Journal of Hydrology 420–421 (2012) 171–182

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

A review of efficiency criteria suitable for evaluating low-flow simulations

Raji Pushpalatha [a], Charles Perrin [a,*], Nicolas Le Moine [b], Vazken Andréassian [a]

[a] Cemagref, UR HBAN, 1 rue Pierre-Gilles de Gennes, CS 10030, 92761 Antony Cedex, France
[b] Université Pierre et Marie Curie, UMR 7619 Sisyphe, 4 Place Jussieu, 75252 Paris Cedex 05, France

ARTICLE INFO

Article history:
Received 5 May 2011
Received in revised form 27 September 2011
Accepted 29 November 2011
Available online 8 December 2011
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Ezio Todini, Associate Editor

Keywords:
Low flows
Simulation efficiency
Model evaluation
Benchmark model
Criteria

SUMMARY

Low flows are seasonal phenomena and an integral component of the flow regime of any river. Because of increased competition between water uses, the demand for forecasts of low-flow periods is rising. But how low-flow predictions should be evaluated? This article focuses on the criteria able to evaluate the efficiency of hydrological models in simulating low flows. Indeed, a variety of criteria have been proposed, but their suitability for the evaluation of low-flow simulations has not been systematically assessed.

Here a range of efficiency criteria advised for low flows is analysed. The analysis mainly concentrates on criteria computed on continuous simulations that include all model errors. The criteria were evaluated using two rainfall–runoff models and a set of 940 catchments located throughout France. In order to evaluate the capacity of each criterion to discriminate low-flow errors specifically, we looked for the part of the hydrograph that carries most of the weight in the criterion computation.

Contrary to what was expected, our analysis revealed that, in most of the existing criteria advised for low flows, high flows still make a significant contribution to the criterion's value. We therefore recommend using the Nash–Sutcliffe efficiency criterion calculated on inverse flow values, a valuable alternative to the classically used criteria, in that on average it allows focusing on the lowest 20% of flows over the study period.

1. Introduction

1.1. In the jungle of efficiency criteria

Hydrological modelling aims at understanding and interpreting catchment hydrological behaviour. It is also used to address a number of practical issues, ranging from flood estimation to water resources management and low-flow forecasting. Whatever model is applied, the model user needs appropriate and meaningful indicators informing on the actual capacity of the model.

However, the evaluation of goodness-of-fit is not as straightforward as it may seem at first glance. Of course, model performance can first be evaluated by the visual comparison of the observed and simulated flow hydrographs, but this remains extremely dependent on the evaluator's experience (Chiew and McMahon, 1993; Houghton-Carr, 1999). A more objective way to evaluate model performance is to use numerical criteria, but then the user may get lost in a jungle of potential criteria. Why is the choice so difficult? Several reasons can be put forward:

1. flows vary by several orders of magnitude that may not be equally useful for the modeller;
2. hydrological models often produce heteroscedastic errors, i.e. their variance is not independent of the flow value;
3. the range of target flows may vary significantly between evaluation periods;
4. the model may be used for different applications, which may require specific criteria.

For these reasons, a large variety of criteria have been proposed and used over the years in hydrological modelling, as shown for example by the lists of criteria given by the ASCE (1993), Dawson et al. (2007), Moriasi et al. (2007) and Reusser et al. (2009). Among these criteria, some are *absolute* criteria such as the widely used root mean square error, while others are *relative* criteria (i.e. normalized) such as the Nash and Sutcliffe (1970) efficiency criterion (*NSE*). In the latter, model errors are compared to the errors of a reference or benchmark model (Seibert, 2001; Perrin et al., 2006). This provides a useful quantification of model performance in that it indicates to which extent the model is better (or worse) than the benchmark. It also facilitates the comparison of performance between catchments.

However, the choice of a benchmark is difficult: different benchmarks are more or less demanding (and thus the comparison more or less informative) depending on the type of hydrological regime or the type of model application. This sometimes makes it difficult to interpret the relative criteria and a bit puzzling for an inexperi-

* Corresponding author. Tel.: +33 140966086; fax: +33 140966199.
E-mail address: charles.perrin@irstea.fr (C. Perrin).

0022-1694/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2011.11.055

enced end-user, who may simply want to know whether the model can be considered as "good", "acceptable" or "bad". Actually, there is no single criterion that can evaluate model performance in all cases (Jain and Sudheer, 2008). Many authors use several criteria simultaneously (see e.g. Efstratiadis and Koutsoyiannis (2010) for a review in the context of multi-objective calibration), but a discussion of these approaches is not within the scope of this article.

The merits and drawbacks of several efficiency criteria have already been discussed and debated in the literature, as well as the links between them. The Nash–Sutcliffe efficiency criterion has probably received the most attention (Garrick et al., 1978; Houghton-Carr, 1999; McCuen et al., 2006; Schaefli and Gupta, 2007; Clarke, 2008; Gupta et al., 2009; Moussa, 2010; Gupta and Kling, 2011). Like many other criteria based on the mean model square error, this criterion is known to put greater emphasis on high flows when calculated on a continuous simulation. Although it was shown to have several limitations (such as its sensitivity to the hydrological regime, sample size or outliers), it remains a valuable and popular means to evaluate models for high-flow simulation.

Comparatively little work has been carried out on the meaning and interpretation of the criteria used to evaluate models in low-flow conditions. The following section summarizes the existing studies.

1.2. Criteria used for the evaluation of low-flow simulation

Table 1 lists some of the studies discussing performance criteria able to judge low-flow simulations. Note that the various criteria formulations listed here depend on at least three factors:

1.2.1. Calculation period

Instead of calculating criteria only over the low-flow periods (which generally requires the subjective choice of a low-flow threshold), most of the existing criteria calculate model errors over the entire test period. Thus, they give some weight to the errors in low-flow as well as to the errors in high-flow conditions.

1.2.2. Target variable

The second major aspect in calculating criteria is the choice of a target variable. Some authors (e.g. Houghton-Carr, 1999) appeal to statistical measures classically used to characterize low flows, such as the base-flow index, a percentile of the flow duration curve (*FDC*) or some minimum accumulated flows over a continuous period (e.g. 7 days). By calculating the ratio between the simulated and observed values, a relative efficiency criterion is obtained. These criteria are very useful when studies focus on specific aspects of low flows. However, one may also continue calculating the sum of errors over the entire test period, provided that the appropriate transformation on flows is used (Box and Cox, 1964; Chiew et al., 1993). This transformation helps put more weight on low flows. The root square or the logarithms are among the most widely used transformations on low-flow values. For example, Smakhtin et al. (1998), Houghton-Carr (1999), Oudin et al. (2006), Jain and Sudheer (2008), and de Vos et al. (2010) used the sum of squared residuals calculated on the logarithms of flow values in order to reduce the biasing towards peak flows. Chiew et al. (1993) used the root squared transform to evaluate the model's performance in low-flow conditions. Krause et al. (2005) proposed using a relative variable as the ratio between simulation and observation, and calculated the distance of this variable from 1. Le Moine (2008) discussed a generalization of these transformations as a power law transformation with positive or negative exponents and proposed a family of squared criteria based on the power transformation of flows (of Box-Cox type), defined by:

$$RMSE(\lambda) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_i^{\lambda} - \widehat{Q}_i^{\lambda})^2} \tag{1}$$

where λ is the power of the flow transformation, n is the number of time steps and Q_i and $\widehat{Q}_i$ are the observed and simulated flows, respectively, at time step i. λ is not necessarily an integer and can take positive and negative values. When λ tends towards zero, the transformation tends towards the logarithm transformation. As noted by Le Moine (2008), low and high values of λ will tend to emphasize the model errors on the minimum and maximum flow values, respectively. For example, Chiew et al. (1993) had used a value of λ equal to 0.2 to give more emphasis on low flows.

1.2.3. Error normalization

Another aspect that differs between criteria is the type of error used and the way model error is normalized. Most of the criteria are based on the squared residuals, but absolute errors can also be considered. A power of these absolute errors may also be used (see e.g. Krause et al., 2005). In terms of model error normalization, most of the efficiency indexes use the form of the *NSE*. Willmot (1984) proposed the index of agreement as another way to normalize model square error, by dividing it by the potential error.

1.3. Are existing criteria appropriate to evaluate low-flow simulations?

Only a few authors have discussed the suitability of the variety of existing criteria to evaluate low-flow simulations. Oudin et al. (2006) compared several objective functions and concluded that the square root transformation provides an all-purpose efficiency measure, not specifically focusing on low flows. Analysing several efficiency indices, Krause et al. (2005) showed that some criteria are closely related while others show very different patterns. They advised using the relative efficiency index for the evaluation of low-flow simulations, noting that the logarithm transformation on flows provides a higher sensitivity to low flows, although they indicate that this criterion remains sensitive to high flows.

Actually, the impact of flow transformations significantly changes the way the hydrograph and model errors are considered in criteria, as shown by Le Moine (2008). This is illustrated in Fig. 1 in the case of natural, root square and logarithmic transformed flows. Fig. 1 shows the series of flow values and model errors as well as the cumulated quadratic error (to facilitate the comparison

Table 1
Studies that used criteria to evaluate low-flow simulation quality.

References	Low flow statistics	Residual errors	Standard deviation	Bias	Coeff. of determination	Nash efficiency (all transforms)	Coeff. of variation	Log MSE	Relative efficiency	Index of agreement
Chiew et al. (1993)						x				
Ye et al. (1998)	x	x				x				
Houghton–Carr (1999)	x					x				
Krause et al. (2005)					x	x			x	x
Oudin et al. (2006)		x				x				
Jain and Sudheer (2008)	x		x	x		x	x			
de Vos et al. (2010)								x		