Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2017-121-RC2, 2017

Hydrology and Earth System Sciences
Discussions

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

A. Efstratiadis (Referee)

andreas@itia.ntua.gr

Received and published: 30 April 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.

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).

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.

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.

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?

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.

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.

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?

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.

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.

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.

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

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?).

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?

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.

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%.

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

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?

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.

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.

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.

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?

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?

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?

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).

Page 3, line 1: Change to read "Sutcliffe".

Page 5, line 4: Remove one of the two parentheses after the year of publication.

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

Page 11, lines 15-16: Please, leave a space between citations and remove parenthesis before "Liu et al.".

Page 14, line 26: Change to read “Discussion”.

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.

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