Absolute and relative measures for evaluating the forecasting performance of time series models for daily streamflows

T. Astatkie

Department of Engineering, Nova Scotia Agricultural College, PO Box 550, Truro, NS, B2N 5E3, Canada
E-mail: *tastatkie@nsac.ca*

Received 27 October 2004; accepted in revised form 23 November 2005

Abstract Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are widely used measures for evaluating the forecasting performance of time series models. Although these absolute measures can be used to compare the performance of competing models, one needs a reference to judge the goodness of the forecasts. In this paper, two relative measures, coefficient of efficiency (*E*) and index of agreement (*d*), and their modified versions (*EM*, *EMP*, *dM* and *dMP*) with desired values of closer to one are presented. These measures are illustrated by comparing the modeling ability and validation forecasting performance of a Nonlinear Additive Autoregressive with Exogenous variables (NAARX), Nested Threshold Autoregressive (NeTAR), and Multiple Nonlinear Inputs Transfer Function (MNITF) models developed for the Jökulsá eystri daily streamflow data. The results suggest that NeTAR describes the system best, and gives better 1- and 2-day ahead validation forecasts. MNITF gives better forecasts for 3-day ahead, and NeTAR and NAARX give comparable performance for 4- and 5-day ahead forecasting. The values of *E* and *d* were larger than those of the modified versions, giving a false sense of model performance, and unlike the modified versions, they decreased as forecast lead times increased. Differences among the values of these six relative measures can reveal the sensitiveness of competing models to outliers, and their potential for long-term forecasting. Accordingly, NeTAR was the least sensitive to outliers and NAARX was the most sensitive, with MNITF in between; and NAARX showed the most potential for long-term streamflow forecasting.

Keywords Coefficient of efficiency; index of agreement; model validation; nonlinear time series

Introduction

Time series models are used for describing a dynamic system, or for short- and/or long-term forecasting, or for simulation. In many areas of application, since the dynamic systems that generate the time series are inherently nonlinear, nonlinear time series models usually do a better job compared to linear models. The book by Tong (1990) provides a comprehensive review of several classes of nonlinear time series models. To choose between competing models for describing a given dynamic system, despite it being an absolute measure (not unit-free), Root Mean Squared Error (RMSE) is commonly used as a performance measure (Armstrong and Collopy 1992). Mean Absolute Error (MAE) is another commonly used absolute measure to compare different forecasting models. Although RMSE and MAE are widely used, their being absolute measures is a major shortcoming because their calculated values depend on the unit of measurement and scale of the series. In the absence of a reasonable reference to compare their values with, it becomes difficult to tell whether the forecasts produced by competing models are acceptably close to the true values. Therefore, as Legates and McCabe (1999) argued, one should consider relative measures as well to compare the forecasting performance of competing time series models.

One such relative performance measure is the Coefficient of Efficiency (*E*). This measure, originally proposed by Nash and Sutcliffe (1970), ranges from minus infinity to one. A value

Nordic Hydrology **Vol 37 No 3 pp 205–215** © IWA Publishing 2006

doi: 10.2166/nh.2006.008

of zero implies that the model is as good as forecasting future values by the average of all values in the series. That is, E uses the average of the series as a reference.

Another relative performance measure, originally proposed by Willmott (1981), is Index of agreement (d). The value of d ranges from zero to one, and values closer to one are desirable. It can be interpreted in a similar way as the coefficient of determination (R^2); however, it should be noted that it is not like R^2 (Legates and McCabe 1999). One potential problem with d is usually that its values are high, and might give a false impression of good performance of the models. This is so because it uses the average of all values in the series as a reference, which could be amiss in some time series.

Using the average of all values as a reference seems unnecessarily primitive in some systems. Instead, using the current value as a reference, especially for short-term forecasting, is better. In this paper, both d and E are modified to dM and EM by using the current value of the series as a reference. The calculation of d and E (as well as dM and EM) involves using squared differences between the observed and predicted values of the series. This can give undue importance to single outlying values. In this paper, dMP and EMP that use absolute differences are proposed to reduce the effect of outliers. It should be noted that, although d and E are used to some extent to compare different, mainly simulation, models in hydrology, they are not common in economics and business.

A second objective of this paper is to develop a Multiple Nonlinear Inputs Transfer Function (MNITF) model for describing the Jökulsá eystri daily streamflow system in Iceland. The nonlinear inputs in a MNITF model are generated according to the underlying nonlinear dynamics. Detailed description and modeling strategy of a MNITF model, as well as its illustration using the Saugeen River (in Canada) daily data, is given in Astatkie and Watt (1998). Although the MNITF model is in the literature, its describing and forecasting ability has never been tried on the Jökulsá eystri benchmark dataset.

The Jökulsá eystri dataset has daily data on streamflow, precipitation and temperature from 1 January 1972 to 31 December 1974 recorded by the Hydrological Survey of the National Energy Authority of Iceland. The meteorological series were recorded at Hveravellir station, which is within the Jökulsá eystri drainage basin. Time series plots of daily streamflow, temperature and precipitation are shown in Figure 1 and a full description of this river is available in Tong (1990) and references cited there.

The daily streamflow, precipitation and temperature data of Jökulsá eystri from 1972–4 have been used by several authors including Tong *et al.* (1985), Tong (1990), Chen and Tsay (1993), Astatkie *et al.* (1997) and Wong and Li (2001). These authors used this benchmark dataset to compare competing time series models. The most recent one, Wong and Li (2001), compared a transfer function with precipitation and temperature inputs, an Open-loop Threshold Autoregressive System (TARSO), a Nonlinear Additive Autoregressive with Exogenous variables (NAARX), a Nested Threshold Autoregressive (NeTAR) and their Logistic Mixture Autoregressive with Exogenous variables (LMARX) models to compare the describing and forecasting abilities of these models. As in Chen and Tsay (1993) and Astatkie *et al.* (1997), they calibrated the models using the 1972–3 data and did validation forecasts for 1974, and concluded that the NeTAR model gives the best result for short-term (1- to 3-days ahead) and NAARX for longer-term (4- and 5-) days ahead forecasts. Following their finding, these two models are compared with MNITF identified for the Jökulsá eystri streamflow system. These three models are compared in terms of the RMSE, MAE, d, E and the modified versions of d and E. These three models are described in the next section.

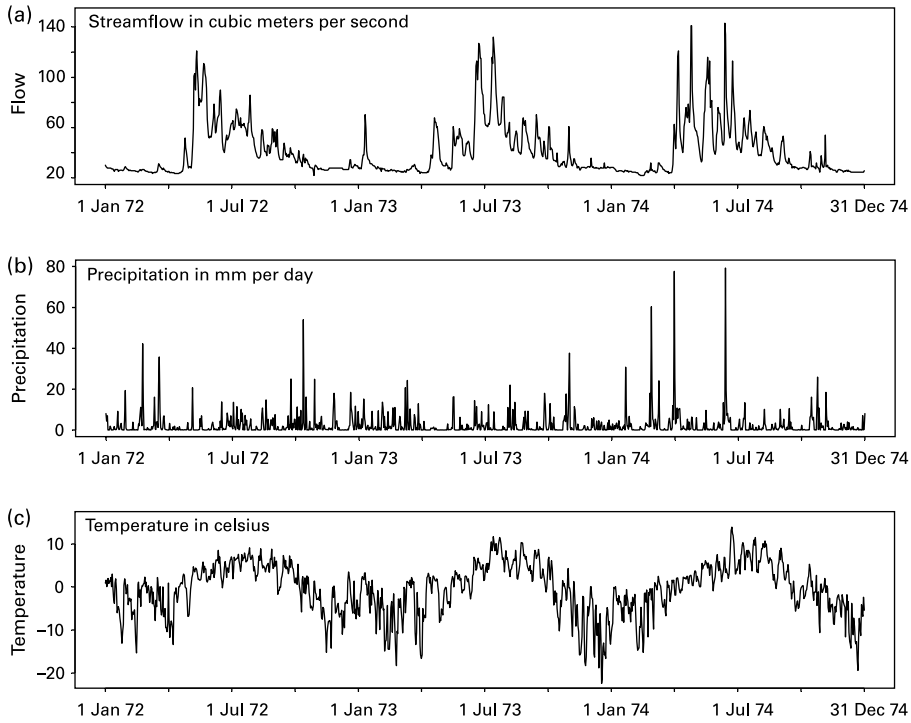

Figure 1 Time plot of (a) daily streamflow, (b) daily precipitation, and (c) daily average temperature of Jökulsá eystri from January 1, 1972–December 31, 1974

NAARX, NeTAR and MNITF models

Chen and Tsay (1993) proposed the NAARX model, which is a simple generalization of the first order nonlinear autoregressive model, to describe the Jökulsá eystri streamflow system. They identified the following model for this river:

$$\begin{aligned} y_t =& \phi_0 + \phi_{1,1} y_{t-1} + \phi_{1,2} y_{t-1} I(y_{t-1} \geq c_1) + \phi_{1,3} y_{t-1} I(y_{t-1} \geq c_2) \\ &+ \phi_2 y_{t-2} + \phi_3 y_{t-3} + \phi_4 y_{t-4} + \beta_1 z_t + \beta_2 z_{t-1} + \omega_{1,1} x_{t-1} \\ &+ \omega_{1,2} x_{t-1} I(x_{t-1} \geq c_3) + \omega_{3,1} x_{t-3} + \omega_{3,2} x_{t-3} I(x_{t-3} \geq c_4) + \varepsilon_t \end{aligned} \quad (1)$$

where y_t is streamflow (in m^3/s) on day t, z_t is precipitation (in millimetres) on day t, x_t is temperature (in °C) on day t, and the structural parameters were estimated to be: $c_1 = 27\,m^3/s$, $c_2 = 100\,m^3/s$, $c_3 = 1$ °C, and $c_4 = 1$ °C. Chen and Tsay (1993) commented that the model identified for the daily streamflow of Jökulsá eystri (Equation (1)) was an NAARX model with simple piecewise linear functionals indicated by the $I(.)$ terms. They also stated that "When the temperature is below 1 °C, streamflow is not significantly affected. On the other hand, when the temperature is higher than 1 °C, the streamflow of the next day increases substantially, presumably due to snow melting. The reverse influence of temperature 3 days earlier on streamflow is understandable, because the effect of snow melting cannot last for long. Of course, one needs to keep in mind the dynamic nature of y_t in interpreting the influence of temperature 3 days earlier. The piecewise function of y_{t-1} indicates that the daily streamflow y_t, after adjusting the effects of precipitation and temperature, has different dynamic properties depending on the level of the previous streamflow y_{t-1}."

The NeTAR model, proposed by Astatkie *et al.* (1997), is a nonlinear time series model useful for describing nonlinear dynamic systems by forming simpler (usually linear) subsystems or zones. The zones are formed in two or more stages using lagged values of the

output and/or the input series. The final NeTAR model for the Jökulsá eystri streamflow system, identified using the procedures described in Astatkie *et al.* (1997), is

$$y_t = \begin{cases} \phi_{10} + \phi_{11} y_{t-1} + \varepsilon_{1t} & (y_{t-2} \le 92, \bar{x}_t \le -2) \\ \phi_{21} y_{t-1} + \phi_{22} y_{t-2} + \beta_{20} z_t + \omega_{20} x_t + \varepsilon_{2t} & (y_{t-2} \le 92, -2 < \bar{x}_t \le 1.8) \\ \phi_{31} y_{t-1} + \phi_{32} y_{t-2} + \beta_{31} z_{t-1}^2 + \omega_{30} x_t & \\ \quad + \omega_{33} x_{t-3} + \varepsilon_{3t} & (y_{t-2} \le 92, \bar{x}_t > 1.8) \\ \phi_{40} + \phi_{41} y_{t-1} + \omega_{40} x_t + \omega_{41} x_{t-1} & \\ \quad + \omega_{43} x_{t-3} + \varepsilon_{4t} & (y_{t-2} > 92) \end{cases} \quad (2)$$

In this threshold model, the threshold values for y_{t-2} (proxy for basin storage) and $\bar{x}_t$ (proxy for basin temperature) were estimated as part of NeTAR modeling to get the four zones: (i) *dry and cold*: $y_{t-2} \le 92\,\mathrm{m}^3/\mathrm{s}$ and $\bar{x} \le -2\,^\circ\mathrm{C}$, (ii) *dry and mild*: $y_{t-2} \le 92\,\mathrm{m}^3/\mathrm{s}$ and $-2 < \bar{x}_t \le 1.8^\circ\mathrm{C}$, (iii) *dry and warm*: $y_{t-2} \le 92\,\mathrm{m}^3/\mathrm{s}$ and $\bar{x}_t > 1.8\,^\circ\mathrm{C}$, (iv) *wet*:$y_{t-2} > 92\,\mathrm{m}^3/\mathrm{s}$, where y_{t-2} is the indicator for basin storage on day t and $\bar{x}_t$is the indicator of basin temperature on day t calculated as $\bar{x}_t = 1/2(x_t + x_{t-1})$. Astatkie *et al.* (1997) commented that:

"The final NeTAR model reveals that temperature plays a dominant role in modulating the dynamics of the Jökulsá eystri. In the low storage condition with below-freezing temperature conditions (zone 1), current flow is low and is influenced by neither temperature nor precipitation. Only the previous day's flow describes it well. Both temperature and precipitation have an effect on current flow when the storage is low and the temperature is above freezing. The lags and amount, however, depend on whether the temperature is slightly above freezing or high. Flow strongly depends on current temperature as well as that of one and three days prior when storage is high (zone 4). One explanation for the heavy dependence on temperature is the presence of a glacier in the basin."

MNITF models, proposed by Astatkie and Watt (1998), account for nonlinearities in the system caused by inputs. For streamflows, this is done by generating the most important inputs (Astatkie *et al.* 1996), snowmelt and effective rain, using nonlinear filters. The model is identified using a vector autoregressive model and the Corner method. Following the procedure detailed in Astatkie and Watt (1998), the MNITF model identified for the Jökulsá eystri streamflow system is

$$y_t = \phi_0 + \frac{(\delta_{10} + \delta_{11}B + \delta_{12}B^2)M_t}{(1 - \omega_{11}B)} + \frac{(\delta_{20} + \delta_{21}B + \delta_{22}B^2)ER_t}{(1 - \omega_{21}B)} + \frac{(1 - \theta_{11}B)e_t}{(1 - \phi_1 B - \phi_2 B^2)} \quad (3)$$

where (i) M_t is snowmelt (in millimetres) on day t obtained as

$$M_t = \begin{cases} C(\bar{x}_t - \bar{x}_0) & \text{if } \bar{x}_t \ge 1,\ x_t > 0,\ SP_{t-1} + S_t > C(\bar{x}_t - \bar{x}_o) > 0 \\ SP_{t-1} + S_t & \text{if } \bar{x}_t \ge 1, x_t > 0,\ C(\bar{x}_t - \bar{x}_0) > SP_{t-1} + S_t \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $C = 2\,\mathrm{mm}/^\circ\mathrm{C}/\mathrm{day}$ is the degree-day factor, $\bar{x}_t$ is basin temperature on day t calculated as the average of the current and the previous day's temperature, $\bar{x}_0 = 1\,^\circ\mathrm{C}$ is the threshold basin temperature for snowmelt to take place, x_t is temperature on day t, S_t is snowfall on day t (since precipitation (z_t) in the data set was not separated into rainfall and snowfall, precipitation is considered as snowfall if the temperature for that day is below zero), and SP_t

is snowpack on day t calculated as

$$SP_t = \begin{cases} SP_{t-1} + S_t - M_t & \text{if } \bar{x}_t > 1,\ SP_{t-1} > 0 \\ SP_{t-1} + S_t & \text{otherwise} \end{cases} \quad (5)$$

The snowmelt model shown in Equation (4) is similar to the degree-day model first discussed by Martinec (1960), and its basis is that the snowmelt rate can be directly related to the degrees of air temperature above a critical temperature for melt to occur. The degree-day model has also been used successfully for estimating the present state of snow depth (see, for example, Schumann and Lauener 2005).

(ii) ER_t is effective rain on day t obtained as

$$ER_t = \begin{cases} z_t & \text{if } \bar{x}_t \geq 1 \text{ and } x_t \geq 3 \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where z_t is precipitation on day t, and $\bar{x}_t$ and x_t are as defined in Equation (4). Note that this input can be improved when data on actual snowfall and potential abstraction are available.

A threshold temperature (basin temperature on day t) of 1 °C, instead of the commonly used threshold of 0 °C, was used to generate snowmelt and effective rain mainly because of the difference between the altitude of the flow station and that of the meteorological station, and the fact that the temperature values are averages. The value of temperature on day t (x_t) is the daily average temperature, and the proxy for basin temperature on day t ($\bar{x}_t$) is the average of the average temperatures on day t and on the previous day. This means that when $\bar{x}_t$ is 1 °C, the actual temperature on day t is somewhat higher than the freezing point, at least for a few hours, leading to possible snowmelt and/or precipitation coming in a form of rain. The coefficient used in Equation (4) indicates that, when the basin temperature is below 1 °C, there will be no snowmelt, but when this temperature is, say 2 °C, snowmelt will be 2 mm provided that snowpack on the previous day is over 2 mm.

(iii) e_t is the noise part, and B is backward shift operator.

Plots of the observed daily flow series together with those fitted by the NAARX, NeTAR and MNITF models are shown in Figure 2.

Evaluation of model performance

To evaluate how the NAARX, NeTAR and MNITF models describe the dynamic system, and forecast future values, the eight measures described below were used. Following Legates and McCabe (1999) who evaluated various performance measures, the n observed values of a time series were denoted by O_t, and the corresponding forecasted (predicted) values were denoted by P_t.

Root mean squared error

Root Mean Squared Error (RMSE) is an extensively used absolute measure for assessing the performance of time series as well as regression models. Its value is given in terms of the unit of the series or the variable being modeled. The RMSE is calculated using

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(O_t - P_t)^2}{n}}. \quad (7)$$

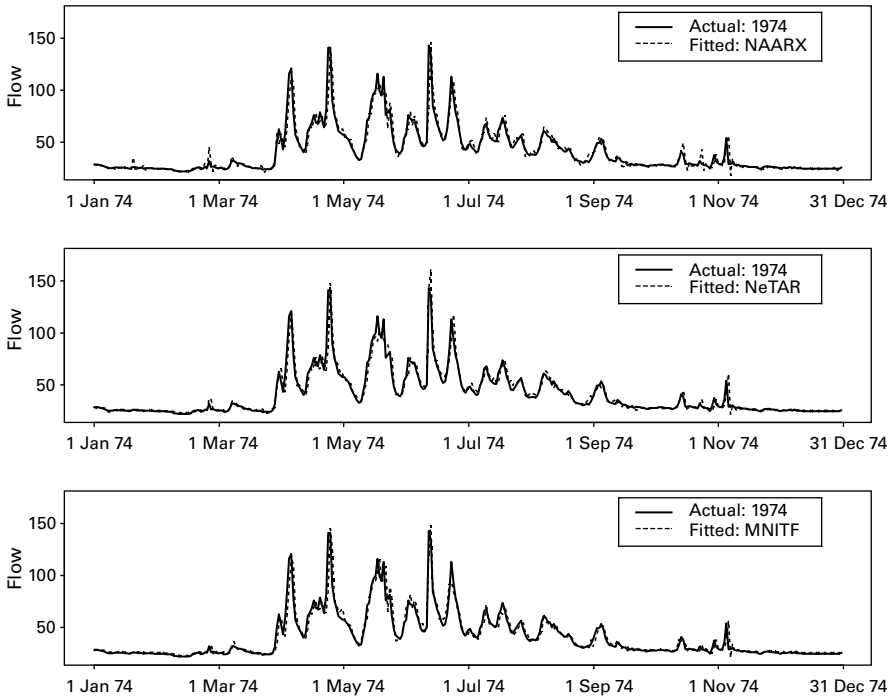

Figure 2 Time plot of the actual (solid) and fitted (broken) flow from the final NAARX, NeTAR and MNITF models for the 1974 data of Jökulsá eystri

Mean absolute error

Mean Absolute Error (MAE) is another widely used absolute measure that takes the unit of the series or the variable. It is given as

$$MAE = \frac{\sum_{t=1}^{n} |O_t - P_t|}{n}. \tag{8}$$

Since the deviations in RMSE are squared, the inflating effect of outliers (extremely large or small values) is more pronounced in RMSE than in MAE. In general, $RMSE \geq MAE$, and the degree to which RMSE exceeds MAE indicates the extent of outliers in the data (Legates and McCabe 1999).

Coefficient of efficiency

Nash and Sutcliffe (1970) proposed an efficiency (E) criterion for objective assessment of streamflow simulation models:

$$E = 1 - \frac{\sum_{t=1}^{n} (O_t - P_t)^2}{\sum_{t=1}^{n} (O_t - \bar{O})^2} \tag{9}$$

where $\bar{O}$ is the mean of the observed values. E ranges from minus infinity to 1, with higher values indicating better performance. If $E > 0$, the model gives better forecasts than forecasting all values by the mean ($\bar{O}$); $E = 0$ means the model forecasts are as good as the mean; and $E < 0$ means that the model is worse than forecasting the values by the mean.

Baseline-modified coefficient of efficiency

Garrick *et al.* (1978) pointed out that even poor streamflow simulation models yield relatively high values for E and noted that, for real-time flood forecasting applications,

the use of the mean discharge as a reference is unnecessarily primitive. Accordingly, Watt and Nozdryn-Plotnicki (1982) proposed a "modified" E, denoted by EM, that uses the present (current) observed streamflow, denoted by O_{tc} instead of the mean. Using the present value as a reference will not require additional resources as it is always available to the forecaster when forecasting future values. Using the present value as a reference is equivalent to forecasting future values by the random walk model (Armstrong and Collopy 1992). EM, denoted by $EM(l)$ for forecast lead l is calculated as

$$EM(l) = 1 - \frac{\sum_{t=1}^{n}(O_{t+l} - P_{t+l})^2}{\sum_{t=1}^{n}(O_{t+l} - O_{tc})^2}. \tag{10}$$

Baseline-modified prime coefficient of efficiency

Garrick *et al.* (1978) as well as Legates and McCabe (1999) argued that squaring the differences increases the sensitivity of the measures to outliers, and hence absolute values of the differences should also be considered. This measure, known as baseline-modified prime coefficient of efficiency (EMP), and denoted by $EMP(l)$ for forecast lead l is calculated as

$$EMP(l) = 1 - \frac{\sum_{t=1}^{n}|O_{t+l} - P_{t+l}|}{\sum_{t=1}^{n}|O_{t+l} - O_{tc}|}. \tag{11}$$

Index of agreement

The index of agreement (d), developed by Willmott (1981), is a relative measure that falls between 0 and 1, with high values indicating better performance. It is given as

$$d = 1 - \frac{\sum_{t=1}^{n}(O_t - P_t)^2}{\sum_{t=1}^{n}(|P_t - \bar{O}| + |O_t - \bar{O}|)^2} \tag{12}$$

where $\bar{O}$ is the mean of the observed values. Although both d and R^2 (coefficient of determination) range from 0 to 1, Legates and McCabe (1999) caution against interpreting their values the same way because "a value of 0.5, for example, has substantially different meanings for R^2 and d." They also suggest that d represents a decided improvement over R^2, but is sensitive to extreme values (outliers).

Baseline-modified index of agreement

Following the same argument as in EM, the baseline-modified index of agreement is denoted by dM, and for forecast lead l by $dM(l)$ is calculated as

$$dM(l) = 1 - \frac{\sum_{t=1}^{n}(O_{t+l} - P_{t+l})^2}{\sum_{t=1}^{n}(|O_{t+l} - O_{tc}| + |P_{t+l} - O_{tc}|)^2}. \tag{13}$$

Baseline-modified prime index of agreement

Like in the baseline-modified prime coefficient of efficiency, the absolute value of differences can be used in the baseline-modified index of agreement to reduce its sensitivity to outliers. This measure is known as the baseline-modified prime index of agreement (dMP). This measure, denoted by $dMP(l)$ for forecast lead l, is calculated as

$$dMP(l) = 1 - \frac{\sum_{t=1}^{n}|O_{t+l} - P_{t+l}|}{\sum_{t=1}^{n}(|O_{t+l} - O_{tc}| + |P_{t+l} - O_{tc}|)}. \tag{14}$$

In this paper, in addition to the commonly used absolute error measures (RMSE and MAE), relative error measures, namely E, EM and EMP are also used to provide comparison

of the errors in a physically meaningful way (i.e. a value of zero implying the model predicts as good as a "reference," namely the observed mean or present value). *d*, *dM* and *dMP* are also used as measures of agreement between the observed and the predicted values.

It should be noted that, although these measures give a good idea about the performance of forecasting models, they do not give different weights for under- and over-predictions. That is, the weight they give to $(O_{t+l} - P_{t+l})$ is the same regardless of the sign (negative for over- and positive for under-prediction) of the difference.

Results and discussion

As was done by Chen and Tsay (1993), Astatkie *et al.* (1997), Wong and Li (2001) and many others who used this benchmark data set, the 1-, 2-, 3-, 4- and 5-day ahead forecasts by the NAARX, NeTAR and MNITF models were computed using the actual values of the inputs (precipitation and temperature). The computed performance measures are presented for the calibration and validation periods. The calibration period covers two years (January 1, 1972–December 31, 1973), and the validation period covers one year (January 1, 1974–December 31, 1974).

As shown in Table 1, NeTAR has the lowest RMSE (5.23 m^3/s). This RMSE, calculated from the model using all 3-year data, is actually the 1-day ahead forecast standard deviation, as the fits are 1-day ahead forecasts. MNITF gave lower RMSE than NAARX. It should be noted that the RMSE for NAARX ($\sqrt{31.28} = 5.59$) computed in this paper is lower than that reported by Chen and Tsay (1993), which was $\sqrt{33.77} = 5.81$. Based on the RMSE, the NeTAR model describes the system best and MNITF comes second.

The RMSE of 1974 validation forecasts, computed using the actual values of the inputs (as used by Chen and Tsay (1993) and Astatkie *et al.* (1997)), suggest that NeTAR performs best for 1-day and 2-day ahead forecasts, MNITF for 3-day ahead forecasts, NeTAR and NAARX for 4-day ahead, and NAARX for 5-day ahead forecasts. Although MNITF gave better results for medium-term (3-day lead) forecasts, its RMSE for a longer run, especially for lead 5, jumped substantially. This could be due to the unreliability of the precipitation values (Tong 1990, p 432) and the unavailability of snowfall and rainfall data separately. The results from RMSE and the coefficient of efficiency (*E*) were strikingly similar because their values are inversely proportional.

The MAE values, also shown in Table 1, give slightly different performance indicator than the RMSE. For 4-day ahead validation forecasts, while RMSE gave comparable performance of NeTAR and NAARX, the MAE identified NAARX as a better performer. Most of the calculated values of MAE are about half of the RMSE values mainly because, as Meade (2000) put it, "MAE does not give undue importance to large errors (as a RMSE would)."

Table 1 RMSE and MAE for 1-day ahead forecasts of the model (using January 1, 1972–December 31, 1974 data), and for 1- to 5-day ahead validation forecasts (January 1, 1974–December 31, 1974) using the model calibrated using January 1, 1972–December 31, 1973 data. The best model for each lead is shown by boldface values

		RMSE (m^3/s)			MAE (m^3/s)		
Data	**Lead (d)**	**NAARX**	**NeTAR**	**MNITF**	**NAARX**	**NeTAR**	**MNITF**
1972–4	1	5.59	**5.23**	5.33	1.17	**1.13**	1.15
1974	1	7.85	**6.99**	7.19	3.52	**3.39**	3.45
1974	2	11.88	**10.78**	11.26	5.68	**5.50**	5.92
1974	3	13.32	12.76	**12.07**	6.61	6.75	**6.60**
1974	4	14.16	**14.14**	14.95	**7.68**	8.00	8.34
1974	5	**14.70**	15.06	18.84	**8.44**	8.95	10.53

The values of *E* as well as its modified versions (*EM* and *EMP*) shown in the first column of Figure 3 are all above zero, suggesting that all 3 models give better validation forecasts than forecasting future values by the mean (for *E*) or by the current value (for *EM* and *EMP*). However, these values were quite different in magnitude. Although the values of *E* declined for longer leads, they were mostly around 0.7, whereas the values of *EM* and *EMP* were mostly around 0.5 and 0.2, respectively.

All versions of *E* identified NeTAR as the best model for describing the system and for short-term (1- and 2-day ahead) forecasting; and MNITF as the best for medium-term (3-day ahead) forecasting and NAARX for long-term (5-day ahead) forecasting. NeTAR and NAARX were equally better for 4-day ahead forecasting.

The difference in the magnitude of *E* and *EM* is clear evidence that using the current value as a reference (as in *EM*) raises performance standards, and gives more room for the measure before it reaches the maximum. When comparing *EM* and *EMP* (Figure 3), one notices that the *EM* values for the three models were more distinct than those of *EMP*. This same pattern was repeated with *dM* and *dMP* as well. The different values on measures that pronounce the effect of outliers differently demonstrate that the three models have different sensitivity to outliers. From the patterns shown in Figures 3(c) and (e), as well as in Figures 3(d) and (f), one can conclude that NeTAR is the least sensitive and NAARX is the most sensitive model to outliers. MNITF falls between the two.

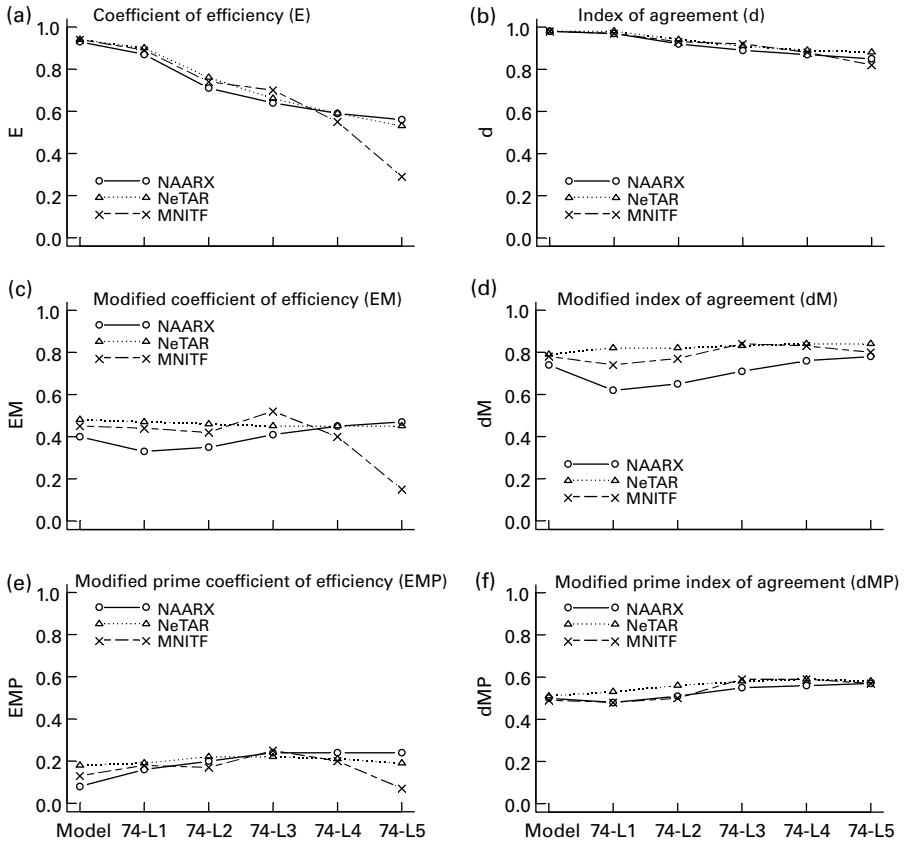

Figure 3 Plot of six performance measures for the NAARX, NeTAR and MNITF models. Model in the *x* axis refers to the 1-day ahead forecasts of the model using all 3 years data; and 74-L1 to 74-L5 refer to 1- to 5-day ahead validation forecasts of 1974 using the model calibrated using 1972–3 data of Jökulsá eystri

According to *d*, *dM*, and *dMP* shown in the second column of Figure 3, NeTAR describes the system best and gives validation forecasts with better agreement with the actual values for 1-, 2- and 5-day lead times. For 3- and 4-day leads, both MNITF and NeTAR performed best according to all three measures (*d*, *dM* and *dMP*).

The differences between the calculated values of *d*, *dM* and *dMP* were not as pronounced as those between *E*, *EM* and *EMP*, which emphasize the tendency of the index of agreement to give a false sense of model performance. The values of *d*, *dM*, and *dMP* were mostly around 0.9, 0.8, and 0.6, respectively. The different versions of the index of agreement suggested similar rankings of the models in most instances except for 1-day and 2-day validation forecasts where *dM* put MNITF in second place and *dMP* put both MNITF and NAARX in second place.

The results shown in Figure 3 also demonstrate a fundamental difference between the unmodified and modified versions of *E* and *d*. While the values of *E* and *d* always decrease with increasing forecast lead time, the values of their modified versions may or may not decrease as forecast lead time increases. This is so because the modified versions have a moving reference (the current value). Obviously, as the forecast lead time increases the forecast error also increases. However, when this increase is smaller than the amount by which the current value (the reference) deviates from the actual value at that lead time, the values of *EM*, *EMP*, *dM* and *dMP* would increase. This increase can be different for different models. For this particular dataset, the modified values of NAARX increased the most, suggesting that it is more suited for a longer term forecasting than NeTAR and MNITF.

Conclusions and recommendations

In this paper, the short term forecasting performances of three nonlinear time series models, namely NAARX, NeTAR and a transfer function with the inputs generated according to the underlying dynamics (MNITF), for a benchmark (Jökulsá eystri daily streamflow) data were compared in terms of RMSE, MAE, *E*, *EM*, *EMP*, *d*, *dM* and *dMP*. This was done using all 3-year data for 1-day ahead forecasts, as well as using 2-year calibration and 1-year validation data for up to 5-day ahead forecasts.

All measures identified NeTAR as the best model to describe the Jökulsá eystri daily streamflow system, and to provide better forecasts for 1- and 2-day lead times. The RMSE and MAE, as well as all the six relative measures, identified MNITF as the best model for 3-day ahead forecasting of the validation period. Both absolute measures, and the unmodified and modified versions of the coefficient of efficiency identified the NAARX model as the best for 5-day ahead forecasting. While the forecasting ability of NAARX was improved and that of NeTAR stayed the same with increasing lead time, the forecasting ability of MNITF model deteriorated substantially. This was more evident in the RMSE, MAE and the three variations of *E* than in *d*, *dM* and *dME*. This is probably because precipitation was not separated into snowfall and rainfall, and hence the effective rain series generated using Equation (6) may have been weak. Among the three models, the NeTAR model can be recommended for describing the system, either the NeTAR or MNITF model for short-term forecasting, and the NAARX model for long-term forecasting of daily streamflows.

For this particular dataset, there was no complete agreement among the eight measures, nor was there alarming differences. In general, the values of the modified measures were lower than those of the unmodified, and the difference is more prominent with *E*. These results demonstrated that the modified versions use a higher standard of model performance, and the differences between *EM* and *EMP*, and between *dM* and *dMP*, articulate the sensitiveness of the different models to outliers. NeTAR was the least sensitive and NAARX was the most sensitive, with MNITF in between. In addition to articulating sensitiveness to outliers and measuring forecast performance against a naive forecast equal to the streamflow

at the time of forecast, the modified versions can be used to compare the potential of competing models for long-term forecasting by looking at how their values increase as the lead time increases. For this particular streamflow system NAARX has the most potential for long-term forecasting.

The results in this study also suggest that the NeTAR and/or MNITF models could be used for short-term forecasting of daily streamflows of drainage basins with seasonal snow accumulation, and that further tests, perhaps real-time, on more watersheds are warranted.

Acknowledgements

I thank the reviewers for their useful comments.

References

Armstrong, J.S. and Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *Int. J. Forecasting*, **8**, 69–80.

Astatkie, T. and Watt, W.E. (1998). Multiple-input transfer function modeling of daily streamflow series using nonlinear inputs. *Wat. Res. Res.*, **34**, 2717–2725.

Astatkie, T., Watt, W.E. and Watts, D.G. (1996). Nested Threshold Autoregressive (NeTAR) models for studying sources of nonlinearity in streamflows. *Nordic Hydrol.*, **27**, 323–336.

Astatkie, T., Watts, D.G. and Watt, W.E. (1997). Nested Threshold Autoregressive (NeTAR) models. *Int. J. Forecasting*, **13**, 105–116.

Chen, R. and Tsay, R.S. (1993). Nonlinear additive ARX models. *J. Am. Statist. Assoc.*, **88**, 955–967.

Garrick, M., Cunnane, C. and Nash, J.E. (1978). A criterion of efficiency for rainfall-runoff models. *J. Hydrol.*, **36**, 75–381.

Legates, D.R. and McCabe, G.J. (1999). Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Wat. Res. Res.*, **35**, 233–241.

Martinec, J. (1960). The degree-day factor for snowmelt runoff forecasting. *IUGG General Assembly of Helsinki, IAHS Commission of Surface Waters* IAHS Publication No. 51, IAHS Press, Wallingford, UK, pp. 468–477.

Meade, N. (2000). Evidence for the selection of forecasting methods. *J. Forecasting*, **19**, 515–535.

Nash, J.E. and Sutcliffe, J.V. (1970). River flow forecasting through conceptual models: I. A discussion of principles. *J. Hydrol.*, **10**, 282–290.

Schumann, G. and Lauener, G. (2005). Application of a degree-day snow depth model to a Swiss glacierised catchment to improve neural network discharge forecasts. *Nordic Hydrol.*, **36**, 99–111.

Tong, H. (1990). *Nonlinear Time Series: A Dynamical Systems Approach*, Oxford University Press, New York.

Tong, H., Thanoon, B. and Gudmundsson, G. (1985). Threshold time series modeling of two Icelandic riverflow systems. In *Water Resources Bulletin, Time Series Analysis in Water Resources*. K.W. Hipel (ed.), vol. 21, American Water Resources Association, Middleburg, VA, USA, pp. 651–661.

Watt, W.E. and Nozdryn-Plotnicki, M.J. (1982). Real-time flood forecasting for flood damage reduction. In *Decision Making for Hydrosystems: Forecasting and Operation*. T.E. Unny and A.E. McBean (eds.), Water Resources Publications, Fort Collins, CO, pp. 551–571.

Willmott, C.J. (1981). On the validation of models. *Phys. Geog.*, **2**, 184–194.

Wong, C.S. and Li, W.K. (2001). On a logistic mixture autoregressive model. *Biometrika*, **88**, 833–846.