INVITED COMMENTARY

HP TODAY

HYDROLOGICAL PROCESSES
Hydrol. Process. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/hyp.6825

Do Nash values have value?

Bettina Schaefli[1]* and Hoshin V. Gupta[2]

[1] *Institute of Geoecology, University of Potsdam, 14 476 Potsdam-Golm, Germany*
[2] *Department of Hydrology & Water Resources, The University of Arizona, Tucson, AZ 85 721, USA*

*Correspondence to:
Bettina Schaefli, Institute of Geoecology, University of Potsdam, 14 476 Potsdam-Golm, Germany.
E-mail: bettina.schaefli@a3.epfl.ch

Received 6 May 2007
Accepted 10 May 2007

How Do We Communicate Model Performance?

The process of model performance evaluation is of primary importance, not only in the model development and calibration process, but also when communicating the results to other researchers and to stakeholders. The basic 'rule' is that every modelling result should be put into context, for example, by indicating the model performance using appropriate indicators, and by highlighting potential sources of uncertainty, and this practice has found its entry into the large majority of papers and conference presentations.

While the question of how to communicate the performance of a model to potential end-users is currently receiving increasing interest (e.g. Pappenberger and Beven, 2006), we–as well as many other colleagues–observe regularly that researchers take much less care when communicating model performance amongst ourselves. We seem to assume that we are speaking about familiar performance concepts and that they have comparable significance for various types of model applications and case studies. In doing so, we do not pay sufficient attention to making clear what the values represented by our performance measures really mean. Even concepts as simple as the bias between an observed and a simulated time series need to be put into proper context: whereas a 10% bias in simulation of simulated discharge may be unacceptable in a climate change impact assessment, it may be of less concern in the context of real-time flood forecasting. While some performance measures can have an absolute meaning, such as the common measure of linear correlation, the vast majority of performance measures, and in particular quadratic-error-based measures, can only be properly interpreted when viewed in the context of a reference value.

For hydrologists, the Nash–Sutcliffe efficiency measure (Nash and Sutcliffe, 1970) (Equation (1)) has become a common part of our everyday jargon when reporting the results of a catchment modelling study. Somehow, we seem to ignore the fact that modellers in other fields of environmental sciences are not often familiar with what a 'Nash value' is. More important, it is worth asking whether we ourselves know what it means when we report that a model has a 'Nash value of 0·87'. The reality is that the Nash efficiency value, while a convenient and normalized (–inf to 1·0) measure of model performance does not provide a reliable basis for comparing the results of different case studies. In stating this, we are not pointing out anything not already well known (e.g. Martinec and Rango, 1989; Legates and McCabe, 1999; Seibert, 2001), but we think that it is worth recalling, with the hope that the following discussion will encourage hydrologists to use performance measures in a more useful manner–which is to always provide appropriate reference values–so that reported Nash values can be properly interpreted.

Nash Values Need a Baseline

The Nash–Sutcliffe performance measure (Nash and Sutcliffe, 1970), called, hereafter, the Nash–Sutcliffe efficiency (NSE), is computed as follows:

$$NSE = 1 - \frac{\sum_{t=1}^{N}[q_{obs}(t) - q_{sim}(t)]^2}{\sum_{t=1}^{N}[q_{obs}(t) - \overline{q}_{obs}]^2} \quad (1)$$

where $q_{obs}(t)$ is the observed discharge at time step t, $q_{sim}(t)$ the simulated discharge, $\overline{q}_{obs}$ the mean observed discharge over the entire simulation period of length N.

The NSE is a normalized measure (–inf to 1·0) that compares the mean square error generated by a particular model simulation to the variance of the target output sequence. In doing so, it represents a form of noise-to-signal ratio, comparing the average 'size' (variability) of model residuals to the 'size' (variability) of the target output. It is implicitly comparing the performance of the particular model to that of perhaps the simplest imaginable model, one that uses as its prediction the (constant) mean value of the observed target. This means that an NSE value = 1·0 indicates perfect model performance (the model perfectly simulates the target output), an NSE value = 0 indicates that the model is, on average, performing only as good as the use of the mean target value as prediction, and an NSE value <0·0 indicates an altogether questionable choice of model. We, therefore, prefer NSE values to be larger than 0·0 and approaching 1·0. This corresponds, however, to an apparent normalization because the implicit reference model has different implications for different case studies.

The NSE does not measure how good a model is in absolute terms. Depending on the case study, the reference model hidden in the NSE value poses completely different constraints on the actual model performance. The use of the mean observed value as a reference can be a very poor predictor (e.g. for strongly seasonal time series), or a relatively good predictor (e.g. for time series that are essentially fluctuations around a relatively constant mean value). Schaefli *et al.* (2005) showed an interesting example. In case studies involving high mountainous catchments having a strong annual discharge cycle, they obtained surprisingly high NSE values (higher than 0·9) just by a simple initial screening (through random generation) of seven model parameters. However, they showed that an extremely simple model corresponding just to the use of the mean observed discharge for each calendar day yields an already high NSE of 0·85.

In the case of strongly seasonal time series, a model that only explains the seasonality but fails to reproduce any smaller time scale fluctuations will report a good NSE value; for predictions at the daily time step, this (high) value will be misleading. In contrast, if the model is intended to simulate the fluctuations around a relatively constant mean value, it can only achieve high NSE values if it explains the small time-scale fluctuations. Clearly, therefore, the definition of an appropriate benchmark model is particularly important when we compare model performance over a variety of hydrologic regimes. This is particularly important in the context of model regionalization studies conducted over widely differing types of eco-hydro-climatic response (e.g. Parajka *et al.*, 2005).

To properly communicate how good a model really is, it seems necessary to establish appropriate reference or benchmark models–models having an easy-to-apprehend explanatory power for a given case study and a given modelling time step (see also Seibert, 2001). For hydrologic case studies, it may be difficult–if not impossible–to establish a general and widely applicable benchmark model (such as an autoregressive process for meteorological time series, see, e.g. Hasselmann, 1976). However, it may be possible to at least decide on benchmark models that 'speak' to the modellers or the end-users, i.e. benchmark models that impose performance constraints that are readily interpretable for a given context. In the following, we will give two illustrative examples for rainfall-runoff models at a daily time step.

Establishing Benchmark Models

For observed time series showing a strong but relatively constant seasonality (for example, related to the climate), a simple benchmark model is the one already mentioned: the interannual mean value for every calendar day. This benchmark model has already been proposed by Garrick *et al.* (1978) and is recommended by the WMO (1986) in their snowmelt model intercomparison report. For the three case studies presented in Schaefli *et al.* (2005), such a benchmark model immediately reveals that the hydrologic model is performing much better for one of the catchments (Table I): for the Lonza River (Figure 1), the NSE of the calibrated hydrologic model is 0·2 or 28% higher than the NSE of the benchmark model. For the other two case studies, the performance improvement of the hydrologic model over the benchmark model is only 13% (Rhone River) and 7% (Drance River). Using such a calendar day benchmark model is equivalent to computing the NSE of pre-treated simulated and observed series from which the seasonality has been removed, which is a standard procedure in time series analysis.

The performance improvement of the hydrologic model over the benchmark model can be measured by defining a normalized benchmark efficiency (BE) defined, in analogy to the NSE, as follows:

$$BE = 1 - \frac{\sum_{t=1}^{N}[q_{obs}(t) - q_{sim}(t)]^2}{\sum_{t=1}^{N}[q_{obs}(t) - q_b(t)]^2} \quad (2)$$

DOI: 10.1002/hyp

Table I. Efficiency measures for the case studies of Schaefli *et al.* (2005) using their conceptual model GSM-SOCONT. For the adjusted smoothed precipitation benchmark (ASPB) model, the precipitation is replaced by the equivalent precipitation (Peq) that equals rainfall plus snow- and ice-melt (NSE = Nash–Sutcliffe efficiency, NSEB = Nash–Sutcliffe efficiency of the benchmark model, BE = benchmark efficiency, win = optimum moving-average window size in days)

	Hydrol. model	Benchmark calendar day		Benchmark ASPB with Peq			
	NSE	NSEB	BE	NSEB	BE	lag	win
Rhone	0·94	0·83	0·55	0·82	0·64	1	13
Lonza	0·92	0·72	0·62	0·78	0·62	1	26
Drance	0·90	0·84	0·22	0·83	0·37	1	7

where $q_b(t)$ is the benchmark model discharge at time step t. Such a calendar day benchmark model will establish whether the hydrologic model has greater explanatory power than already contained in the seasonality of the driving forces (the climate).

In a similar vein, we can construct benchmark models that measure whether the hydrologic model has more explanatory power than already contained in the frequency content of the dominant driving process, i.e. in the rainfall. Recall that the 'function' of a catchment is (by a process of storage and time-delayed release) to transform the variability of the driving signal (the rainfall) into an output response (the streamflow) that has reduced amplitude and variability, and is diffused over time. Therefore a considerable part of the variability in the output comes from the driving signal, and our interest is in evaluating the ability of the model to correctly replicate the transforming function of the catchment.

It is common for catchment modellers to show hydrograph time series plots in which the model simulation 'goes up—and down' in a manner similar to that of the measured catchment hydrograph, as an indication of modelling success. Clearly, however, the vast majority of the 'up—and down' hydrologic model response is caused by the driving variables, and what we need to measure is how well the catchment process modification of this behaviour has been reproduced. A very simple benchmark model, therefore, would be to simply scale the rainfall to match the mean discharge (analogous to a Φ-index operation, see, Chow *et al.*, 1988) and to shift the sequence in time by some optimum lag that reflects the time of concentration of the basin. The idea is that this benchmark 'model' projects the frequency variation of the driving variables into the output while having the correct runoff ratio. This adjusted precipitation benchmark (APB) model is constructed as follows:

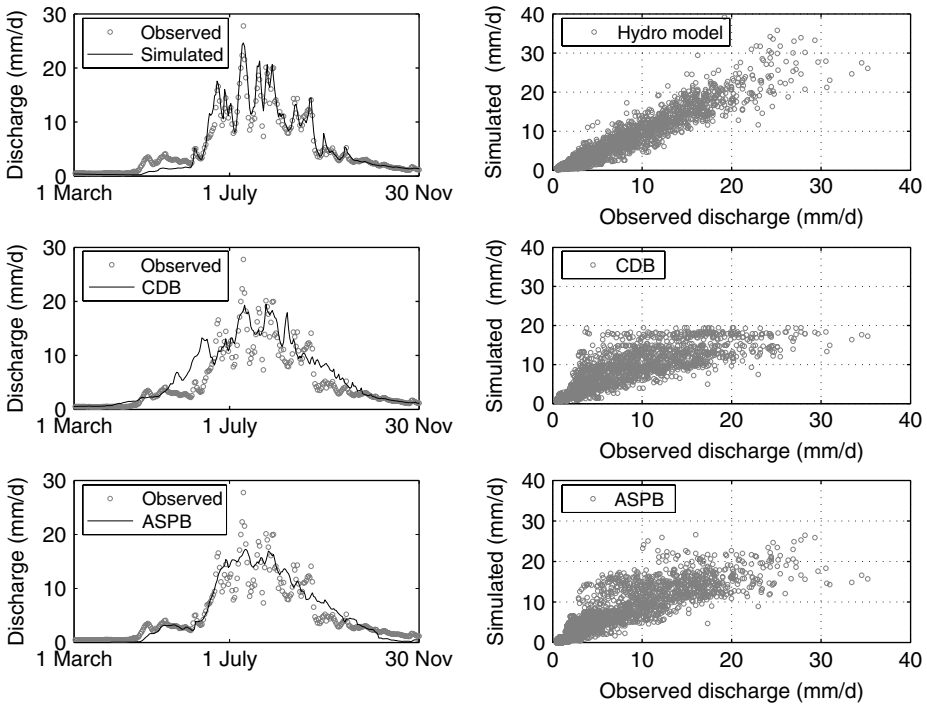

Figure 1. Lonza River case study from Schaefli *et al.* (2005); left: observed, simulated and benchmark discharge time series (last year of calibration period); right: scatter plots of simulated or benchmark discharge values versus observed discharge (for the same year); top: discharge simulated with hydrologic model; centre: calendar day benchmark model (CDB); bottom: adjusted smoothed precipitation benchmark model (ASPB); the efficiency measures are given in Table I

Hydrol. Process. (in press)
DOI: 10.1002/hyp

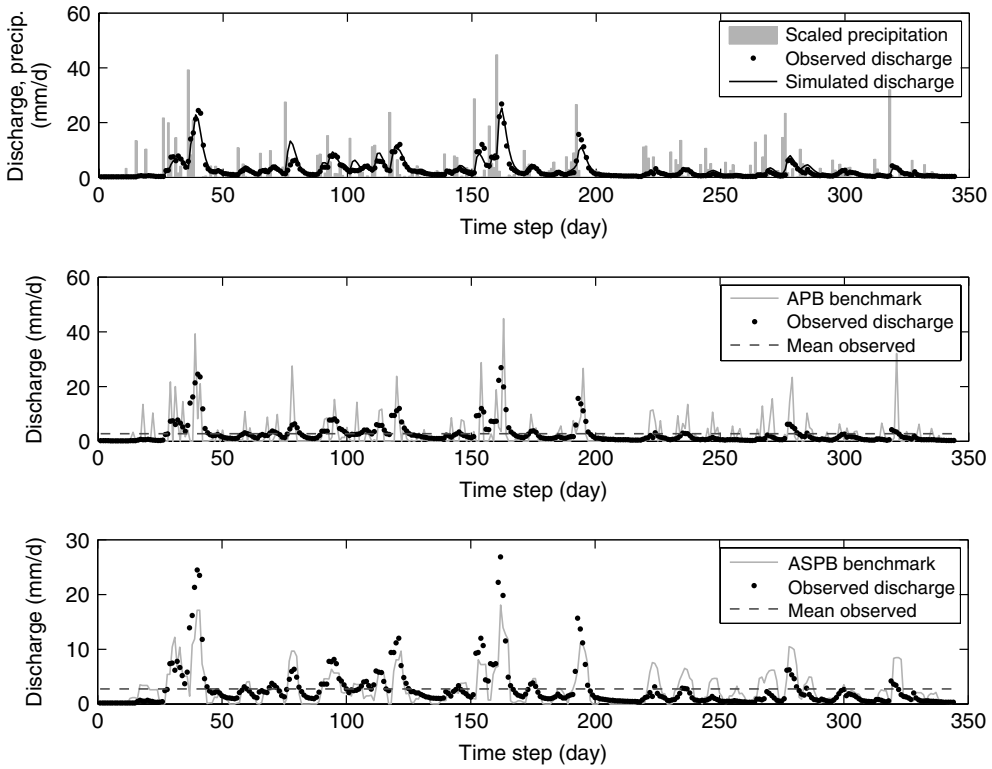

Figure 2. Rainfall adjusted benchmark models for the Leaf River case study (year 1948); top: observed and simulated discharge and scaled precipitation (precipitation multiplied by runoff ratio); centre: Adjusted Precipitation Benchmark (APB) simulation; bottom: adjusted smoothed precipitation benchmark (ASPB) simulation (note the scale difference). The benchmark efficiencies and the benchmark model parameters are given in Table II

1. **Take the precipitation $p(t)$ and multiply by the runoff ratio r (mean discharge/mean precipitation) to make the adjusted precipitation volume equal to the observed discharge volume; this is called adjusted precipitation.**
2. **Shift the adjusted precipitation to the right by an optimum lag (lag_{opt}) which minimizes the mean squared error-of-fit to observed discharge.**
3. **Hence the APB model becomes**

$$q_{APB}(t) = r \cdot p(t - lag_{opt}) \tag{3}$$

We illustrate the use of this benchmark model (Figure 2 top and centre) for a case study involving daily simulation of the rainfall-evapotranspiration-streamflow behaviour of the Leaf River Basin, near Collins, Mississippi, using the simple 5-parameter model HyMod (Boyle, 2000) calibrated with the algorithm presented in Vrugt *et al.* (2003). Clearly, the APB model prediction gives a more realistic reference time series than just the mean discharge: it fluctuates in a way that reflects the behaviour of the basin averaged precipitation time series. Note that the NSE value of the APB model prediction (called NSEB) *is negative* (Table II, year 1948), incorrectly indicating that the mean discharge is a better predictor–arguably this inference using the NSE measure is not reasonable. However, the benchmark efficiency (BE) measure (BE = 0·93) indicates clearly that the 5-parameter HyMod model provides a significant improvement over the 2-parameter (runoff ratio and lag) APB benchmark model.

A further important characteristic of catchments is to filter (smooth) the rainfall to remove higher frequency variability. We can, therefore, further add a simple dispersion process (a moving average) to adjust the smoothness of the scaled-down and translated precipitation to match the smoothness of the observed discharge. One simple way to choose the degree of smoothness (the size of the moving-average window) is so as to maximize the correlation between the adjusted precipitation and the observed flow (Morin *et al.*, 2002). Figure 2 (bottom) shows the performance of this benchmark model (which we refer to as the adjusted smoothed precipitation benchmark (ASPB) model for the same Leaf River Basin study. As expected, the ASPB benchmark model provides an improved baseline simulation of the observed discharge. In fact, the NSEB value of this model is 0·54 and the new BE with respect to the ASPB model is 0·72, indicating that the 5-parameter hydrologic model represents a significant improvement over the simple 3-parameter (runoff ratio, lag and window size) benchmark model. Table II shows the Nash efficiencies as well as the BEs for the Leaf River simulated for a different year (1978). Considering only the NSE, we would come to the conclusion that the model performs equally well as for the calibration period;

Hydrol. Process. (in press)
DOI: 10.1002/hyp

Table II. Efficiency measures for the Leaf River case study simulated with the HyMod model and the Swiss pre-alpine river Rietholzbach (see Gurtz *et al*., 2003) simulated with the PREVAH model (win = optimum moving-average window size in days, calib = calibration period)

		Hydrol. model	Benchmark APB			Benchmark ASPB			
	Period	NSE	NSEB	BE	lag	NSEB	BE	lag	win
Leaf River	1948 (calib)	0·87	−0·90	0·93	3	0·54	0·72	2	4
Leaf River	1978	0·87	−0·31	0·90	3	0·71	0·56	1	5
Rietholzbach	1987–2000	0·87	0·01	0·87	0	0·16	0·85	1	7

the BEs, however, indicate that the model performs considerably worse for the year 1978 than for the calibration period (1948).

The ASPB, therefore, yields a stringent benchmark model as it corresponds to the simplest hydrologic filter for rainfall driven catchments: two linear processes plus a time translation. The BE, with respect to this model, enables us to compare different case studies and to judge how good the calibrated model for the given case study really is. To illustrate this, we applied the ASPB also to a rainfall-driven catchment from the Swiss pre-alpine region simulated with the PREVAH model (Gurtz *et al*., 1999). The case study has a NSE of 0·87. How good is this performance compared to the Leaf River example? The ASPB benchmark efficiency is 0·85; this indicates that the model does a better job than HyMod for the Leaf River, i.e. it explains comparatively more variation of the discharge time series than obtained by a simple linear filtering of the precipitation time series.

We also computed the ASPB for the case studies of Schaefli *et al*. (2005) but for these high mountainous catchments we replaced the precipitation in the ASPB by the so-called equivalent precipitation that equals the precipitation plus simulated snow- and ice-melt. This equivalent precipitation ASPB benchmark model has four parameters (lag, window size, snowmelt factor and ice melt factor (Schaefli *et al*., 2005). The results obtained with the calendar day benchmark model are confirmed (Table I and Figure 1): the hydrologic model has a much poorer performance for the Drance River. But for at least two of the case studies (the Rhone and the Lonza River, see Figure 1) we can conclude that the equivalent precipitation–runoff transformation completed by the hydrologic model using five additional parameters represents a substantial improvement over the benchmark model.

Conclusion

The purpose of this paper is to argue that the definition of an appropriate baseline for model performance, and in particular, for measures such as the NSE values, should become part of the 'best practices' in hydrologic modelling. Every modelling study should explain and justify the choice of benchmark. Of course, the appropriate benchmark model will necessarily be different for different types of case studies. However, for efficient communication, the benchmark should fulfill the basic requirement that every hydrologist can immediately understand its explanatory power for the given case study and, therefore, appreciate how much better the actual hydrologic model is. We encourage further research aimed at establishing a comprehensive set of benchmark models that could also be of further use as a Null hypothesis for hydrologic significance testing, and invite further open dialogue on this topic.

Acknowledgement

We would like to thank Massimiliano Zappa from the Swiss Federal Institute for Forest, Snow and Landscape Research for the data of the Rietholzbach case study.

References

Boyle DP. 2000. *Multicriteria calibration of hydrological models*. PhD Dissertation, Department of Hydrology and Water Resources, The University of Arizona, Tucson.

Chow V-T, Maidment DR, Mays LW. 1988. *Applied Hydrology*. McGraw-Hill: New York, 572.

Garrick M, Cumane C, Nash JE. 1978. A criterion of efficiency for rainfall-runoff models. *Journal of Hydrology* 36: 375–381.

Gurtz J, Baltensweiler A, Lang H. 1999. Spatially distributed hydrotope-based modelling of evapotranspiration and runoff in mountainous basins. *Hydrological Processes* 13: 2751–2768.

Gurtz J, Zappa M, Jasper K, Lang H, Verbunt M, Badoux A, Vitvar T. 2003. A comparative study in modelling runoff and its components in two mountainous catchments. *Hydrological Processes* 17: 297–311.

Hasselmann K. 1976. Stochastic climate models. Part I: theory. *Tellus* 28: 473–485.

Legates DR, McCabe GJ Jr. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35: 233–241, DOI: 10·1029/1998WR900018.

Martinec J, Rango A. 1989. Merits of statistical criteria for the performance of hydrological models. *Water Resources Bulletin* 252: 421–432.

Morin E, Georgakakos KP, Shamir U, Garti R, Enzel Y. 2002. Objective, observations-based, automatic estimation of the catchment response timescale. *Water Resources Research* 38: 1212, DOI: 10·1029/2001WR000808.

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models. Part I: a discussion of principles. *Journal of Hydrology* 10: 282–290.

Hydrol. Process. (in press)
DOI: 10.1002/hyp

Pappenberger F, Beven KJ. 2006. Ignorance is bliss: or seven reasons not to use uncertainty analysis. *Water Resources Research* 42: W05 302, DOI: 10·1029/2005WR004820.

Parajka J, Merz R, Bloschl G. 2005. A comparison of regionalisation methods for catchment model parameters. *Hydrology and Earth System Sciences* 9: 157–171.

Schaefli B, Hingray B, Niggli M, Musy A. 2005. A conceptual glacio-hydrological model for high mountainous catchments. *Hydrology and Earth System Sciences* 9: 95–109.

Seibert J. 2001. On the need for benchmarks in hydrological modelling. *Hydrological Processes* 15: 1063–1064. DOI: 10.1002/hyp.446.

Vrugt JA, Gupta HV, Bouten W, Sorooshian S. 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic models. *Water Resources Research* 39: 1201, DOI:10·1029/2002WR001642.

World Meteorological Organisation. 1986. *Intercomparison of models of snowmelt runoff*, Operational Hydrology Report No. 23. Secretariat of the World Meteorological Organization, Geneva, Switzerland.