Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment–MOPEX. IAHS Publ. 307, 2006.

A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins

THIBAULT MATHEVET[1,2], CLAUDE MICHEL[1], VAZKEN ANDRÉASSIAN[1] & CHARLES PERRIN[1]

1 *Cemagref, Parc de Tourvoie, BP 44, F-92163 Antony cedex, France*
thibault.mathevet@edf.fr

2 *now at: EDF-DTG, 21, Avenue de l'Europe, BP 41, F-38040 Grenoble cedex 09, France*

Abstract Rainfall–runoff models are useful tools for hydrological research, water engineering and environmental applications. Given the large number of available rainfall–runoff models, many comparative studies have been done to compare models performances, to specify their domain of application and provide guidance to end-users. Although most existing comparative studies tested models only on a few basins, we believe that effective model evaluation requires large samples of test catchments. However, large test samples raise the issue of appropriate criteria to quantify model performances. This paper shows that the widely used Nash and Sutcliffe criterion may be difficult to apply for large test samples and that a bounded version of this criterion (called C_{2M}) is better suited for extensive model assessment.

Key words large sample of basins; Nash-Sutcliffe criterion; rainfall–runoff modelling

INTRODUCTION

Since the 1960s, hydrologists have developed a large number of more or less complex rainfall–runoff (RR) models. As a consequence of this proliferation, the need for comparative studies appeared quite early (WMO, 1975). Linsley (1982) suggested that *"because almost any model with sufficient free parameters can yield good results when applied to a short sample from a single basin, effective testing requires that models be tried on many basins of widely differing characteristics, and that each trial cover a period of many years"*. Few modellers have, however, followed these recommendations, and most of the RR modelling studies reported in the literature present the performances of one RR model on a single basin (or on a small number of similar basins). If studies include too few watersheds, the validity of their conclusions will be limited to the hydro-climatic domain of the test sample. Conversely, a large set of basins provides a general overview of the efficiency of one or several models, in a wide range of hydro-meteorological conditions and catchment physical characteristics (geology, soil, vegetation, topography, land use, etc.).

With large basins sets, however, assessing RR model performances becomes complex since a complete distribution of results is obtained, often over a large range of performances. In such cases, one must find ways to compare these distributions, or to summarize them into proper statistics, provided that the formulation of the selected criterion allows deriving such a summary.

In this paper we propose a criterion formulation suitable for comparing model performances on large basin samples. An application is made on a sample of 313 basins. We also show the usefulness of large basin sets for model assessment.

In the first section we discuss the need for using large basin samples. Then we introduce a new formulation of the classical Nash & Sutcliffe (1970) criterion, better suited to the assessment of models on large samples. In the following section we present the basin sample, the tested models and the assessment methodology. Last, we present the results of model tests and discuss the usefulness of this new criterion formulation for large basin samples.

WHY ARE LARGE BASIN SAMPLES NEEDED TO ASSESS RR MODELS EFFICIENCIES?

Most existing comparative studies rely only on a small number of watersheds. Among exceptions are the studies by Vandewiele *et al. (*1992), Makhlouf & Michel (1994) and Xu & Vandewiele (1995) at the monthly time-step, and by Perrin *et al.* (2001) at the daily time-step. Nash & Sutcliffe (1970), Linsley (1982) and Klemeš (1986) stressed that generality should be a fundamental requirement for any RR model. Linsley (1982) argued that *"it seems axiomatic that the fundamental processes of hydrology are the same in all catchments.[…] In some cases a process may not be present.[…] However, these differences do not mean that a single model cannot be applied in all cases. The model must represent the various processes with sufficient fidelity so that irrelevant processes can be "shut off" or will simply not function. Differences […] must be represented by model parameters which can be preset to represent these characteristics.* (p. 14)*"*.

Thus, to assess model generality, it seems obvious that model assessment cannot rely only on a single or few basins. Given the large differences that exist between basins, a large number of basins is required to judge of the actual adaptability of a RR model to different conditions.

However, large sets of data raise some problems. It is difficult to systematically check data quality since data come from many different sources and result from different collection practices, so errors may affect data. There may also be influences (e.g. due to human activities) difficult to detect by visual inspection of time series. For these reasons, there will inevitably be a number of basins in the test set where the models will fail. These basins should not be excluded from the test set since, as argued by Linsley (1982), all models will suffer equally from these problems and this will not bias the comparative model assessment. However, these model failures may cause problems in the quantification of the average model performance over the test sample if appropriate criteria are not chosen. This need for appropriate criteria is discussed in the next section.

AN APPROPRIATE CRITERION FOR MODEL ASSESSMENT OVER LARGE BASIN SAMPLES

The Nash & Sutcliffe (1970) criterion is widely used in hydrological modelling. However, for model assessment on large basins sets, this criterion raises some problems.

The Nash & Sutcliffe criterion: advantages and drawbacks

The approach followed by Nash & Sutcliffe (1970) is to build a relative index of agreement (or disagreement) between observed and computed runoff that could be used to compare model performances between periods or basins.

They start from the sum of square errors given by:

$$F = \sum_{i=1}^{n} \left(Q_{obs,i} - Q_{sim,i} \right)^2 \tag{1}$$

where F is the index of disagreement, $Q_{obs,i}$ and $Q_{sim,i}$ are the observed and simulated discharges at time step i, the sum being taken over n time steps of a pre-selected period. F is analogous to the residual variance of a regression analysis. The initial variance F_0 is given by:

$$F_0 = \sum_{i=1}^{n} \left(Q_{obs,i} - \overline{Q_{obs}} \right)^2 \tag{2}$$

where $\overline{Q_{obs}}$ is the mean of the observed discharge over the pre-selected period. Nash & Sutcliffe (1970) then define the efficiency of the model E as the proportion of the initial variance accounted for by the model:

$$NS = 1 - \frac{F}{F_0} \tag{3}$$

NS can take values between $-\infty$ and 1. A value of 1 indicates a perfect agreement and a value of zero indicates that the model does not explain any part of the initial variance.

The Nash & Sutcliffe criterion can also be interpreted as a criterion that determines the improvement made by a given model in simulating flows in comparison with a reference model that would simulate a flow equal to $\overline{Q_{obs}}$ at each time step. The value of zero for the criterion therefore means that the model is not better than this basic one-parameter model, and a negative value indicates that the model is worse than this basic model.

This criterion is very useful in model assessment since its adimensional form is hoped to allow comparison of performances on different catchments or periods. Several authors mention, however, that this criterion has some drawbacks. Garrick *et al.* (1978) and Martinec & Rango (1989) showed that the NS criterion may produce relatively high values, even for quite poor models. This is mainly due to the fact that the basic model ($Q_i = \overline{Q_{obs}}$ for all i) can be very primitive in some instances, so it becomes easy to be better than this basic reference ($F << F_0$). To overcome this problem, Garrick *et al.* (1978) proposed to use another reference model, such as a "seasonal model", to compare in F_0 the measured runoff to the long-term average measured runoff for each Julian day. Conversely, it is difficult to obtain high values for periods or basins where flow does not vary much in time.

These considerations indicate that this criterion may not be similarly demanding in all circumstances and that it can yield wide ranges of performances when models are assessed on large basin samples that include many different characteristics. This is all

the more true as this criterion has no lower bound and may give strongly negative values when the model fails.

The problem of criteria formulation for extensive model assessment

If one focuses on comparative assessments, the simplest case is to test two models on one basin. There are numerous ways of comparing performances in such a simple case. However, when the number *n* of models and the number *m* of test basins increase, the classification of models becomes more complex. Indeed, one has to compare *n* lists of *m* values of the NS criteria. To summarize such a large amount of information, Perrin *et al.* (2001) used:

– the mean value of the *m* NS criteria,
– the distribution of NS criteria obtained by each model over the basin sample and the percentiles (0.1, 0.5, 0.9) of this distribution.

The mean performance is actually the best measure to have a synthetic overview of model performance. Unfortunately, the mean of NS criteria can be heavily influenced by a few strongly negative values obtained on a small number of basins. Therefore the mean value can be artificially biased, which may impair the conclusions of the comparison exercise. The use of distribution percentiles (e.g. the median value) could be a remedy to this problem. Unfortunately the distributions of NS criteria may cross each other and therefore the relative value of two models will depend on the selected percentile. It could also be argued that basins where some models do not work should be discarded. However, this would bias the comparison in favour of the models used for this prior screening.

C_{2M}, a bounded formulation for the Nash & Sutcliffe criterion

To avoid these problems, we propose to adopt a new formulation of the NS criterion to make it vary between –1 and +1, like a correlation coefficient. This will generate less skewed criteria distributions and it will be possible to compute significant mean values. To keep the same zero value as the NS criterion, we propose the following formulation, called C_{2M}:

$$C_{2M} = \frac{1 - \frac{F}{F_0}}{1 + \frac{F}{F_0}} \tag{4}$$

The NS criterion is related to C_{2M} as follows (Fig. 1):

$$NS = \frac{2.C_{2M}}{1 + C_{2M}} \tag{5}$$

$$C_{2M} = \frac{NS}{2 - NS} \tag{6}$$

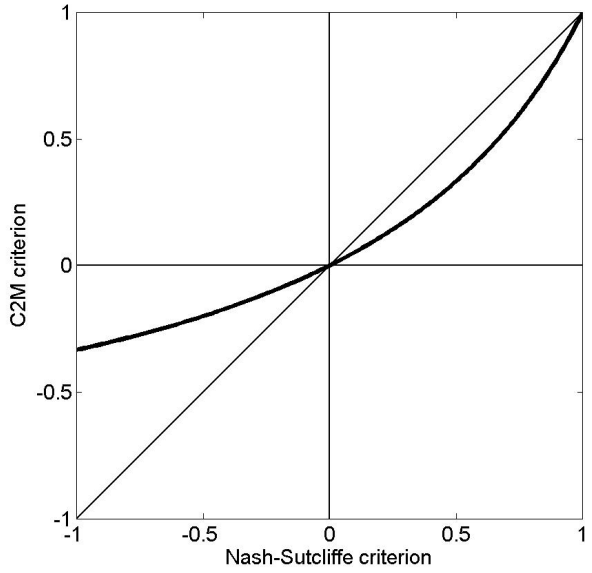

Fig. 1 Relation between the Nash-Sutcliffe and the C_{2M} criteria.

Note that the C_{2M} is less optimistic than the NS criterion for positive values, which partially provides an answer to the criticism made by Garrick *et al*. (1978) who argued that the NS criterion produces too high values. Since the criteria distributions are now bounded, it should be possible to derive more meaningful statistics to summarize model performances.

DATA AND METHODS

Basin sample test

To test the usefulness of this new criterion formulation, we used a sample of 313 basins (see characteristics in Table 1). Most basins are in France (227) and in the USA (70), some in Australia (12), Spain (2) and Slovenia (2). Hydro-climatic conditions are varied: semiarid, Mediterranean, oceanic, temperate, mountainous and continental. Snowmelt influences are generally limited. Data were collected at the hourly time step.

Rainfall–runoff models

We used three model structures derived from existing RR models (see Table 2). All three are lumped and continuous, and were run at the hourly time step. More details can be found in Mathevet (2005). As the purpose of this study is to discuss methodological aspects and not to compare original models, the models are called M1, M2 and M3 hereafter. The models were fed with exactly the same input data, i.e. rainfall time-series and potential evapotranspiration estimates, and their free parameters were calibrated against observed runoff.

Table 1 Minimum–mean–maximum characteristics of the 313 basins.

Country	France	USA	Australia	Spain	Slovenia
Number of watersheds	227	70	12	2	2
Annual runoff (mm)	*35*–44–*1655*	*0*– 279–*1612*	*9*–35–*96*	429–736	1024–1182
Annual rainfall (mm)	*403*–963–*2067*	*193*–1163–*2996*	*569*–674–*1025*	1183	1384
Annual PE (mm)	*595*–791–*1252*	*1104*–1545–*2085*	1226	639	735
Annual runoff–rainfall ratio(%)	*0.05*–0.44–*2.59*	*0*– 0.20 –*0.81*	*0.01*–0.05–*0.13*	0.36–0.62	0.74–0.85
Type of climate	Temperate, Mediterranean, oceanic, continental	Temperate, oceanic, semiarid	Semiarid	Mountainous Mediterranean	Mountainous
Watershed area (km²)	*1.1*–280–*4978*	*1.2*–33–*334*	*2.7*–48–*2538*	0.56–4.17	457–1385
Length of the time series (year)	*3*–8–*34*	*3*–11–*43*	6	3–4	5

Table 2 Characteristics of the three RR models used in this study. The details of the modified versions tested here can be found in Mathevet (2005).

Tested model	Number of free parameters	Number of reservoirs	Original model	Reference of original model
XINANJ	8	4	XINANJIANG	Zhao *et al*. (1995)
IHAC	6	3	IHACRES	Jakeman *et al*. (1990)
GR4H	4	2	GR4J	Perrin *et al*. (2003)

Assessment methodology

To assess the performances of the selected RR models, we applied the split-sample test procedure (Klemeš, 1986): the models can thus be tested in simulation mode, under meteorological conditions different from those of the calibration period. For each basin, the available time-series was split into several independent sub-periods. The models were successively calibrated on each sub-period and tested in validation mode on the remaining ones. The results are based on a total of 2093 validation tests, with 2.2 year long periods on average, for the 313 basins used here. Periods are quite short but this is not a problem since all models were strictly compared in the same conditions.

RESULTS AND DISCUSSION

In this section, we use model results to demonstrate: (i) that it is interesting to resort to large basins samples in order to compare RR models; and (ii) that the C_{2M} criterion is more valuable for such a comparison. We present here the results of the three selected RR models. They were tested on the whole sample of 313 basins and on randomly generated sub-samples as explained below.

To demonstrate the interest of large basin sets to compare the efficiencies of RR models, 500 random sub-samples of 5, 10 and 50 watersheds were drawn from the initial sample (313 basins). Fig. 2 clearly shows the high dispersion of mean NS criteria for the 500 random samples of 5, 10 and 50 watersheds, when compared to the mean NS efficiency over 313 basins (black cross). This figure shows that:

– when using a small number of basins, it is always possible to find samples of a limited number of basins, where model A is better than model B on one sample, and conversely model B is better than model A on the other;
– even with random sub-samples of 50 basins, it remains possible to find cases where M2 and M3 are better than M1 and M3 is better than M2, whereas the opposite conclusions are drawn on the whole sample.

The high dispersion of mean NS values averaged over 10 or even 50 basins is mainly caused by the basins where models fail dramatically. When the sample size increases, the weight of these basins is reduced.

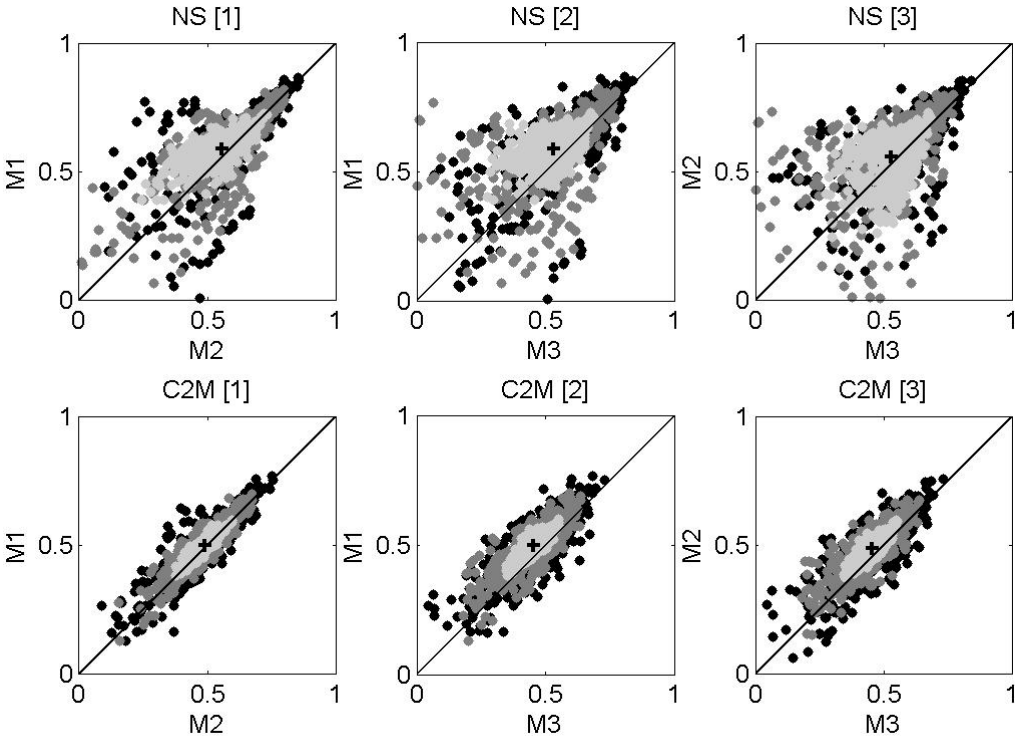

Fig. 2 Comparison of M1, M2 and M3 efficiencies of 500 random samples of 5, 10, 50 basins, with Nash-Sutcliffe and C_{2M} criteria (note that basins samples whose mean NS efficiency is lower than 0 are not shown). Black dots, 5 watersheds; dark grey dots, 10 watersheds ; Light grey dots, 50 watersheds.

The use of the C_{2M} criterion eases the comparison since the dispersion of efficiencies in Table 3 is much more limited with the C_{2M} than with the NS criterion. With random sub-samples of 10 or 50 basins, it is still possible to find samples of basins where any model is better than another. However, with the C_{2M} criterion, the dispersion of efficiencies' difference between two RR models decreases greatly

Table 3 Mean model efficiency using NS and C_{2M} criterion, over 313 watersheds, as a function of the chosen differentiation criterion.

	M1	M2	M3
Minimum NS	–815.2	–648.5	–1778.2
Mean NS	55.7	52.9	38.5
Maximum NS	89.5	87.0	85.4
NS Percentile			
0.1	10.3	0	–11.2
0.5	71.5	68.6	62.3
0.9	87.0	85.0	82.6
NS standard deviation	69.4	60.4	120.9
% of NS			
< –100 %	1.5	2.2	3.5
< 0 %	6	10.5	11.5
Mean C_{2M}	48.9	45.5	38.7
C_{2M} Percentile			
0.1	5.4	0	–5.0
0.5	55.6	52.2	45.2
0.9	77.0	74.0	70.3
C_{2M} standard deviation	28.0	28.2	31.2

when the sample size increases. In this case, 50 watersheds seem to be sufficient to discriminate the efficiencies of M1 and M3 models, or M2 and M3 models.

Fewer basins are needed to rank two models when using C_{2M} than with the NS criterion: when the difference of C_{2M} efficiency between two models over the whole sample is about 10 points (the case of M1 *vs* M3 and M2 *vs* M3), a sample of about 50 basins seems sufficient to differentiate these models, whereas 50 are clearly not enough in the case of the NS criterion. When the difference is smaller, the size of the required basin sample will increase, but there are less cases of obvious misinterpretation with the C_{2M} than with the NS criterion.

CONCLUSION

The objectives of this article were to advocate the use of large samples of basins to assess and compare RR models and to present the usefulness of a bounded version of the Nash-Sutcliffe criterion (called C_{2M}) in such an assessment.

The lack of a lower bound of the NS criterion is a real drawback and yields strongly skewed distribution of efficiencies over large test samples when, for any reason, very low performances are obtained on a few basins. As a consequence, it is hard to compute meaningful statistics to summarize the whole distribution. The C_{2M} criterion introduced here provides a bounded formulation of the NS criterion, varying within the interval [–1, +1]. The test of three model structures on a sample of 313 basins clearly shows that the C_{2M} criterion allows one to compute more meaningful mean model efficiencies over large test samples than the NS criterion, and therefore provides a more reliable comparison of model efficiencies. It is also shown that, given the variability of model efficiencies over an heterogeneous sample of basins, a large sample size (at least 50 basins) is clearly required to differentiate the efficiencies of

RR models. The smaller the difference between models, the larger the sample size required to warrant that the difference is significant.

Acknowledgements The authors would like to thank the institutions and researchers who provided data sets for model testing: Christian Scherer (French Ministry of Ecology and Sustainable Development); Bruno Rambaldelli (Météo France); Christian Zammit (Water and River Commission of Western Australia), Muguresu Sivapalan (Center for Water Research of the University of Western Australia); Mira Kobold (Environmental Agency of the Republic of Slovenia) Fransesc Gallart and Jerôme Latron (Jaume Almera Institute of Earth Sciences). Thanks are also due for the huge work of those who did the field data collection.

REFERENCES

Garrick, M., Cunnane, C. & Nash, J. E. (1978). A criterion of efficiency for rainfall–runoff models. *J. Hydrol*. **38**, 375–381.

Jakeman, A. J., Littlewood, I. G. & Whitehead, P. G. (1990) Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. *J. Hydrol*. **117**, 275–300.

Klemeš, V. (1986) Operational testing of hydrological simulation models. *Hydrol. Sci. J.* **31**(1), 13–24.

Linsley, R. K. (1982) Rainfall–runoff models—an overview. In: *Proc. Int. Symp. on Rainfall–Runoff Modelling* (ed. by V. P. Singh), 3–22. Water Resources Publications, Littletown, Colorado, USA.

Makhlouf, Z. & Michel, C. (1994) A two-parameter monthly water balance model for French watersheds. *J. Hydrol*. **162**(3–4), 299–318.

Mathevet, T. (2005) Which rainfall–runoff model at the hourly time-step ? Empirical development and intercomparsison of rainfall–runoff models on a large sample of watersheds. PhD thesis, ENGREF University, Paris, France (in French).

Martinec, J. & Rango, A. (1989) Merits of statistical criteria for the performance of hydrological models. *Water Resour. Bull*. **25**(2), 421–432.

Nash, J. E. & Sutcliffe, J. V. (1970) River flow forecasting through conceptual models. Part I—A discussion of principles. *J. Hydrol.* **27**(3), 282–290.

Perrin, C., Michel, C. & Andréassian, V. (2001) Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *J. Hydrol*. **242**(3–4), 275–301.

Vandewiele, G. L., Xu, C. Y. & Win, N. L. (1992) Methodology and comparative study of monthly models in Belgium, China and Burma. *J. Hydrol*. **134**, 315–347.

WMO (1975) Intercomparison of conceptual models used in operational hydrological forecasting. *Operational Hydrology Report no. 7, WMO no 429.* World Meteorological Organization, Geneva, Switzerland.

Xu, C. Y. & Vandewiele, G. L. (1995) Parsimonious monthly rainfall–runoff models for humid basins with different input requirements. *Adv. Water Resour.* **18**, 39–48.

Zhao, R J. & Liu, X. R. (1995) The Xinanjiang model. In: *Computer Models of Watershed Hydrology*, Chap. 7 (ed. by V. P. Singh), 215–232. Water Resources Publications, Littletown, Colorado, USA.