All that glitters is not gold: the case of calibrating hydrological models

Vazken Andréassian, Nicolas Le Moine, Charles Perrin, Maria-Helena Ramos,
Ludovic Oudin, Thibault Mathevet, Julien Lerat, Lionel Berthet

All that glitters is not gold: the case of calibrating hydrological models

Vazken Andréassian[1], Nicolas Le Moine[2], Charles Perrin[1], Maria-Helena Ramos[1], Ludovic Oudin[2], Thibault Mathevet[3], Julien Lerat[4], Lionel Berthet[5]

[1] Irstea (formerly Cemagref), Hydrosystems and Bioprocesses Research Unit, Antony, France

[2] Université Pierre et Marie Curie, UMR Sisyphe, Paris, France

[3] Electricité de France, Direction Technique Générale, Grenoble, France

[4] CSIRO Land and Water, Canberra, Australia

[5] DREAL Auvergne, Clermont-Ferrand, France

All that glitters is not gold is one of those universal truths that also applies to hydrology, and particularly to the issue of model calibration, where a glittering *mathematical* optimum is too often mistaken for a *hydrological* optimum. This commentary aims at underlining the fact that calibration difficulties have not disappeared with the advent of the latest search algorithms. While it is true that progress on the numerical front has allowed us to quasi-eradicate miscalibration issues, we still too often underestimate the remaining hydrological task: screening mathematical optima in order to identify those parameter sets which will also work sufficiently outside the calibration period.

1

Model calibration and gold mining: an analogy

The calibration process can be looked at as a task of sorting potential parameter sets, just as gold mining can be looked at as one of sorting minerals. To implement this sorting, calibration requires a method (often a search algorithm) and a specific objective function. Similarly, gold miners search river sediments for gold flakes: they use a shovel to dig sand, a classifier to separate stones from gold-bearing sand, and a pan to separate heavy minerals from the sand.

In the search for gold, the miner may be lured by fool's gold, i.e. pyrite crystals which glitter like gold but are not by any means of the same worth. Similarly, the hydrologist may be lured by parameter sets which shine over a short calibration period, but prove dull when judged either over a longer so-called calibration period, or a different validation period. These parameter sets can be considered analogous to fool's gold.

A hydrological-optimum is what we as hydrologists wish to identify through calibration. It is not simply a parameter set which allows maximizing one or more objective functions over the calibration period: it is one that ideally would permit representing the catchment under all possible calibration periods encompassing climate forcings of interest, i.e. one allowing extrapolation. However, search algorithms only provide numerical optima at best, and their level of optimality is, by definition, only guaranteed for the calibration period.

2

Numerical MIScalibration and hydrological OVERcalibration

Let us now detail the distinction between the two main types of calibration problems. Calibration consists of sorting alternative parameter sets, to maximize one (sometimes several) numerical criteria (sometimes qualitative criteria and/or the expert opinion of the modeller are also used). Two distinct reasons can explain that a promising parameter set (i.e. one showing a good fit over the calibration period) proves disappointing in the validation phase: miscalibration and overcalibration.

- **mis**calibration occurs where the search algorithm has failed to localize the exact mathematical optimum. This problem is often referred to as 'being trapped around a secondary optima'. In our mining example, this is analogous to mistaking sand for gold.

- **over**calibration occurs when one has indeed identified the mathematically optimum parameter set over the calibration period, but it is one which does not remain mathematically optimum over different periods. In our mining example, this is analogous to mistaking fool's gold for true gold.

Classical examples of mis- and over- calibration are widespread in the hydrological literature. Already in their famous study, Johnston and Pilgrim (1973) related the numerous disappointments caused by an extensive search for the optimum values of the parameters of Boughton's model. They listed all the problems which have since been recognized as the major impediments to the calibration of hydrological models (discontinuities of the response surface, multiplicity of equifinal solutions, unidentifiability, lack of robustness of calibrated parameter values…). More recently, Berthet et al. (2010) have showed how a small number of large events can have a

3

Published online 20 March 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hyp.9264

major impact on the criterion value and therefore on the identification of the optimum parameter set.

Causes of overcalibration

While miscalibration is essentially a numerical problem, overcalibration is a purely hydrological one, which manifests itself for several reasons (see e.g. Bergström, 1991; Sorooshian and Gupta, 1995; Sorooshian et al., 1983; Sorooshian and Gupta, 1983):

- ***Because of the presence of noise:***

During the calibration process, the model may not only digest the time-invariant specificities of catchment behaviour but also some of the time-varying noise existing in the observed time series. As a consequence, the parameter set identified by calibration may also be representative of the characteristics of the noise and thus lack robustness.

- ***Because of lack of information:***

We never observe the catchment over the whole range of possible climatic situations. Our calibration time-period is *always* shorter than we wished it were. Some of the functions of the catchment and hence the model may thus not be significantly activated during this period. As an extreme example, consider the parameters of a snow routine, part of a generic hydrological model. For many catchments in the warmer part of the world, a significant snow cover will not occur every year. Thus, if calibrated on a time period lacking sufficient snowfall, the parameters of the snow

4

routine will take erratic values, and be poorly representative of the long-term behaviour of the catchment. In the systems theory literature, this situation is referred to as an 'insufficient excitation of the system's modes', which is known to disturb the model identification process (Ljung, 1998).

- ***Because of structural problems:***

The structure of the hydrological model has an impact on the above-mentioned problems. A model of a complex non-engineered system is always an imperfect representation, and there is no alternative to the structure containing a certain level of aggregation of physical processes as well as of time and space scales. Thus, it would be unrealistic to expect to escape parameter identification problems entirely. At the limit, however, if provided with a sufficiently long time series that allows the activation of all of its processes, the ideal model should have optimal parameters independent from the calibration period.

Although none of our models is 'ideal' in that sense, we know that some are less ideal than others: the fact that structural problems are widespread does not mean that we cannot avoid them in large part by choosing a sound model structure (Michel et al., 2006)! A particular attention should be given to the number of parameters (Perrin et al. 2003). Jakeman and Hornberger (1993) suggested that the maximum number of parameters that can usually be identified is much lower (4-6) than what is found in most hydrological models.

5

Possible solutions to avoid overcalibration: screening mathematical optima to retain hydrological optima

Hydrologists have long been looking for ways to avoid overcalibration and identify those mathematical optima that are also hydrologically optimal. The proposed measures, which we will now discuss, can roughly be grouped in four classes (Figure 1). Remember that we purposely exclude approaches seeking to fight secondary optima, because they are not strictly speaking a way to deal with overcalibration, but rather a way to address miscalibration.

1. Changing the objective function used in calibration

Some authors have proposed addressing the overcalibration problem by changing the objective (or cost) function on which optimization is conducted. Although from a rigorous statistical point of view model calibration should include an analysis of the structure of model errors, most modellers trust standard criteria based on squared errors (typically the RMSE). Different objective functions have been proposed in the search for robustness (Schoups and Vrugt, 2010; Sorooshian et al., 1983; Thyer et al., 2009; Yang et al., 2007). Oudin et al. (2006) discussed the merit of the Nash-Sutcliffe criterion computed on a square root transformation of flows, which they compared with a multi-objective calibration scheme, while Gupta et al. (2009) discussed a decomposition of the Nash-Sutcliffe criterion, proposing an alternative that should give more robust parameter estimates. Several authors also advocated multi-objective strategies, about which a large literature now exists (see e.g. Efstratiadis and Koutsoyiannis, 2010; Vrugt et al., 2003), or empirical objective

6

functions aimed at reproducing human expertise (Ehret and Zehe, 2011; Ewen, 2011).

2.   Ensemble approaches: replacing the estimation of an optimum value by the estimation of a statistical distribution

As an alternative approach to the difficulties of model parameterisation, several hydrologists have suggested abandoning the concept of optimal value, and estimating either a family of parameter sets (Bardossy and Singh, 2008; Beven, 1993) or a statistical distribution of possible parameter values (see e.g. Thiemann et al., 2001; Thyer et al., 1999). This approach considers parameter sets as random variables that can be characterised by a distribution, which makes sense from a statistical point of view. Note however that this approach, often Bayesian, will not solve all problems: equifinal parameter sets will not disappear… they will just transmute into a characterisation of the parameters as multi-modal distributions!

3.   Guided calibration approaches: looking outside of the rainfall-runoff time series for complementary constraints

Quite surprisingly, promoters of guided calibration approaches have been coming from two apparently opposite directions:

- on one side, experimentalists (see e.g. Seibert and McDonnell, 2002) initially argued that model parameterisation should only be based on physical field measurements. Facing the disappointingly low efficiency of no-calibration strategies, they advocated the use of 'soft data' (qualitative information that could not be used directly in quantitative terms) to at least guide the calibration;

7

- on the other side, one finds the operationally-focused hydrologists interested in making the calibration task more robust, and in better extracting the information found in the calibration data. Among them, Koren et al. (2003) and Leavesley et al. (2003) argued that a good means to avoid overcalibration with distributed models was to force some a priori level of spatial and physical consistency into parameter estimates (for a lumped model, this is somewhat equivalent to the concept of regional calibration, see e.g. Fernandez et al., 2000; Parajka et al., 2007). Kuzmin et al. (2008) took a position against global optimization algorithms and suggested a local calibration approach, starting from physically-relevant a priori parameter guesses. Perrin et al. (2008) proposed a quite similar approach, with a parameter search restricted to a library of previously obtained parameter sets.

Note that guided calibration approaches can naturally be given a Bayesian interpretation, with the prior parameter distribution being seen as the main guide.

4. Questioning of the model structure

Approaches consisting of questioning the structure of a model are more difficult to find in the literature: this is something mostly done in the initial stages of model development and modellers thus rarely write about it. It is, however, sometimes mentioned *en passant*, i.e. for example by Johnston and Pilgrim (1973) who, at the very end of their calibration study, mentioned that one of the solutions to the numerous problems they had listed could be to 'review the structure of the model' (p. 135). Jakeman and Hornberger (1993) insisted on our unavoidably limited capacity to identify parameters, suggesting that it could be impossible to identify more than 4 to

8

6 parameters in a rainfall-runoff model. More recently, some hydrologists have been raising the question of the responsibility of model structures for the existence of secondary optima (Kavetski and Clark, 2010; Kavetski and Kuczera, 2007), suggesting that before casting the stone on the optimization algorithm, modellers should improve the numerical representation of their model, while others have argued for adapting the model structure to each new catchment on which a model on which a model is to be applied to (Fenicia et al., 2008).

A softer way to question model structure consists in discussing its strengths and weaknesses, and looking for an explicit characterization of the conditions under which a model performs adequately and poorly. Wagener et al. (2003) have tried it with their 'dynamic identifiability analysis' method that depicts parameter variations through time as an aid to model improvement. Young (2011) has stressed the virtues of recursive time series methods for indicating model parameter variation and hence model structure inadequacy. It seems, however, that one can go a long way by first identifying what a model is good at and what it is not so good at. This would involve not only assessing which parts of the hydrograph are predicted well but also how the model performs under different types of conditions (an interesting package to help in this aim is presented by (Andrews et al., 2011): it allows separation of a hydrologic or rainfall time series into events and establishing model performance for different types of events.
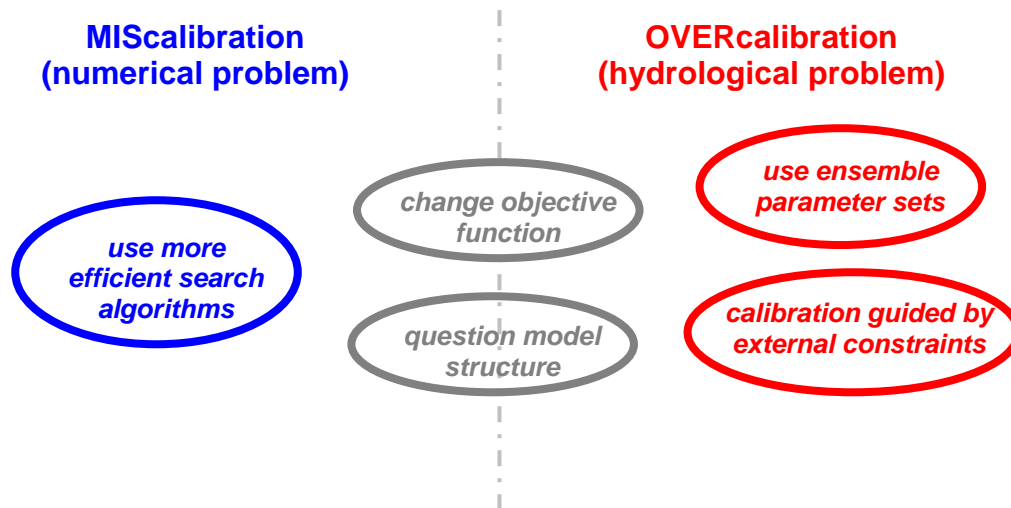
9

Figure 1: Possible solutions to miscalibration and overcalibration problems in hydrological modelling

Perspectives

This commentary has attempted to highlight the difference existing between miscalibration and overcalibration in hydrological modelling. It has reviewed some of the major solutions, successively proposed over the last few decades. For many years, hydrologists have been focusing on the miscalibration issue, and research has focused mostly on numerical methods. One could say that secondary optima have sometimes been the trees for which many hydrologists have been unable to see the forest. Today, miscalibration has been solved for most models, and the effect of overcalibration is more apparent. Research is still needed on the solutions listed above, either separately or in combination, in order to ensure that our mathematically-optimal parameter sets are also hydrologically-optimal.

Acknowledgements

References

Andrews, F.T., Croke, B.F.W., Jakeman, A.J., 2011. An open software environment for hydrological model assessment and development. Environmental Modelling & Software, 26(10): 1171-1185.

Bardossy, A., Singh, S.K., 2008. Robust estimation of hydrological model parameters. Hydrology and Earth System Sciences, 12: 1273-1283.

Bergström, S., 1991. Principles and confidence in hydrological modelling. Nordic Hydrology, 22: 123-136.

Berthet, L., Andréassian, V., Perrin, C., Loumagne, C., 2010. How significant are quadratic criteria? Part 2. On the relative contribution of large flood events to the value of a quadratic criterion. Hydrological Sciences Journal, 55(6): 1063-1073.

Beven, K.J., 1993. Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources, 16: 41-51.

Efstratiadis, A., Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrological Sciences Journal, 55(1): 58-78.

Ehret, U., Zehe, E., 2011. Series distance - an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. Hydrology and Earth System Sciences, 15(3): 877-896.

Ewen, J., 2011. Hydrograph Matching Method for Measuring Model Performance. J. Hydrol., 408(1-2): 178-187.

Fenicia, F., Savenije, H.H.G., Matgen, P., Pfister, L., 2008. Understanding catchment behavior through stepwise model concept improvement. Water Ressources Research, 44(1).

Fernandez, W., Vogel, R.M., Sankarasubramanian, S., 2000. Regional calibration of a watershed model. Hydrological Sciences Journal, 45(5): 689-707.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol., 377(1-2): 80.91.

Jakeman, A.J., Hornberger, G.M., 1993. How much complexity is warranted in a rainfall-runoff model? Water Resour. Res., 29(8): 2637-2649.

Johnston, P.R., Pilgrim, D.H., 1973. A study of parameter optimisation for a rainfall-runoff model, Water Research Laboratory of the University of new South Wales, Manly Vale, N.S.W.

Kavetski, D., Clark, M.P., 2010. Numerical troubles in conceptual hydrology: approximations, absurdities and impact on hypothseis testing. Hydrological Processes, 25(4): 661-670.

Kavetski, D., Kuczera, G., 2007. Model smoothing strategies to remove microscale discontinuities and spurious secondary optima im objective functions in hydrological calibration. Water Resour. Res., 43(3).

Koren, V., Smith, M., Duan, Q., 2003. Use of a priori parameter estimates in the derivation of spatially consistent parameter sets of rainfall-runoff models. In: Duan, Q., Gupta, H., Sorooshian, S., Rousseau, A., Turcotte, R. (Eds.), Calibration Watershed Models, Water Sci. Appl. 6. AGU, pp. 239- 254.

Kuzmin, V., Seo, D.J., Koren, V., 2008. Fast and efficient optimization of hydrologic model parameters using a priori estimates and stepwise line search. J. Hydrol., 353: 109-128.

Leavesley, G., Hay, L.E., Viger, R.J., Marstrom, S.L., 2003. Use of *a priori* parameter estimation methods to constrain calibration of distributed-parameter models. In: Duan, Q., Gupta, H., Sorooshian, S., Rousseau, A., Turcotte, R. (Eds.), Calibration Watershed Models Water Sci. Appl., 6, AGU, pp. 255-266.

Ljung, L., 1998. System Identification-Theory for the User, 2nd Edition. Prentice Hall.

Michel, C., Perrin, C., Andréassian, V., Oudin, L., Mathevet, T., 2006. Has basin-scale modelling advanced beyond empiricism? In: Andréassian, V., Hall, A., Chahinian, N., Schaake, J. (Eds.), Large sample basin experiments for hydrological model parameterization. results of the Model Parameter Experiment - MOPEX. IAHS Publication 307. IAHS, Wallingford.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., Michel, C., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterization. Water Resour. Res., 42(7).

Parajka, J., Blöschl, G., Merz, R., 2007. Regional calibration of catchment models: Potential for ungauged catchments. Water Resour. Res., 43(W06406): doi:10.1029/2006WR005271.

Perrin, C., Andréassian, V., Mathevet, T., Le Moine, N., 2008. Discrete parameterization of hydrological models: evaluating the use of parameter sets libraries over 900 catchments. . Water Resour. Res., 44: W08447, doi:10.1029/2007WR006579.

Schoups, G., Vrugt, J.A., 2010. A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res., 46: W10531.

Seibert, J., McDonnell, J.J., 2002. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. Water Resour. Res., 38(11): 1241.

Sorooshian, S., Gupta, H., 1995. Model calibration. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, pp. 23-68.

Sorooshian, S., Gupta, H., Fulton, J.L., 1983. Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall-runoff models: Influence of calibration data variability and length on model credibility. Water Resour. Res., 19(1): 251-259.

Sorooshian, S., Gupta, V.K., 1983. Automatic calibration of conceptual rainfall-runoff models: the question of parameter observability and uniqueness. Water Resour. Res., 19(1): 260-268.

12

Thiemann, M., Trosset, M., Gupta, H., Sorooshian, S., 2001. Bayesian recursive parameter estimation for hydrologic models. Water Resour. Res., 37(10): 2521-2535, doi:10.1029/2000WR900405.

Thyer, M., Kuczera, G., Bates, B.C., 1999. Probabilistic optimization for conceptual rainfall-runoff models: A comparison of the shuffled complex evolution and simulated annealing algorithms. Water Resour. Res., 35(3): 767-773.

Thyer, M. et al., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resources Research, 45: W00B14.

Vrugt, J.A., Gupta, H.V., Bastidas, L.A., Bouten, W., Sorooshian, S., 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water Resour. Res., 39(8): SWC51-SWC519.

Wagener, T., McIntyre, N., Lees, M., Wheater, H., Gupta, H., 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. Hydrological Processes, 17(2): 455-476.

Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., 2007. Hydrological modelling of the Chaohe basin in China: Statistical model formulation and Bayesian inference. J. Hydrol., 340(3-4): 167-182.

Young, P.C., 2011. Recursive estimation and time-series analysis. Springer, Heidelberg, 504 pp.