HYDROLOGICAL PROCESSES
Hydrol. Process. **15**, 1063–1064 (2001)
DOI: 10.1002/hyp.446

INVITED COMMENTARY

HP TODAY

On the need for benchmarks in hydrological modelling

Jan Seibert

Department of Forest Engineering, Oregon State University, Corvallis, OR 97331, USA

Correspondence to:
Jan Seibert, Swedish University of Agricultural Sciences, Department of Environmental Assessment, Box 7050, SE-750 07 Uppsala, Sweden.
E-mail: jan.seibert@ma.slu.se

WHAT IS 'ACCEPTABLE ACCURACY'?

Hydrological models are applied frequently to scientific or practical problems. For many applications it is concluded that the model has been able to reproduce the measurements with 'acceptable accuracy'. The question is what we mean by this term; the meaning of 'acceptable accuracy' can be quite subjective. We might compute statistical goodness-of-fit measures such as model efficiency, but even the use of such a measure does not necessarily allow an objective judgment of model performance. Does an efficiency of 0.8 for the runoff simulations indicate good or poor model performance? The answer depends on whom you ask. But it also depends on what could be achieved given the specific catchment and the observed data. What might be a poor fit for a watershed with excellent measurements might rightly be considered to be good in a watershed where the available data are of poor quality. To truly assess model performance, it is important to compare one's results with results obtained in some other way, i.e. to choose an appropriate benchmark series.

BENCHMARK SERIES

Using benchmark series, Q_{bench}, we can compute the goodness-of-fit with respect to the benchmark, G_{bench}, using Equation (1), where $Q_{\text{obs}}(t)$ and $Q_{\text{sim}}(t)$ are the observed and simulated runoffs respectively at time step t. G_{bench} is negative if the model performance is poorer than the benchmark, zero if the model performs as well as the benchmark, and positive if the model is superior, with a highest value of one for a perfect fit.

$$G_{\text{bench}} = 1 - \frac{\sum_t (Q_{\text{obs}}(t) - Q_{\text{sim}}(t))^2}{\sum_t (Q_{\text{obs}}(t) - Q_{\text{bench}}(t))^2} \quad (1)$$

This equation is a more general formulation of the model efficiency, where we compare model errors with the errors of the simple method of using the mean observed runoff as (constant) prediction, i.e. using the mean observed runoff as a benchmark. The choice of the benchmark does not influence the calibration of a model, since, in any case, the sum of squared errors is minimized. However, similar formulations can be used if we want to evaluate the goodness-of-fit by another objective function.

Received 29 November 2000
Accepted 4 December 2000

Obviously there are more rigorous benchmarks that could be used instead of the prediction of a constant mean runoff. For instance, we might use the long-term seasonal variation instead of one constant value. We can also use the observed runoff shifted backwards by one or more time steps. In this case, we use the observed runoff at time step t as a prediction of the runoff at time step $t + n$. This type of benchmark is especially suitable for forecast models.

The goodness-of-fit of model simulations for runoff can also be compared with the results of another model or a simple method (e.g. rational method). Simple, lumped models with about four to six parameters are often capable of explaining a large part of runoff variability. The runoff series simulated by such a model provides a valuable benchmark because it accounts for the difficulty in simulating certain watersheds. Data quality, for instance, affects both the tested model and the benchmark model.

In the case where a model with parameter values that have been calibrated for one catchment is applied to another (sub)catchment, results often will be 'acceptable'. But again, to be able to draw any conclusion about the worth of a model we need a benchmark. Both simulated and observed runoff series from the catchment that has been used for calibration provide suitable benchmarks. In both cases the runoff must, of course, be scaled for differences in catchment area. Comparison of a simulation with the first benchmark will reveal the value of taking catchment-specific characteristics, such as the percentage or pattern of different land-use classes, into account. If the model fails to be better than the first benchmark, the implementation of catchment-specific characteristics has to be reconsidered. The second benchmark evaluates the usefulness of a model approach compared with a very simple method to obtain runoff series at ungauged locations. The benefit of this benchmark is that it accounts for observed variation between (sub)catchments. It is useful in discerning whether the results are good because the model is good (and takes differences between the catchments into account), or because runoff dynamics are similar. A comparable approach can be used to assess the usefulness of regionalized parameter values. A suitable benchmark is, for instance, the mean specific runoff of the watersheds on which the regionalization has been based.

CONCLUDING REMARKS

When a model does not provide better runoff simulations than some benchmark, it does not necessarily mean that the model is worthless. The tested model may simulate internal variables (appropriate benchmarks should be used to judge such simulations) or provide opportunities not available otherwise (e.g. simulation of different land-use scenarios). However, with low values of G_{bench} the need to demonstrate the worth of a model increases.

The bottom line is that appropriate benchmarks should be used more often for evaluating model results and justifying our conclusions concerning model performance. It would be beneficial if the hydrological modelling community could elaborate standards on which benchmarks to use. Obviously there is the risk of discouraging results when a model does not outperform some simpler way to obtain a runoff series. But if we truly wish to assess the worth of models, we must take such risks. Ignorance is no defence.