Journal of Hydrology (NZ) 54 (1): 63-66 2015

NOTE

Goodness of fit indices for discharge forecasts in real time

Earl Bardsley

School of Science, University of Waikato, PO Box 3105, Hamilton. Corresponding address: e.bardsley@waikato.ac.nz

Abstract

Discharge forecasting models are often used as a means to provide regular updates on subsequent river discharges, taking into account the most recent information as it becomes available. It is then of interest to use an index of fit to quantify how well the model was able to predict the actual discharge sequence. With sequential forecasts, two quite different approaches can be used to obtain a goodness of fit index. Firstly, a naïve forecast, such as using present discharge to forecast future discharge, could serve as a reference in a benchmark-based index of fit like the Nash-Sutcliffe efficiency (NSE). Alternatively, a fit value like percentage error or maximum difference might be calculated as a direct comparison of observed and predicted discharges. For the same data it is possible for the former index to indicate no predictive ability beyond a naïve model, while at the same time prediction errors might be small enough for the model to be useful. Both approaches to fit measurement have their place in different contexts. It is suggested that a benchmark-based index like the NSE is best used during model development to ensure that the scientific basis is sufficiently developed to enable better forecasts than a naïve alternative. On the other hand, an index using some measure of direct observed and predicted discharge comparison is more appropriate as part of evaluating a discharge forecasting model for practical use.

Keywords

Nash-Sutcliffe fit, flood forecast evaluation, benchmark model, naïve model.

Introduction

This brief communication is motivated by a discussion at the 2014 New Zealand Hydrological Society conference. The question arose over how best to present a goodness of fit index for a discharge forecasting model that produces regular sequential updates of subsequent discharges during the course of a flood event. Specifically, should the discharge forecasts be compared to the actual discharges that occurred or should the forecast accuracy be measured relative to discharges predicted by a benchmark model? This question will arise in any situation that requires evaluation of model-derived sequential estimates of a time series. In this paper a short overview of the use of benchmarks in goodness of fit indices is followed by an example, with some concluding comments about the relative merits of the two approaches.

Benchmarks in goodness of fit

Hydrological models do not exist in isolation but might be loosely thought of as points along a continuum of complexity. Any proposed new model needs to be shown to be better than some existing or 'obvious' model of greater simplicity, which serves as the benchmark for comparison. The benchmark could be particularly simple in which case it could be referred to as a 'naïve model'. For example, the original version of the Nash-Sutcliffe efficiency (NSE) for goodness of fit utilises the mean of the observed data as the prediction benchmark for comparison with model predictions (Nash and Sutcliffe, 1970). Therefore, a model is said to fail if the degree to which the model predictions match observed data is no better than a naïve model repeatedly making predictions using just the mean of the observed data. There are a number of possible variations of such an index, including use of squared deviations or absolute deviations. However, the choice of deviation type is of no particular relevance with respect to the discussion here. The original NSE fit measure incorporates squared deviations and is used here because it gives weight to the largest discharges, which are of most interest in flood studies (Krause *et al.*, 2005). The NSE expression can be written in a more general form as:

$$NSE = 1 - \frac{\sum (O_i - P_i)^2}{\sum (O_i - B_i)^2} \qquad (1)$$

where the O_i, P_i and B_i are respectively the observations, the model predictions, and the specified benchmark values. An NSE of 1.0 denotes perfect fit to a set of recorded data and NSE < 0 means the benchmark values give better predictions than the model in the sense of having a smaller value of the mean squared deviation.

As an example of using a benchmark other than the mean of the observed values, a model seeking to forecast monthly rainfalls in a seasonal climate might predict high rainfalls for wet season months and low rainfalls for dry season months, yielding an apparent good fit to monthly data. However, the seasonal hydrology means that 'month of year' is itself a predictive model and the respective monthly means are therefore the more appropriate benchmark. The original good fit might then be seen to disappear when the new fit value is calculated using the new benchmarks.

The question of suitable benchmarks for different hydrological models has been discussed from time to time in the literature. Legates and McCabe (1999) mention the seasonality effect and other time-varying benchmark possibilities. Seibert (2001) notes using recorded discharge as one benchmark for application to model predictions for sequential river flow forecasting. Schaefli and Gupta (2007) emphasise the importance of benchmarks for model evaluations generally. NSE benchmarks for the specific case of evaluating flood forecasting models are discussed by Moussa (2010) and a selection of naïve models for evaluating flood forecasting models are considered by Dawson *et al.* (2012).

Flood hydrograph example

Figure 1 shows the time series for a flood event on the Leith River (Dunedin) together with predicted discharges obtained from a forecasting model. The model in this case happens to be a linear function of recent rainfalls and does not include current discharge as part of the model input (Mohssen, 2014). Forecasts were made at hourly intervals for discharge four hours into the future. A number of models of similar structure were utilised in that study but with different choices of prior rainfalls as predictor variables. The example considered here used previous rainfalls from four hours lag through to eight hours lag. Results from one of the other models are displayed in Mohssen (2014).

Figure 2 shows the data as a scatter of observed and predicted values, with the model evidently suffering from the under-prediction seen in Figure 1. Some fit measure based on direct comparisons of observed and predicted discharges could be presented at this point. For example, if the mean of all the positive deviations is divided by the peak discharge then a dimensionless value of 0.11 is obtained, with 0.06 being the corresponding value for the negative deviations. A value of 0.0 for both deviations would represent a perfect match.

Visual inspection suggests that a roughly equivalent degree of fit might achieved by a naïve model using just present discharge to predict the discharge four hours subsequent (Fig. 3). This suggests the alternative of a benchmark-based index, using discharge four hours previously as the benchmark values in Equation 1. It happens that this results in an NSE value of -0.44, so this particular discharge forecasting model fails the benchmark test because present discharge will on average give a more accurate prediction of river discharge four hours on. That is, incorporating recent rainfalls into the predictive model adds no forecasting skill beyond using just present discharge as the forecaster. The negative NSE value here arises because present discharge is a reasonable estimator of future discharges that are not too far into the future. In contrast, the NSE would improve to an unrealistically high value of 0.71 if the mean of the recorded flows was used as the benchmark. This is because the mean is a much poorer estimate of future discharges due to the 'forecasts' being just repeats of the same constant value.

The Leith River predictive model was under development at the time of writing and there is no implication that the general modelling approach involved is deficient in any way. Inevitably, present discharge will have less predictive value further into the future and forecasting models then are better able to demonstrate their worth. In particular, present discharge will increasingly under-predict peak discharges as forecasting time extends. Also, the focus here has just

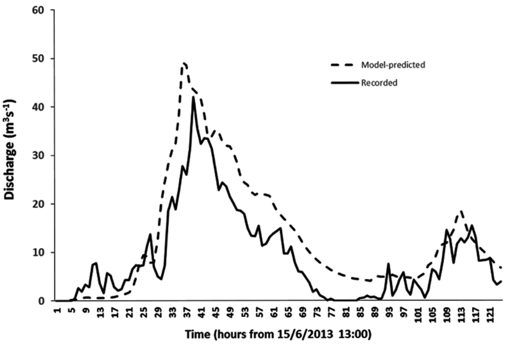

Figure 1 – Hourly observed and model-forecast discharges during a flood event on the Leith River, Dunedin.

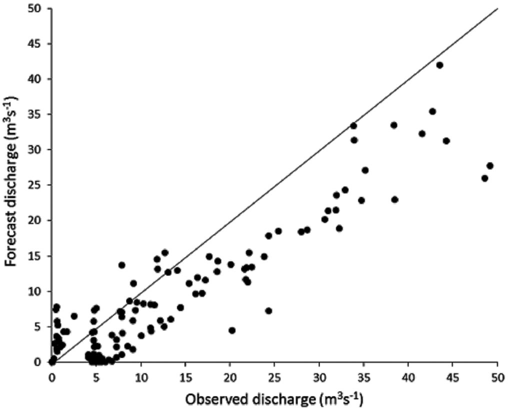

Figure 2 – The data of Fig. 1 plotted as an observed/forecast scatter. Solid line is the 1:1 relation.

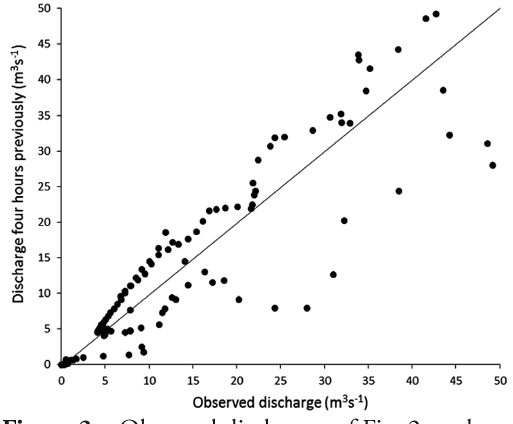

Figure 3 – Observed discharges of Fig. 2, and discharges four hours previously. Solid line is the 1:1 relation.

been on methods of comparing observed and forecast discharge magnitudes and even a minor change in model timing can sometimes convert a low fit value into a much higher one when magnitude comparisons are involved (Moussa, 2010). The point is also made that comparisons of a sequence of observed and predicted discharges is not the same as evaluating a flood forecasting model, where the critical aspects for evaluation are how well the peak flow is anticipated and how much advance warning can be given.

Conclusion

Returning to the original question – how best to present a goodness of fit index for a discharge forecasting model – it is evident that both representations of flow forecasting goodness of fit values (i.e., benchmark-based and indices of direct comparison of observed and predicted discharges) can have their place in different contexts. When developing a flow forecasting model, it is desirable to have a naïve estimator as a goal to surpass so as to avoid proposing a new model that actually gives worse forecasts than the naïve model. On the other hand, an accurate final model will yield a misleadingly poor fit value from a benchmark-based index like NSE if the naïve model happens to give even better estimates. In practice, any users of forecasting models are concerned with accuracy and not with referencing to benchmarks. Therefore, conditional on first passing a benchmark comparison, it would seem appropriate that a completed flow forecasting model should be finally evaluated for practical use by reference to some index of direct comparison like percentage error, maximum difference, or mean difference.

Acknowledgement

I am grateful to Magdy Mohssen of Lincoln University for an interesting discussion on this topic and for making available the discharge and prediction data of Fig. 1.

References

Dawson, C.W.; Mount, N.J.; Abrahart, R.J.; Shamseldin, A.Y. 2012: Ideal point error for model assessment in data-driven river flow forecasting. *Hydrology and Earth System Sciences 16*: 3049-3060.

Krause, P.; Boyle, D.P.; Bäse, F. 2005: Comparison of different efficiency criteria for hydrological assessment. *Advances in Geosciences 5*: 89-97.

Legates, D.R.; McCabe G.J. 1999: Evaluating the use of 'goodness of fit' measures in hydrologic and hydroclimatic model validation. *Water Resources Research 35*: 233-241.

Mohssen, M. 2014: Flood forecasting of river flows. *Water Symposium 2014 Abstracts* p.189, Blenheim, Nov. 24-28.

Moussa, R. 2010: When monstrosity can be beautiful while normality can be ugly: assessing the performance of event-based flood models. *Hydrological Sciences Journal, 55*: 1074-1084.

Nash, J.E.; Sutcliffe, J.V. 1970: River flow forecasting through conceptual models. Part I: a discussion of principles. *Journal of Hydrology 10*: 282-290.

Schaefli, B.; Gupta, H.V. 2007: Do Nash values have value? *Hydrological Processes 21*: 2075-2080.

Seibert, J. 2001: On the need for benchmarks in hydrological modelling. *Hydrological Processes 15*: 1063-1064.

Manuscript received 22 December 2014; accepted for publication 7 May 2015