommon situation and we believe needs to be discussed.

Page 3, lines 1-2: "Lumped metrics such as Nash Suttclife Efficiency (NSE) and coefficient of determination (R2) have been established as key model performance benchmarks." Apparently, I agree for NSE, but not for R2, i.e. the square of the Pearson's correlation coefficient. To my knowledge, R2 is not often used as performance metric, since this choice implies that the relationship between the examined variables is expected to be linear, which is obviously not the case in hydrological modeling. Anyway, the limitations of both NSE and R2 have been widely discussed in the literature, and the authors have to significantly enhance the associated references and better explain their novelties.

Response: Figure 1 in the Fatichi et al. (2016) does use R2 as do many other papers (e.g., (Abbaspour et al., 2007)). Perhaps there are some geographical differences as to which metrics are most commonly used around the world, but that issue falls outside the scope of our paper. We believe R2 is sufficiently common to mention in our paper and its inclusion certainly does not detract from the paper in any way.

Page 4, line 7: The delineation to 1,823 HRUs (and 17,174 for Root River basin)

through ArcSWAT implies that a tremendous number of properties have to be identified. It is strongly recommended to make a comment on how did you take advantage from this spatial information within the calibration problem.

Response: The following information has been added to the supplemental information along with a reference to it in the manuscript. “There are 19 parameters contained in the HRU file that directly defines HRU properties and also control streamflow. The four parameters representing topographic characteristics were derived using a 10m DEM. We did not calibrate these parameters as these were classified as derived parameters. There are thirteen parameters that represent land cover characteristics of which some were extracted from land use data, some others specified from literature and the rest were treated as pure calibration parameters when either measured data was not available or in case of conceptual parameters, where it can only be estimated through calibration. For example, OV_N defined as the Manning’s n for overland flow was specified a range and then varied within that range. In other words, it was treated as derived calibration parameter. DEP_IMP, the depth to impervious layer was treated as a pure calibration parameter. The ESCO and EPCO defined in Table 2 in SI were treated at the basin scale. The soil parameters were extracted from detailed soils information from the SSURGO database. The landscape is predominantly agricultural and land use and crop management properties were obtained from government documents of fertilizer use such as timing, type and rate.”

Page 3, line 20: What do you mean by “modern water quality problems”?

Response: Recent changes in flow regimes have contributed to increased turbidity in several Minnesota Rivers that are attributed to modern land management practices. For example, extensive tile drainage utilization in Minnesota have contributed to highly erosive rivers (Belmont et al., 2011;Schottler et al., 2014). We have changed “modern water quality problems” to “anthropogenically introduced water quality problems” to better convey the idea.

Page 4, line 12: Please, provide the mathematical expressions of the three metrics or at least a description of them (particularly, what do you mean by PBIAS?).

Response: We have included explanation and mathematical expressions for the three metrics in the SI and referenced them in the manuscript. We don’t believe it is useful to include such trivial information in the manuscript itself.

Page 6, line 16: NSE values have been estimated by contrasting simulated vs. observed flows at the outlet station? How do these values change across the rest of basin’s stations?

Response: We provide NSE, R2 and PBIAS values for all of the gages in Table 1 for the calibrated models. If we repeat our exercise of 1000 runs varying flow routing parameters we would obtain results similar to those shown in Figure 2. However, we believe the point is sufficiently made using the outlet gage results.

Page 6, line 33, to page 7, line 2: “The same three plots were generated with R2 and PBIAS as the performance metric and are shown in the SI (Fig. S3 and Fig. S4). These plots indicate that multiobjective criteria may not fully resolve the challenges with equifinality.” In order to take full advantage of multiobjective calibration benefits, the individual performance criteria have to be mutually uncorrelated (e.g. Efstratiadis and Koutsoyiannis, 2010). This important requirement is not generally fulfilled when using the above statistical expressions, which are generally based on least square error hypothesis.

Response: We thank the reviewer for the comment. We have mentioned this caveat and cited the paper to clarify this point.

Page 7, lines 13-25: CN is by definition associated with maximum soil retention, S (in mm), though the well-known formula S = 254 (100/CN – 1). From the above expression we easily recognize that the sensitivity of S is maximized for extreme CN values, while it becomes is less sensitive for medium values. Therefore, a specific percentage change

of the reference CN does not imply that the soil retention will also change by the same ratio. For instance, a 10% increase of CN from 50 to 55 will result to a decrease of S by 18%, while a 10% increase of CN from 80 to 88 will result to a decrease of S by 45%.

Response: We thank the reviewer for the comment. We have added the mentioned point to highlight range-dependent sensitivity.

Page 7, section 3.4: The tile "Choosing parameters through physics" does not correspond to the procedures described herein.

Response: Thank you for the suggestion. This section discusses how parameters representing physical processes affect streamflow in the time and frequency domains. But to clarify, we have altered the title of this section 'Choosing parameters based on physical processes'.

Page 8, lines 6-7: "Several SWAT parameters are specified at the spatial resolution of the basin, e.g., TIMP and SMTMP". Please, provide a brief explanation of the second parameter, as done for TIMP. Page 9, section 3.5: Parameter classification, which is, to my opinion, the most important step (and very well-written in the text), shouldn't be implemented before the aforementioned analyses in time and frequency domain?

Response: We have added information in the manuscript explaining SMTMP in detail.

Page 10, lines 13-14: [Derived parameters] ". . . theoretically contain not only the uncertainties and errors associated with the measured values on which they are based, but also uncertainties and errors due to their specification." This sentence requires clarification. I suppose that you mean that these parameters contain two types of uncertainties, with respect to their literature definition (based on experimental data) as well as the selection made by the user for the specific problem.

Response: Yes. We have made the change to clarify the sentence. "These parameters theoretically contain three types of uncertainties, (i) with respect to their literature definition (based on experimental data), (ii) the uncertainties of the measured data used

to specify, as well as (iii) the selection made by the user for the specific problem, which can depend on user expertise".

Page 10, lines 24-25: "Examples of these parameters include soil hydraulic conductivity and soil texture." I do not agree that these parameters correspond to real physical quantities, since they do not refer to the point scale but to the computational element (e.g., grid cell) scale, which is much larger (cf. Nalbantis et al., 2011). Additionally, it is not possible to have measured information about these properties across the entire watershed. My opinion is that such parameters should be better characterized as physically-based rather than measured.

Response: We appreciate the suggestion, but prefer to stick with 'measured.' Referring to third group as physically-based could be misinterpreted that other parameters are not, which is not quite correct. In addition, precipitation, temperature and numerous other quantities are measured and assumed constant for practical purposes for spatial extents over which the model is lumped. In that context, both soil hydraulic conductivity and soil texture are measured quantities. The SSURGO database contains this information for the entire watershed and most of United States and similar databases exist elsewhere.

Page 12, lines 15-17: "Nevertheless, given the current level of sophistication of hydrologic modeling and increasing demand for models that target specific hydrologic metrics (e.g., summer base flows, timing and rate of the rising limb of the snowmelt hydrograph), we propose that the community is poised to move beyond these basic metrics." This recommendation, although correct, is already adapted in state-of-the-art hydrological practices.

Response: All most all if not all papers report R2 or NSE or similar such lumped metric as a metric for evaluating model performance (e.g., (Fatichi et al., 2016;Eckhardt and Arnold, 2001;Gupta et al., 2009)). In most cases they are the only metrics reported and they are reported for just the calibration and validation periods for a particular gage.

With that said, we agree with the author that the idea can be presented in different form by capturing the reviewer's point of view.

Page 12, lines 23-24: "Fig. 7 shows that model performance varies considerably during the fall and winter months and is captured by both metrics" The graphs indicate that the obtained values vary within an extremely large range, while the average model efficiency in fall and winter is very bad. Are these results extracted from calibrations or simply through Monte Carlo sampling through the feasible parameter space? How do you explain this performance?

Response: They are from the calibrated model. We wanted to illustrate that when we use the entire flow record for the time period when the model was calibrated, the NSE and R2 values are 0.69 and 0.70 respectively. However, when take this same data and segregate them into seasons we see a different picture. So, if a modeler is interested in a particular season, it may be critical to verify model performance for the season of interest or in a general sense, the period where they are interested. So, they are not from just sampling through the feasible parameter space. When we lump large time periods, the NSE is biased because of very large flows. However, when we separate them, we see the differences between smaller flows clearly.

Page 13, lines 14-15: "The uncalibrated model deviates from measured flows across the full range of flows" Apparently, a calibrated model is substantially superior to an uncalibrated one. Could one expect something different?

Response: The authors have fixed the figure identified as 7 in the manuscript (which should be 8). We have made the correction. Now the statement is true.

Page 14, section 5.2: I do not understand the interpretation of 32-day and 128-day bands. Which is the physical explanation of these frequencies?

Response: 32 day or 128 day corresponds to period which is the inverse of frequency. Wavelet analysis is a multiresolution signal analysis. The vertical axis indicates period

in days and represents the width of the dilated basis function used to evaluate the signal. The procedure for obtaining the coherence plot entails generating a continuous wavelet transform (cwt) of the individual signals which are then compared. The cwt can decompose the signal to show if a particular frequency is present for a specific day. If a particular frequency is present in the signal for the evaluating window at a particular time (horizontal axis), then we will obtain a higher coefficient. Then we compare the two cwt's to obtain a coherence value, which reveals if the two cwt's are similar. So, 32 day or 128 day in a coherence plot means if those frequencies were present in both the signals. Now with that background, frequency is how often some event is repeating. So, by comparing to a known function with known frequency (here complex Morlet), we can say if a particular event is repeating. The use of the term band is to refer to a range of frequencies. We need infinite number of scaling and translations to compute the wavelet transform, which is a problem. Hence, wavelets are applied as a band pass filter (Valens, 1999). Hence, the use of the term 'band' as we are not evaluating the transform at all scalings. MINOR EDITORIAL COMMENTS Page 1, line 5: Please, change "simulating hydrology and water quality" by "simulating hydrological and water quality processes" (hydrology refers to the discipline, not the water cycle processes). Response: We thank the reviewer for catching the error. We have made the change and replace it with reviewer suggestion.

Page 3, line 1: Change to read "Sutcliffe". Response: We thank the reviewer for catching the error. We have fixed it.

Page 5, line 4: Remove one of the two parentheses after the year of publication. Response: The second parenthesis is because the reference is used as an example.

Page 7, lines 27 and 33: Change to read "physically-based" and "threshold-based", respectively. Response: Done.

Page 11, lines 15-16: Please, leave a space between citations and remove parenthesis before "Liu et al.". Response: The second parenthesis is because the references cited

are used as an examples. We thank the reviewer for catching the error. We have added space between citations.

REFERENCES Eckhardt, K. and Arnold, J. G.: Automatic calibration of a distributed catchment model, J. Hydrol., 251(1–2), 103–109, doi:10.1016/S0022-1694(01)00429-2, 2001.

Efstratiadis, A., and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: A review, Hydrol. Sci. J., 55, 58-78, doi:10.1080/02626660903526292, 2010.

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., Downer, C. W., Camporese, M., Davison, J. H., Ebel, B., Jones, N., Kim, J., Mascaro, G., Niswonger, R., Restrepo, P., Rigon, R., Shen, C., Sulis, M. and Tarboton, D.: An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, J. Hydrol., 537, 45–60, doi:10.1016/j.jhydrol.2016.03.026, 2016.

Gupta, H.V., Kling, H., Yilmaz, K.K., and Martinez, G.F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. of Hydrol., 377(1-2), 80-91, 2009, 2009.

Nalbantis, I., Efstratiadis, A., Rozos, E., Kopsiafti, M., and Koutsoyiannis, D.: Holistic versus monomeric strategies for hydrological modelling of human-modified hydrosystems, Hydrol. Earth Syst. Sci., 15, 743-758, doi:10.5194/hess-15-743-2011, 2011.

Ritter, A., and Munoz-Carpena, R.: Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments, J. Hydrol., 480, 33-45, 2013.

References Abbaspour, K. C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J., and Srinivasan, R.: Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT, Journal of Hydrology, 333, 413-430, https://doi.org/10.1016/j.jhydrol.2006.09.014, 2007. Belmont, P., Gran, K. B.,

Schottler, S. P., Wilcock, P. R., Day, S. S., Jennings, C., Lauer, J. W., Viparelli, E., Willenbring, J. K., Engstrom, D. R., and Parker, G.: Large Shift in Source of Fine Sediment in the Upper Mississippi River, Environmental Science & Technology, 45, 8804-8810, 10.1021/es2019109, 2011. Eckhardt, K., and Arnold, J. G.: Automatic calibration of a distributed catchment model, Journal of Hydrology, 251, 103-109, https://doi.org/10.1016/S0022-1694(01)00429-2, 2001. Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., Downer, C. W., Camporese, M., Davison, J. H., Ebel, B., Jones, N., Kim, J., Mascaro, G., Niswonger, R., Restrepo, P., Rigon, R., Shen, C., Sulis, M., and Tarboton, D.: An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, Journal of Hydrology, 537, 45-60, https://doi.org/10.1016/j.jhydrol.2016.03.026, 2016. Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80-91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009. Moriasi, D. N., Gowda, P. H., Arnold, J. G., Mulla, D. J., Ale, S., Steiner, J. L., and Tomer, M. D.: Evaluation of the Hooghoudt and Kirkham Tile Drain Equations in the Soil and Water Assessment Tool to Simulate Tile Flow and Nitrate-Nitrogen, Journal of Environmental Quality, 42, 1699-1710, 10.2134/jeq2013.01.0018, 2013. Schottler, S. P., Ulrich, J., Belmont, P., Moore, R., Lauer, J. W., Engstrom, D. R., and Almendinger, J. E.: Twentieth century agricultural drainage creates more erosive rivers, Hydrological Processes, 28, 1951-1961, 10.1002/hyp.9738, 2014. Valens, C.: A really friendly guide to wavelets, 1999.

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