Selecting reference streamflow forecasts to demonstrate the performance of NWP-forced streamflow forecasts

J.C. Bennett [a], **D.E. Robertson** [a], **D.L. Shrestha** [a] **and Q.J. Wang** [a]

[a] *CSIRO Land and Water, Graham Road, Highett, Victoria*
Email: james.bennett@csiro.au

Abstract: We assess four different reference forecasts for the purpose of measuring the skill of streamflow forecasts generated from Numerical Weather Prediction (NWP) model rainfall forecasts. The reference forecasts we investigate are 1) streamflow climatology, 2) persistence, 3) a hydrological model forced by zero rainfall and 4) a hydrological model forced by an ensemble of resampled historical rainfall. We assess performance of reference forecasts to lead-times of 9 days. Reference forecasts should be simple to produce, but also must be reasonably accurate to establish a robust performance threshold. We show that because streamflows are strongly autocorrelated, streamflow climatology is a very low performance hurdle to clear for any NWP-forced streamflow forecasts, particularly at short lead-times (<2 days). Conversely, because the shape of hydrographs is broadly predictable, persistence forecasts generally perform very poorly at longer lead times (>1 day). Using a hydrological model substantially improves the accuracy of reference forecasts, with resampled–historical-rainfall forced forecasts outperforming zero-rainfall-forced forecasts, particularly at longer forecast lead times. We argue that streamflow climatology and simple persistence are not accurate enough to be used as reference forecasts. We recommend the use of reference forecasts generated by resampled historical rainfalls as a robust performance benchmark of NWP-forced streamflow forecasting systems. We demonstrate the use of resampled–historical-rainfall forced reference forecasts to assess the performance of a new Australian ensemble streamflow and flood forecasting system developed by CSIRO and the Bureau of Meteorology.

Keywords: *Streamflow forecast, reference forecast, forecast verification, numerical weather prediction*

1. INTRODUCTION

Numerical weather prediction (NWP) model rainfall forecasts have the potential to extend real-time streamflow forecasts (forecasts generated only from observed rainfalls and streamflows) to lead times of the order of 10 days, with attendant benefits to water resource and flood managers. NWP-forced streamflow forecasting systems (here abbreviated to NSFS) are usually made up of a conceptual hydrological model that is spun up to initialise its states and then forced by NWP forecast rainfalls. Often an error correction (or error updating) model is also employed as part of an NSFS to align streamflow forecasts with observed streamflows when the forecast is issued.

Traditionally, the simulation performance of hydrological models is assessed against observed streamflows, under the assumption that rainfall observations will be available whenever the model is used to predict flows (as is the case in real-time forecasting systems). However, using observed streamflows to assess forecasts from an NSFS is an unreasonably stringent test, as NWP rainfall forecasts cannot be expected to predict rainfall perfectly, even as they offer useful information. Accordingly, the performance of NSFS is usually assessed against *reference forecasts*: alternative (usually simpler) estimates of future streamflows.

There is no consensus on which reference streamflow forecasts should be used to test NSFS. Climatology has traditionally been used as a reference forecast to verify meteorological forecasts, while Pappenberger et al. (2008) tentatively advocate the use of persistence forecasts (i.e., using the last available observation as the forecast) for assessing flood forecasts. To establish a robust performance threshold for NSFS, reference forecasts should be as accurate as possible. The relative performance of different reference forecasts for NSFS has not, to our knowledge, been investigated in published literature. We rectify this omission by comparing the performance of four reference forecasts: streamflow climatology, persistence, the hydrological component of an NSFS forced with zero rainfall (rather than NWP rainfall) and the hydrological component of an NSFS forced with resampled historical rainfall. We test the utility of these reference forecasts by comparing them to streamflow forecasts from an NSFS.

2. CATCHMENTS

We present results for two catchments, the Cotter River in south-eastern Australia's Great Dividing Range and the South Esk River in north east Tasmania. Gauge site locations are given in Figure 1 and catchment characteristics are listed in Table 1.

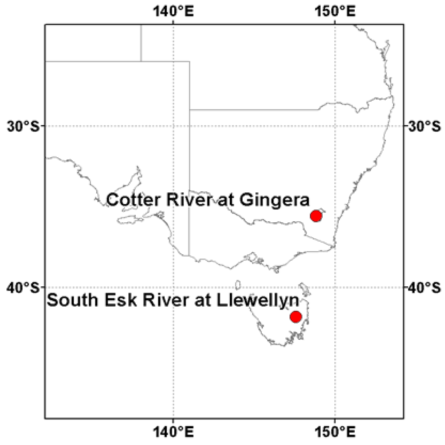

Figure 1. Locations of gauge sites (red dots).

Table 1. Catchment characteristics

Site Name	Period of available hourly observations	Drainage area (km^2)	Annual runoff (mm)	Annual rainfall (mm)	Annual PET (mm)
Cotter River at Gingera	01/01/1990-01/06/2012	145	276	876	1117
South Esk River at Llewellyn	01/01/2001-01/06/2012	2284	234	686	950

3. NUMERICAL WEATHER PREDICTION FORCED STREAMFLOW FORECASTS

For this study, we use the new ensemble NSFS that is being jointly developed by CSIRO and the Bureau of Meteorology. This system produces ensemble 9-day streamflow forecasts at an hourly timestep. The system applies a rainfall post-processor (RPP) to rainfall forecasts from the ACCESS-G NWP model to correct biases and quantify rainfall forecast uncertainty (Robertson et al., 2013). Post-processed NWP rainfall forecasts are then used to force the GR4H rainfall-runoff model (an hourly variant of the daily GR4J model described by Perrin et al., 2003), and runoff is routed with the well-known linear Muskingum channel routing algorithm.

GR4H differs from the daily GR4J model described by Perrin et al. (2003) as follows:

1. The X_2 parameter is multiplied by 0.67, and the X_3 parameter is multiplied by 2.21.

2. The percolation function for GR4H assumes an original maximum store capacity of $4 \times X_1$.

3. GR4H has a unit-hydrograph exponent of 5/4.

We use a modified version of the dual-pass error correction model (Pagano et al., 2011) to update hydrological model predictions and correct biases. Only the second pass of the dual-pass model is used. The correction is decayed exponentially so the forecast gradually reverts back to the 'raw' streamflow forecast.

For convenience, we refer to the combination of GR4H, channel routing and error correction models as the *hydrological model*.

ACCESS-G forecasts are available from 1 August 2010 to 30 April 2012, which means the NSFS could only be evaluated for this period (accordingly, we refer to this as the *evaluation period*). Independent post-processed rainfall forecasts are obtained for the evaluation period using a leave-one-month-out cross-validation procedure (Robertson et al., 2013). The hydrological model parameters are calibrated using all available hourly rainfall and streamflow data before 1 January 2010 (data availability is listed in Table 1).

4. REFERENCE FORECASTS

4.1. Streamflow climatology

Streamflow climatology is calculated by taking long-term averages of observed streamflow at each ordinal date across years. This results in a very noisy climatology. To smooth the series, we apply a 31-day moving average. We calculate streamflow climatology from all data available before 1 January 2010.

4.2. Persistence

Persistence takes the most recent streamflow observation available at the forecast issue date and assumes this flow will continue for the duration of the forecast (in this case, 9 days).

4.3. Hydrological model forced by zero rainfall

The performance of the hydrological modelling component of any forecasting system can be readily evaluated with traditional comparisons to observed hydrographs. When initialised, a well-functioning and hydrological model should be able to accurately represent hydrograph recessions, even without rainfall forcing, during the forecast period. This is particularly true when an error correction model is applied as part of the hydrological model. Accordingly, we test if reference forecasts generated by the hydrological model forced by zero rainfalls will outperform persistence forecasts. We refer to these as *zero-rainfall-forced* reference forecasts.

4.4. Hydrological model forced by resampled historical rainfall

In addition to zero-rainfall forcing, we generate reference forecasts with the hydrological model forced by resampled historical rainfall. The method we use to resample historical rainfall is similar to that use to generate the ensemble streamflow prediction forecasts long used by the United States National Weather Service (see, e.g., Day, 1985). Rainfalls are resampled from the historical record for each ordinal forecast issue date. For each ordinal date, we first sample 9-day rainfall sequences beginning at this ordinal date for each year in the observation record. We then sample 9-day rainfall sequences from the previous ordinal day and then the following ordinal day from each year in the observations record. We repeat this process until we have 100 sequences of 9-day rainfalls from the historical record. To generate reference forecasts, we force the hydrological model with the 100-member ensemble of resampled historical rainfall. We refer to these as *historical-rainfall-forced* reference forecasts.

5. VERIFICATION MEASURES

The lack of consensus on which reference forecasts to employ arises in part because the choice of reference streamflow forecasts may depend on the objectives of the forecast: for example, forecasts may be intended either for flood prediction or for water allocation, or forecasters may be more interested in shorter or longer lead times, all of which can influence the choice of reference forecast. Accordingly, we compare reference forecasts using three verification measures that can be used to assess the performance of an NSFS in forecasting: 1) floods (the Nash-Sutcliffe efficiency, NSE), 2) overall flows, including the performance of ensemble forecasts (the Continuous Ranked Probability Score, CRPS), and 3) flow volumes for water resource allocation (bias). These are described in more detail below.

We present all verification measures conditioned on lead-time. Verification is performed for the evaluation period 1 August 2008 to 30 April 2010.

5.1. Nash-Sutcliffe Efficiency (NSE)

The NSE (Nash and Sutcliffe, 1970) is commonly employed as a deterministic verification measure in hydrology. NSE is defined as

$$NSE = 1 - \frac{\sum_{t=1}^{n} \left(Q_{Fcst,t} - Q_{Obs,t}\right)^2}{\sum_{i=1}^{n} \left(Q_{Fcst,t} - \overline{Q_{Obs}}\right)^2} \tag{1}$$

where Q_{Fcast} and Q_{Obs} are the forecast and observed discharges at time *t*. To apply NSE to ensemble forecasts we take the mean of the ensemble at each time step. Because the error term is squared in NSE, this puts greater emphasis on instances where errors are very large. Very large errors tend to occur at larger flows, meaning NSE is something of a *de facto* measure of forecast performance at large flows.

5.2. The Continuous Ranked Probability Score (CRPS)

The CRPS is regularly used as an overall measure of ensemble forecast performance. The CRPS measures the error of all ensemble members with respect to observations, in the units of the forecast (in our case, m^3/s), by calculating the area between the cumulative distribution functions (CDFs) of the forecast and observation. Mathematically, the CRPS is expressed as:

$$CRPS = \int_{-\infty}^{\infty} [F(Q) - H(Q - Q_{Obs})]^2 .dQ \tag{2}$$

Where *F(Q)* is the forecast CDF and $H(Q\text{-}Q_{obs})$ is the Heaviside function, which takes the value 0 when $Q\text{-}Q_{Obs}<0$, and 1 otherwise.

Larger CRPS values indicate a poorer forecast. A highly useful characteristic of the CRPS is that it collapses to the mean absolute error for deterministic forecasts, allowing comparison of ensemble and deterministic forecasts.

5.3. Bias

Bias measures the difference in total volume between simulated and observed streamflows, as follows:

$$Bias = \frac{\overline{Q_{Fcst}} - \overline{Q_{Obs}}}{\overline{Q_{Obs}}} \tag{3}$$

Bias is a deterministic measure. As with NSE, we extend bias to ensemble forecasts by defining Q_{Fcst} as the average of all ensemble members for each time step. We do not measure the bias of persistence forecasts. For persistence, $\overline{Q_{Fcst}} \approx \overline{Q_{Obs}}$ during the evaluation period (by definition), and this renders bias a meaningless performance measure for persistence.

Bias gives no indication of how well forecasts represent the shape or variation of observed hydrographs, and consequently is likely to be the performance measure of least interest to users of short-medium term forecasts. Nonetheless, bias is useful for diagnosing long-term forecast errors, and we include it accordingly.

6. RESULTS AND DISCUSSION

NSE, bias and CRPS for both catchments are shown for all reference forecasts in Figure 2. In addition, Figure 2 shows performance metrics for NSFS forecasts and for perfect rainfall forecasts (hydrological model forced by observed rainfall). Perfect rainfall forecasts show the best forecast performance attainable given the limitations of the hydrological model.

At lead-times of 3 days or fewer, streamflow climatology is often the poorest performing forecast for all verification measures. This is unsurprising, for two major reasons. First, the other reference forecasts are updated with streamflow information available at the forecast issue time. Because streamflows are usually highly autocorrelated, this information allows all the other forecasts to outperform streamflow climatology at

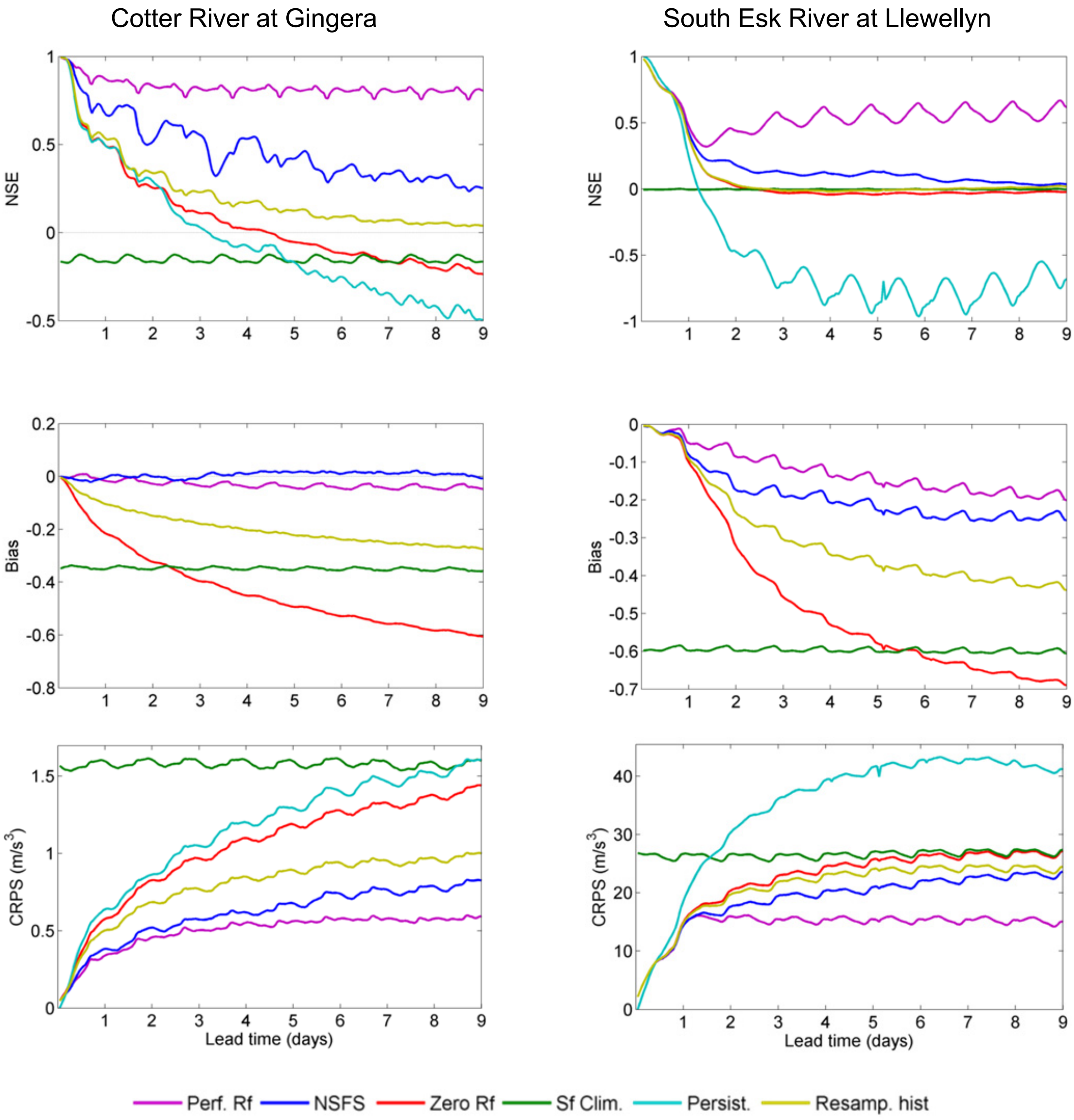

Figure 2. Performance measures for a range of reference forecasts calculated at each lead time for the Cotter River (left column) and the South Esk River (right column). Top row shows NSE, middle row shows bias, bottom row shows CRPS. Forecasts shown are: hydrological model forced with perfect (observed) rainfall forecasts (purple); hydrological model forced with post-processed NWP rainfall forecasts (blue); hydrological model forced with zero rainfall forecasts (red); streamflow climatology (green); persistence (light blue); hydrological model forced with resampled climatology rainfall forecasts (yellow).

short lead times. Second, the period over which streamflow climatology is calculated included 2001-2010, the period of the so-called 'Millenium Drought' in south-east Australia, one of the driest periods in south-east Australia in recorded history. Conversely, the evaluation period experienced above-average rainfalls. As a consequence, streamflow climatology chronically underestimates streamflows in the evaluation period, as shown by strongly negative biases in Figure 2.

Streamflow climatology might be expected to be a reasonably stringent measure of performance at longer lead-times, where catchment memory has a reduced influence on the forecast skill of the other reference forecasts. It is, however, one of the poorest performing reference forecasts at longer lead times because it is so strongly biased, as shown particularly by the CRPS (Figure 2). While streamflow climatology will not necessarily be biased for a given evaluation period, our example shows that this can occur. At both longer and shorter lead-times, then, streamflow climatology is a very low performance hurdle for an NSFS to clear.

Persistence reference forecasts perform consistently well at short lead times, a product of the high autocorrelation of streamflows. The forecast horizons for which persistence performs well will vary between catchments, as we show here: persistence shows some skill (NSE>0) for lead-times of 3 days in the Cotter River, but is only skilful for lead times of <1 day in the South Esk River (Figure 2). We note, however, that the performance of persistence at short lead times is no better than the hydrological model forced by zero rainfalls for both NSE and CRPS.

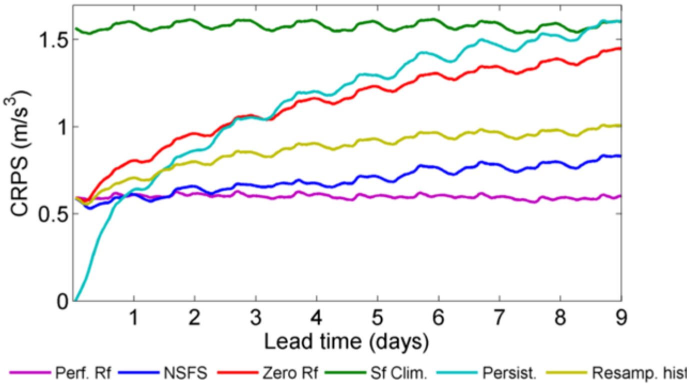

Figure 3. CRPS for the Cotter River for forecasts without hydrological error correction. Forecasts shown are: GR4H forced with perfect (observed) rainfall forecasts (purple); GR4H forced with post-processed NWP rainfall forecasts (blue); GR4H forced with zero rainfall forecasts (red); streamflow climatology (green); persistence (light blue); GR4H forced with resampled climatology rainfall forecasts (yellow).

Zero-rainfall-forced reference forecasts perform well in relation to persistence in Figure 2 largely because the hydrological model includes an effective error correction model. When the error correction model is not included as part of the hydrological model, persistence outperforms zero-rainfall-forced forecasts for lead-times up to ~3 days (Figure 3). Indeed, without hydrological error correction, persistence outperforms even perfect-rainfall forced forecasts for lead-times up to 1 day (Figure 3). This underscores the need for error correction (or other forms of data assimilation) to make use of recent streamflow observations in any NSFS. When recent observations are used to inform the hydrological model, however, zero-rainfall-forced reference forecasts perform as well at short lead times as persistence, and substantially better at longer lead times. We contend, accordingly, that persistence is always an insufficiently stringent reference forecast at longer lead-times, and is no better than a well-functioning hydrological model at shorter lead times.

Resampled-historical-rainfall forced reference forecasts are the best performing reference forecasts for all lead-times and for all measures of performance (Figure 2). Zero-rainfall-forced forecasts tend to underestimate flows at longer lead times when soil moisture stores in the hydrological model become depleted. Resampled-historical-rainfall forced forecasts add historically plausible quantities of water to the system, leading to less negatively-biased reference forecasts (*cf.* zero-rainfall-forced forecasts) at longer lead times. As expected, when rainfalls are sampled from a significantly drier period than the evaluation period (as occurs in both catchments), resampled-historical-rainfall forced forecasts tend to be negatively biased in the evaluation period (Figure 2). Similarly, the reverse effect can be expected if the evaluation period is significantly drier than historical resampling period. Despite this, the resampled-historical-rainfall forced forecasts show the highest NSE values, are least biased, and have the lowest CRPS of any reference forecasts, particularly at long lead times. CRPS tends to advantage ensemble forecasts, because any reasonable estimate of uncertainty tends to reduce the areal difference between the ensemble CDF and the observed CDF. Nonetheless, we have shown that the resampled-historical-rainfall forced reference forecasts outperform the other (deterministic) reference forecasts in the deterministic verification measures (NSE and bias), validating the CRPS results.

As with zero-rainfall-forced forecasts, at shorter lead-times (<2 days) resampled-historical-rainfall forced forecasts performs comparably well with persistence only when hydrological error correction is applied. When hydrological error correction is not applied to the Cotter River, for example, persistence outperforms resampled-historical-rainfall forced forecasts at lead times of 36 hours or fewer (Figure 3).

When hydrological error correction is applied, the performance of persistence, zero-rainfall and reference forecasts is very similar for lead-times of one day or fewer. Accordingly, any of these three is a sufficiently stringent test of NSFS at very short lead-times. However, at longer lead-times the performance of resampled-historical rainfall is clearly superior to other reference forecasts, and offers the most rigorous test of any NSFS for a range of verification measures. Accordingly, we recommend that NWP forecast streamflow forecast systems be assessed against resample-historical-rainfall forced forecasts.

In addition to outperforming other reference forecasts, resampled-historical-rainfall forced reference forecasts have the considerable benefit of generating a reference forecast ensemble. Much attention has recently been directed at generating ensemble forecasts from NWP inputs to represent uncertainties in the forecasts: for example, the NSFS presented here generates ensemble forecasts that reflect uncertainty in NWP rainfall forecasts. Having an ensemble reference forecast allows comparison of ensemble-specific verification measures, most notably those related to statistical reliability (i.e., measures that demonstrate whether the uncertainty in the forecasts is correctly represented by the ensemble).

Figure 2 shows that the NSFS outperforms even the stringent resampled-historical-forced reference forecasts at all lead times for these two catchments in all verification measures. This gives a strong indication that the forecasts from this NSFS are likely to be useful to end-users. We note, however, that the performance of the NSFS would be overstated at long lead times had we chosen, for example, persistence as the reference forecast. This would give an unreasonably inflated assessment of value to end-users for long lead-time forecasts.

7. SUMMARY AND CONCLUSIONS

We have investigated four reference forecasts for the purposes of assessing the performance of NWP-forced streamflow forecasting systems: streamflow climatology, persistence, a hydrological model forced by zero rainfall; and a hydrological model forced by resampled historical rainfalls. We aimed to find the best performing reference forecast for a range of verification measures, in order to offer the most stringent test to NWP-forced streamflow forecasts. Streamflow climatology gives by far the poorest reference forecasts, even at long lead times. We show that a well-functioning hydrological model (in this case, one that applies hydrological error correction) is able to produce similarly accurate streamflow forecasts to persistence at short lead-times, and perform markedly better than persistence at longer lead-times, even when forced by zero rainfall.

Forecasts generated by a hydrological model forced with resampled historical rainfalls offered the best overall performance. They performed similarly to persistence at short-lead times (<2 days) and clearly outperformed all other reference forecasts at longer lead-times. Accordingly, we recommend that resampled-historical-rainfall forecasts be used to measure the performance of NWP-forced streamflow forecasts in preference to persistence or other reference forecasts.

We note that for any reference forecast that relies on a hydrological model, extreme care must be taken to ensure the hydrological model performs well. This includes, in particular, employing an effective method for updating forecasts with recently observed streamflows.

ACKNOWLEDGMENTS

This research has been supported by the Water Information Research and Development Alliance (WIRADA) between the Australian Bureau of Meteorology and CSIRO Water for a Healthy Country Flagship. Thanks to David Enever and Prasantha Hapuarachchi (Bureau of Meteorology) for providing data and catchment delineations.

REFERENCES

Day GN. 1985. Extended streamflow forecasting using NWSRFS. Journal of Water Resources Planning and Management 111: 157–170

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models part I - A discussion of principles. Journal of Hydrology 10: 282–290. DOI: 10.1016/0022-1694(70)90255-6.

Pagano TC, Wang QJ, Hapuarachchi P, Robertson DE. 2011. A dual-pass error-correction technique for forecasting streamflow. Journal of Hydrology 405: 367-381. DOI: 10.1016/j.jhydrol.2011.05.036.

Pappenberger F, Scipal K, Buizza R. 2008. Hydrological aspects of meteorological verification. Atmospheric Science Letters 9: 43-52. DOI: 10.1002/asl.171.

Perrin C, Michel C, Andréassian V. 2003. Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology 279: 275-289. DOI: 10.1016/S0022-1694(03)00225-7.

Robertson DE, Shrestha DL, Wang QJ. 2013. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrology and Earth System Sciences Discussions 10: 6765-6806. DOI: 10.5194/hessd-10-6765-2013.