Discussion

The fallacy in the use of the "best-fit" solution in hydrologic modeling

K. C. Abbaspour

Texas A&M University, Department of Biological and Agricultural Engineering, College Station, USA.

Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600, Dübendorf, Switzerland.

Corresponding author

K.C. Abbaspour

Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600, Dübendorf, Switzerland.

abbaspour@eawag.ch

Abstract

Using the parameters associated with the *best-fit* simulation (i.e., the simulation with the highest objective function value) to represent a calibrated hydrological model is inadequate. The reason is that the calibrated model's *best objective function value* is usually not significantly different from the *next best value* or the *values after that*. This non-uniqueness of the objective function values causes a problem because the *best solution's parameters* are often significantly different from the *next best set of parameters*. Therefore, *only* using the *best simulation parameters* as the calibrated model's sole parameters to interpret the watershed processes or perform further modeling analyses could produce misleading results. Furthermore, the lack of pristine watersheds makes the task of watershed-scale calibration increasingly challenging.

Subjective thresholds of *acceptable performance criteria* suggested by some researchers, based on comparing the measured and the best solution signals, are often not achievable. Hence, to obtain a *satisfactory fit*, researchers and practitioners are often forced to compromise the science behind their work. This article discusses the fallacy in using the *best-fit* solution in hydrologic modeling. A two-factor statistic to assess the goodness of calibration/validation is discussed, considering model output uncertainty.

Keywords: Model Uncertainty, Calibration, Performance criteria, Stochastic calibration, Deterministic Calibration

Distributed watershed models are input-intensive, requiring inherently uncertain data. These data include soil and landuse maps and databases, climate data, water use, watershed management data, and at the minimum river discharge data for model calibration. These data are affected by almost all watershed activities such as agricultural activity, point sources, dam operation, river controls, various constructions, and water transfers. Given the highly uncertain input data, a watershed model's calibration must be *stochastic* and consider uncertainties. However, *deterministic* approaches, which use a single set of model parameters associated with the best fit, are still widely used. These calibrated models are assessed by performance measures such as Nash-Sutcliffe (NSE), R^2, and PBIAS, which compare only two signals.

In contrast, stochastic solutions are generated by parameters treated as random variables and deemed acceptable if they fall within a behavioral threshold and have statistically similar objective function values. In other words, if there is one good solution, then there are many, and they all constitute the solution to the calibration problem.

The problem with the deterministic solution is not with the best fit but rather with taking the best fit's parameter set as the actual parameters of the watershed model and using it in the interpretation of the watershed hydrology. Subjective Criteria rating the goodness of calibration or validation often include statements such as: (Very good: 0.75 < NSE < 1.00), (good: 0.65 < NSE < 0.75), (satisfactory: 0.5 < NSE < 0.65), or (Unsatisfactory: NSE < 0.50) (e.g., Moriasi et al., 2007). These criteria could be quite misleading if used in a deterministic way - that is, to search for one solution within the above ranges to regard a model as satisfactory or good. Instead, *all solutions* within these ranges should be sought and the associated parameters used for further analysis to quantify the uncertainties in the modeling works.

The following points summarize the pitfalls of model performance criteria to rate the performance and uncertainty in a calibrated model. A SWAT (Soil and Water Assessment Tool) (Arnold et al., 2012) model example from a watershed in the Danube basin is used for illustration.

First, model performance criteria only compare two signals, mainly observed versus the best-fit simulation (Fig. 1a). The implicit assumption here is that the best-fit solution (Table 1, first row) represents the calibrated watershed model. Parameters associated with this solution are then used in subsequent analyses, such as calculating water resources, crop yield, and climate change impacts. This assumption is not correct as many significantly different parameter sets can produce statistically similar model performance criteria used as an objective function (Table 1, all ten rows). Taking only one of them, albeit the best one, to represent the watershed could lead to entirely erroneous and misleading results. For example, calculating the watershed's blue water resources represented by the top ten parameter sets in Table 1 leads to significantly different numbers ranging from 543 to 1575 mm.

Second, model performance criteria, by their deterministic nature, ignore model uncertainty. Therefore, the deterministic subjective criteria cited above are not adequate for hydrologic models that consider model uncertainties.

Third, watersheds are being increasingly disturbed with dams, reservoirs, water transfers, and accelerated landuse changes; hence, matching the output of a deterministic model with observation is becoming difficult. It is, therefore, necessary to compare an observation signal with uncertain model outputs distributions.

A procedure to calibrate a model stochastically is summarized here and detailed in the references provided.

Initially, model simulation is compared with observation to decide if the model is adequate for calibrating. Not complying with the *correct neglect* principle (Abbaspour et al., 2018) could render calibration meaningless if essential processes are missing from the model. Next, physically meaningful ranges are assigned to parameters chosen for calibration based on the initial model result (Abbaspour et al., 2015). Following a calibration protocol outlined in the latter reference, it will take a few iterations of around 200-500 simulations to calibrate a model. The final parameters have smaller ranges centered on the best model performance. At each iteration, the 95% prediction uncertainty (95PPU) is calculated (Fig. 1b) to quantify the effect of parameter uncertainties on model outputs, such as river discharge. Two statistics, referred to as *P-factor* and *R-factor*, quantify the calibration performance or the goodness of fit after each iteration. *P-factor* represents model accuracy and ranges from 0 to 1. In other words, it is the percentage of measured data that falls inside the 95PPU band. By definition, (1 - *P-factor*) is the model error. *R-factor* is the average thickness of the 95PPU divided by the standard deviation of the measured data and depicts model uncertainty. It can range from 0 to a relatively large value. A value around 1 for the *R-factor* is equal to the standard deviation of the observation and is desirable. These two factors fully describe the performance of the calibrated model. The closer the *P-factor* is to 1 and the *R-factor* to 0, the better the calibrated model represents the measurements. Based on experience and only as a reference and not a criterion, we should bracket about 70% of the measured data in the 95PPU band (*P-factor* ≥0.7, R-factor ≤1.5) for river discharge. Due to more considerable uncertainties in measuring and modeling sediment and nitrate loads, the reference *P-factors* could be smaller (≥0.5 or 0.4) and the *R-factors* larger ≤2 to 3).

The example in Figure 1a shows a deterministic case with an NSE of 0.47, an unsatisfactory model based on the subjective thresholds mentioned above. While taking model uncertainties into account (Fig. 1b), the calibrated model is more acceptable with *P-factor* = 0.73 and *R-factor* = 1.1, assuming a 10% error in the flow measurement.

In the above example, the subjective criteria of satisfactory, good, very good, or unsatisfactory are meaningless if model uncertainty is not quantified. A model with a best-fit NSE of 0.8 but with considerable uncertainty in the prediction could be unsatisfactory. Facing the difficulty of satisfying the subjective criteria leaves researchers in a predicament. On the one hand, they need to maintain their work's scientific integrity by reporting the actual calibration results. On the other hand, they need to produce an *acceptable* calibration result to publish their work. Unfortunately, it is always the former that is sacrificed. Therefore, it is prudent to use schemes that compare a measured signal (or a distribution if considering measurement errors) with a model output distribution.

References

Abbaspour KC, et al. 2015. Modelling hydrology and water quality of the European Continent at a subbasin scale: calibration of a high-resolution large-scale SWAT model. Journal of Hydrology. 524:733-752. https://doi.org/10.1016/j.jhydrol.2015.03.027.

Abbaspour KC, et al. 2018. A guideline for successful calibration and uncertainty analysis for soil and water assessment: a review of papers from the 2016 international SWAT conference. Water. 10:6.

Arnold JG, et al. 2012. SWAT: Model use, calibration, and validation. Transactions of the ASABE. 55:1491-1508. https://digitalcommons.unl.edu/biosysengfacpub/406

Moriasi DN, et al. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE. 50:885-900. https://doi.org/10.13031/2013.23153

Figure 1. a) Deterministic model results comparing the best-fit signal with observed data. NSE=0.47. b) Stochastic model results comparing the 95% prediction uncertainty (95PPU) with observed data. *P-factor*=0.73, *R-factor*=1.1.

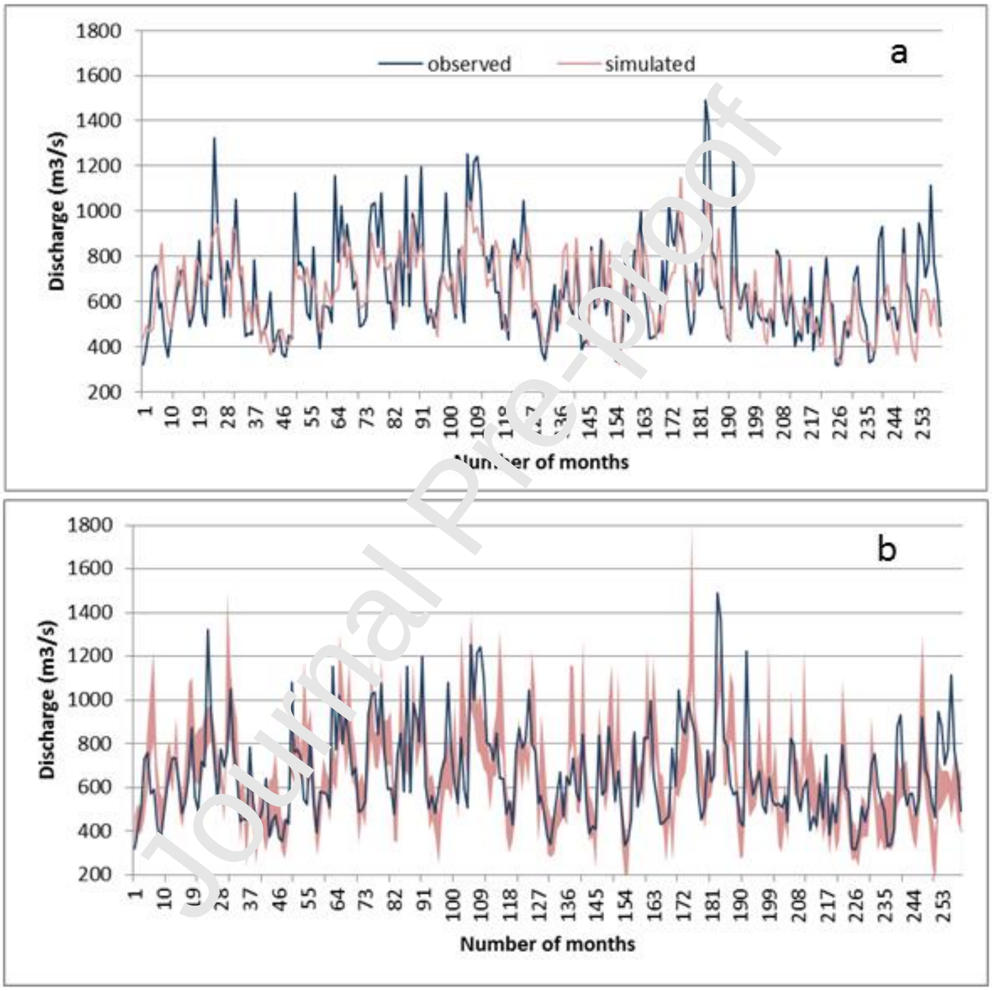
a
observed
simulated
Discharge (m3/s)
1800
1600
1400
1200
1000
800
600
400
200
1
10
19
28
37
46
55
64
73
82
91
100
109
118
127
136
145
154
163
172
181
190
199
208
217
226
235
244
253
Number of months
b
Discharge (m3/s)
Number of months

Table 1. Model parameters and their associated objective function values (NSE) showing similar objective functions obtained with significantly different parameters.

r__CN2	v__ESCO	v__GWQMN	v__GW_DELAY	r__SOL_K	r__SOL_BD	others	NSE
0.03	0.72	558	77	0.14	0.82	-	0.470
-0.08	0.85	779	53	-0.12	0.76	-	0.466
-0.07	0.87	544	61	0.32	0.69	-	0.460
0.13	0.80	333	64	-0.15	0.01	-	0.460
0.11	0.70	1250	74	0.05	0.55	-	0.460
-0.02	0.87	1232	41	0.00	0.05	-	0.445
-0.08	0.78	890	76	-0.42	0.31	-	0.445
0.22	0.72	1214	77	0.17	0.81	-	0.445
0.11	0.73	337	53	-0.5	0.53	-	0.445
0.28	0.71	811	49	0.09	0.39	-	0.445

r__ represents a relative change, v__ represents a value change (see Abbaspour et al., 2007 for details).

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.