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

Karim C. Abbaspour[1]

[1]Swiss Federal Institute of Environmental Science and Technology (EAWAG)

November 22, 2022

**Abstract**

The use of the parameters associated with the "best-fit" criterion to represent a calibrated hydrological model is inadequate. Furthermore, assessing the goodness of model calibration or validation based on performance criteria, such as NSE, $R^2$, or PBIAS, is misleading because they only compare two signals, i.e., measurement and the best-fit simulation (i.e., simulation with the best objective function value). The reason is that the calibrated model's best objective function value is usually not significantly different from the next best value or the value after that. This non-uniqueness of the objective function causes a problem because the best solution's parameters are always significantly different from the next best parameters. Therefore, only using the best simulation parameters as the calibrated model's sole parameters to interpret the watershed processes or perform further model analyses could lead to erroneous results. Furthermore, most watersheds are increasingly changing due to human activities. The lack of pristine watersheds makes the task of watershed-scale calibration increasingly challenging. Subjective thresholds of acceptable performance criteria suggested by some researchers to rate the goodness of calibration are based on the comparison of the two signals, and in most cases, the thresholds are not achievable. Hence, to obtain a satisfactory fit, researchers and practitioners are forced to massage and manipulate the input or simulated data, compromising the science behind their work. This article discusses the fallacy in using the "best-fit" solution in hydrologic modeling. It introduces a two-factor statistics to assess the goodness of calibration/validation while taking model output uncertainty into account.

# 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.

## Commentary

## Corresponding author

K.C. Abbaspour

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

abbaspour@eawag.ch

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


## Abstract

The use of the parameters associated with the "best-fit" criterion to represent a calibrated hydrological model is inadequate. Furthermore, assessing the goodness of model calibration or validation based on performance criteria, such as NSE, $R^2$, or PBIAS, is misleading because they only compare two signals, i.e., measurement and the best-fit simulation (i.e., simulation with the best objective function value). The reason is that the calibrated model's best objective function value is usually not significantly different from the next best value or the value after that. This non-uniqueness of the objective function causes a problem because the best solution's parameters are always significantly different from the next best parameters. Therefore, only using the best simulation parameters as the calibrated model's sole parameters to interpret the watershed processes or perform further model analyses could lead to erroneous results. Furthermore, most watersheds are increasingly changing due to human activities. The lack of pristine watersheds makes the task of watershed-scale calibration increasingly challenging. Subjective thresholds of acceptable performance criteria suggested by some researchers to rate the goodness of calibration are based on the comparison of the two signals, and in most cases, the thresholds are not achievable. Hence, to obtain a satisfactory fit, researchers and practitioners are forced to massage and manipulate the input or simulated data, compromising the science behind their work. This article discusses the fallacy in using the "best-fit" solution in hydrologic

modeling. It introduces a two-factor statistics to assess the goodness of calibration/validation while taking model output uncertainty into account.

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. Watershed data could include information about everything in a watershed affecting water regime and its quality; for example, agricultural activity, point sources, dam operation, river controls, road building, and water transfers. Given the highly uncertain input data, a watershed model's calibration must be stochastic. However, deterministic approaches, which use a single set of parameters associated with the best-fit, are widely used. In a stochastic solution, parameters are treated as random variables, with distributions representing all the solutions that fall within a behavioral threshold or within statistically similar objective function values.

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 that watershed and using it for subsequent analysis and 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 are misleading on many levels. A SWAT (Soil and Water Assessment Tool) (Arnold et al., 2012) model example from a watershed in the Danube basin is used to illustrate some points.

First, NSE or similar model performance criteria (MPC) only compare two signals, mainly observed versus the best-fit simulation (Fig. 1). 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 objective function values (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, MPCs, by their deterministic nature, ignore model uncertainty. Therefore, the deterministic subjective criteria cited above are not adequate for hydrologic models considering model uncertainties.

Third, as 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. Hence, it is necessary to compare an observation signal with uncertain model outputs.

Facing the difficulty of satisfying the subjective criteria for "very good," "good," or "satisfactory" calibration results 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. A procedure is summarized here and detailed in the references provided.

Calibration begins with a set of optimizing parameters chosen based on the initial model result before calibration. The parameters are initially quantified by uncertainty ranges (uniform distributions) based on prior experience and knowledge of the physical parameter values. Following a calibration protocol (Abbaspour et al., 2015), it will take a few iterations of around 500 simulations each for a model to be calibrated. The result is a smaller parameter ranges centered on the best model performance in each iteration. At each iteration, the 95% prediction uncertainty (95PPU) is calculated at the 2.5% and 97.5% levels of the cumulative distribution of output variables obtained through the Latin hypercube sampling scheme (Fig. 2). Two statistics, referred to as *P-factor* and *R-factor*, are used to quantify the calibration performance or the goodness of fit. *P-factor* represents model accuracy and ranges from 0 to 1. It is the percentage of measured points that fall inside the 95PPU band; in other words, these points are "correctly" simulated by the model. *R-factor* depicts model uncertainty and can range from 0 to a very large value. It is the average thickness of the 95PPU divided by the standard deviation of the measured data. A value of around 1 for the *R-factor* is in the range of standard observation deviation and is desirable. These two factors fully describe the strength of the calibrated model. The closer the *P-factor* is to 1, and the *R-factor* is to 0, the better the calibrated model represents the measurements. Based on experience and only as a reference, for river discharge, we should want to bracket about 70% of the measured data in the 95PPU band (*P-factor* >0.7, R-factor <1.5 ). Due to larger uncertainties in the measured data and modeling errors, for sediment load, we recommend *P-factor* >0.5, and for nitrate and phosphate loads, *P-factor* >0.4 with an *R-factor* around 1.5 to 2.5.

The example in Figure 1 shows a determinist case with an NSE of 0.47, an unsatisfactory model based on the subjective thresholds mentioned above. While taking model uncertainties into account, the calibrated model has acceptable results 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 for a calibrated model being good, very good, or unsatisfactory are irrelevant if model uncertainty is not quantified. A model with a best-fit NSE of 0.9 with considerable prediction uncertainty could also be unsatisfactory. Based on the existing evidence, it is time to abandon using the "best-fit" as a criterion for assessing model calibration results and adopt an uncertainty-based approach as described above.

## References

Abbaspour, K. C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R. & Klöve, B. (2015). Modelling hydrology and water quality of the European Continent at a subbasin scale: calibration of a high-resolution large-scale SWAT model. *Jornal of Hydrology*, *524*, 733-752.

Arnold J. G., Moriasi D. N., Gassman P. W. Abbaspour, K. C.  White, M. J., Srinivasan, R., et al. (2012). SWAT: Model use, calibration, and validation. *Transactions of the ASABE*, 55, 1491-1508.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of  the ASABE*, 50, 885-900.

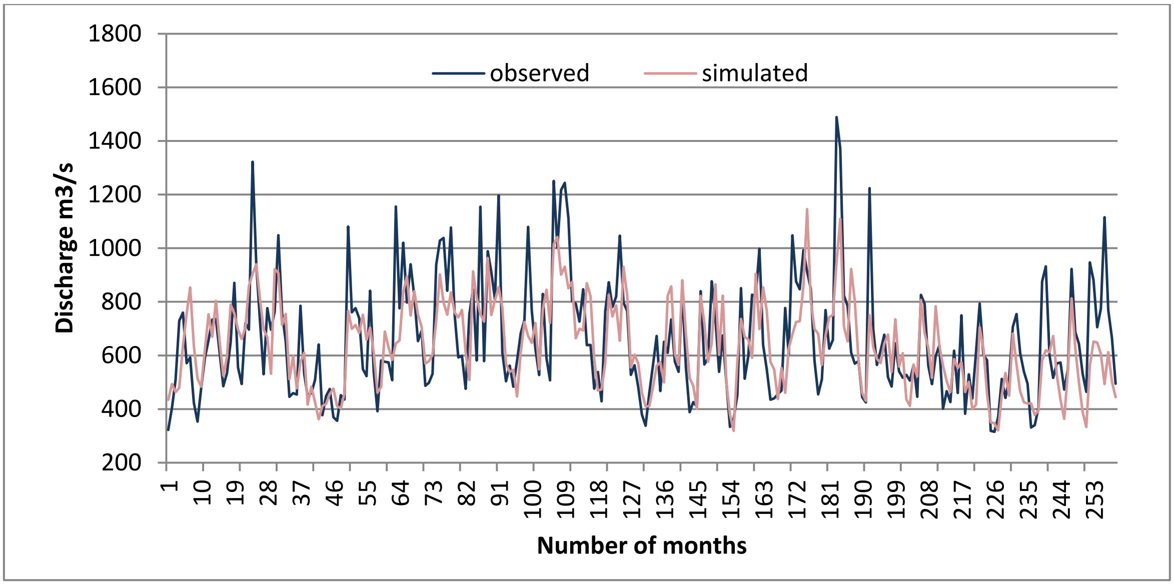


Figure 1. Deterministic model results comparing the best-fit signal with observed data. NSE=0.47.

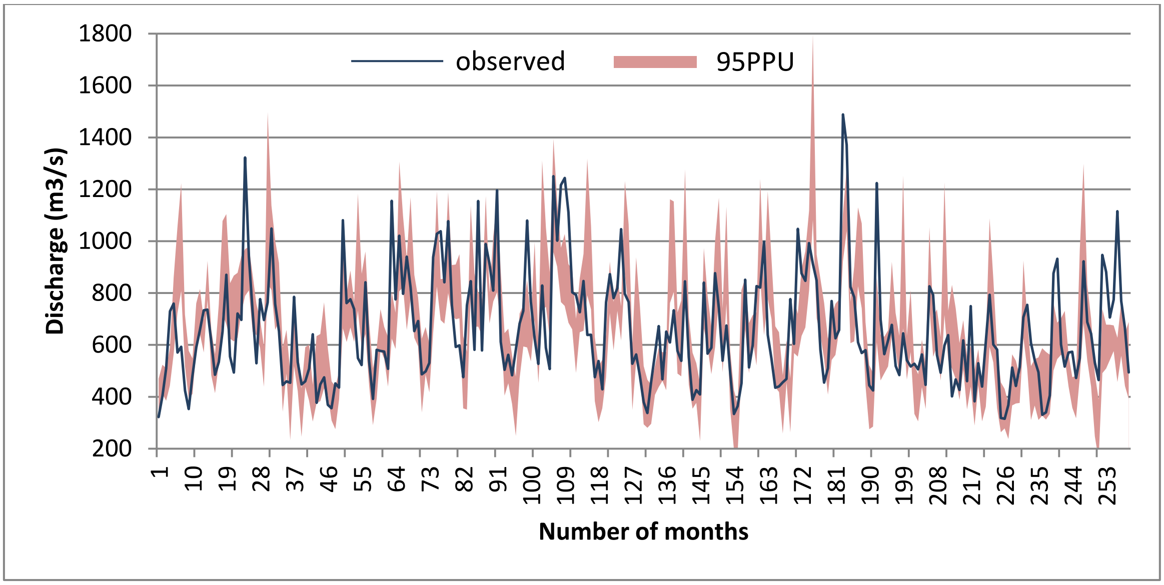


Figure 2. Stochastic model results comparing the 95% prediction uncertainty (95PPU) with observed data. *P-factor*=0.73, *R-factor*=1.1.

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 | 557.98 | 77.44 | 0.14 | 0.82 | . | 0.470 |
| -0.08 | 0.85 | 779.12 | 53.24 | -0.12 | 0.76 | . | 0.466 |
| -0.07 | 0.87 | 543.71 | 60.59 | 0.32 | 0.69 | . | 0.460 |
| 0.13 | 0.80 | 322.57 | 64.26 | -0.15 | 0.01 | . | 0.460 |
| 0.11 | 0.70 | 1249.94 | 73.77 | 0.05 | 0.55 | . | 0.460 |
| -0.02 | 0.87 | 1232.11 | 40.70 | 0.00 | 0.05 | . | 0.445 |
| -0.08 | 0.78 | 889.69 | 75.92 | -0.42 | 0.31 | . | 0.445 |
| 0.22 | 0.72 | 1214.27 | 77.31 | 0.17 | 0.81 | . | 0.445 |
| 0.11 | 0.73 | 336.84 | 52.36 | -0.50 | 0.53 | . | 0.445 |
| 0.28 | 0.71 | 811.22 | 48.81 | 0.09 | 0.39 | . | 0.445 |

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