20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Evaluation of Downscaled POAMA M24 for Monthly and 3-Monthly Streamflow Forecasts

Hongxing Zheng[a], QJ Wang[b], Kabir Aynul[c], Quanxi Shao[d], Daehyok Shin[c], Narendra Tuteja[c]

[a] *CSIRO Water for a Healthy Country Flagship / CSIRO Land and Water, Canberra, Australia.*
[b] *CSIRO Water for a Healthy Country Flagship / CSIRO Land and Water, Highett, Victoria Australia.*
[c] *Climate and Water Division, Bureau of Meteorology, Australia.*
[d] *CSIRO Mathematics, Informatics and Statistics, Floreat, Western Australia, Australia.*
Email: Hongxing.zheng@csiro.au

Abstract: This paper evaluates skill of ensemble forecasts of monthly and three-monthly streamflows for three catchments from different hydrological regions, using a conceptual hydrological model - the GR4J. The latest available POAMA-M24 rainfall predictions for the period of 1980-2008 are downscaled and used as forcing inputs of the model to produce streamflow forecasts. In dealing with model uncertainty, 200 parameter sets derived through BATEA are used for each downscaled rainfall forcing. The results show that skill scores are both catchment and season dependent. In the Biggara catchment (SMD region), Jan., Apr., and Nov. are the months with the highest skill scores; for the Picnic Crossing catchment (QLD region), the best forecasts are for June and July, while for the Tinderry catchment (SEC region), the best forecast months are Sept. to Nov. and Feb. as well. The skill scores of monthly steramflow forecasts are higher than that of three-monthly forecasts except for reliability. Slight difference between M24-E33 and M24-E99 is found when additional forcings with lead times of 1 and 2 months are used in the ensemble forecasting.

Keywords: *streamflow forecasts, ensemble forecasts, GR4J, POAMA*

1. INTRODUCTION

Ensemble streamflow prediction (ESP) produces multiple streamflow forecasts based on an ensemble of meteorological forcings (e.g. rainfall and potential evapotranspiration), each member of the ensemble is a possible realization of future weather condition. The ensemble forcings can be that re-sampled conditionally from historical records or that downscaled from the GCM outputs [Day, 1985; Wood et al, 2005; Wang et al, 2011a; Tuteja et al, 2011]. In a comparison with using historical ensemble inputs, Wood et al [2005] showed that GCM forecasts for regionally averaged variables did not improve streamflow forecasts in the USA. However, during strong ENSO years, forecasts may or may not benefit from using GCM forecasts, depending on regions having strong or weak tele-connection with the ENSO.

As reported in a preliminary investigation on selected catchments in East Australia, the use of the ESP approach with historical ensemble forcings showed useful skills for monthly and three-monthly streamflow forecasts, especially for the months and seasons following wet seasons [Wang et al, 2011a]. An evaluation of the ESP approach with rainfall predictions downscaled from POAMA-P24 (Predictive Ocean Atmosphere Model for Australia) indicated that the ESP approach was comparable to those obtained from historical ensemble forcings [Wang et al., 2011b].

POAMA is a state-of-the-art seasonal climate forecast system developed by the Australia Bureau of Meteorology based on a coupled ocean/atmosphere model and ocean/atmosphere/land observation assimilation systems [Alves et al,. 2003]. The most recent launched POAMA M24 is now available to public. The major difference between M24 and its previous version (P24) is that M24 is updated at the frequency of 10 days (i.e., the model is updated on the 1st, 11th, and 21st days of each calendar month), while the previous versions (P15 and P24) are updated at the frequency of one month (i.e., the model is updated at the beginning of each calendar month).

The objective of this paper is to evaluate the skills of ensemble streamflow forecasting by using downscaled POAMA M24 rainfall predictions as forcings of a conceptual hydrological model for selected catchments. The performance of ESP using M24 is investigated via cross validation.

2. SELECTED CATCHMENTS

Three catchments locating at SMD, SEC and QLD in east Australia are selected for this study. The locations of the catchments are shown in Fig. 1. Catchment average rainfall data are derived from AWAP (Australian Water Availability Project) grid data with about 5 km spatial resolution [Raupach et al., 2009]. The inter-annual variation of streamflow of the three catchments shows that Biggara at SMD region is spring dominant and Picnic Crossing is summer dominant, whilst the streamflow regime of Tinderry at SEC region is uniform.

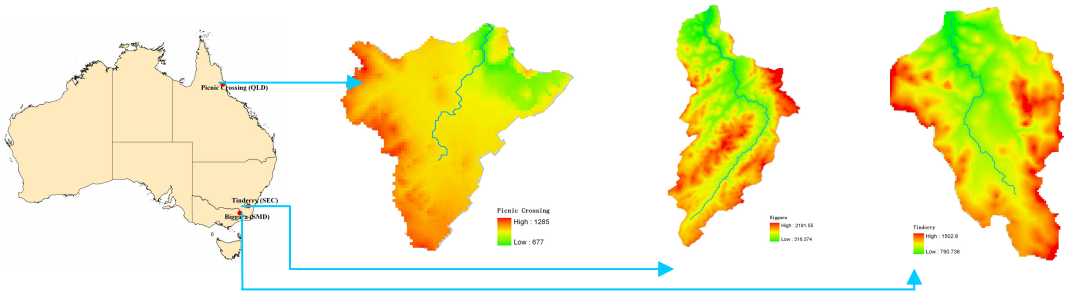

Figure 1 Locations of the selected catchments

3. METHODOLOGY

3.1. Downscaled ensemble forcings

To predict streamflow using a dynamic hydrological model, the POAMA output rainfall is first downscaled to the catchment scale based on a modified statistical analogue method developed by Shao and Li [2013], where a two-stage bias correction was applied in conjunction with Timbal's analogue method [Timbal, 2008]. The two-stage bias correction method first transforms the non-normal POAMA data to normal distribution and the transformed data is then standardized. The method has been proved be able to largely retain the correlation between NCEP/NCAR and POAMA data. The downscaled daily rainfall from POAMA M24 includes 33 ensemble members for the period 1980-2008 with lead time of zero month. Two ensemble streamflow prediction schemes are applied for the monthly streamflow forecast. One of the schemes (E33)

includes 33 members of downscaled M24 rainfall with lead time of zero-month, while the other scheme (E99) includes 99 members of downscaled M24 rainfall with lead time of 0, 1 and 2 months. For three-monthly streamflow forecast, only the E33 scheme is applied owing to the availability of the POAMA predictions.

3.2. Streamflow forecasting

The GR4J model [Perrin et al., 2003] is used for ensemble streamflow prediction. With the consideration of the uncertainty in hydrological model, the GR4J model is firstly run under the BATEA framework to derive multiple parameter sets for each prediction period (herein, 200 parameter sets are derived). The initial condition of each prediction is generated accordingly using the derived parameter sets and the real rainfall before the prediction period. Subsequently, for each ensemble member of forcing, the GR4J model runs 200 times and produces 200 forecasts of the same period. Therefore, for the E33 schemes the ensemble streamflow prediction procedure produces 6,600 forecasts, while for the E99 scheme there are 19,800 forecasts for every prediction period.

3.3. Evaluation criteria

To evaluate the performance of ensemble streamflow forecast using POAMA M24, a leave 5-year out cross validation scheme is implemented. In addition to R^2 representing correlation between mean of ensemble forecast and observed streamflow (1980-2008), the skill scores based on NSE_{ref}, RMSEPS and SSCRPS are used. The NSE_{ref} is defined as:

$$NSE_{ref} = 1 - \sum_{t=1}^{N} (y_{f,t}^{mean} - y_{o,t})^2 / \sum_{t=1}^{N} (\bar{y}_{ref,t} - y_{o,t})^2 \tag{1}$$

where, $\bar{y}_{ref,t}$ is the mean monthly streamflow of the reference period (before 1980), $y_{o,t}$ is the observed streamflow, while $y_{f,t}^{mean}$ is the mean of the ensemble forecasts. The root mean square error in probability score (SSRMSEP) proposed by Wang et al. [2009]. The value of SSRMSEP is not larger than 1. The lager the SSRMSEP value is, the better the ensemble forecasts is. Negative SSRMSEP means that the ensemble forecast performs poorer than the climatology average of the reference period.

The CRPS is formulated for verification of probabilistic forecasts of continuous variables [e.g., Brown, 1974; Matheson and Winkler, 1976; Bouttier, 1994; Hersbach, 2000; Gneiting et al., 2007; Laio and Tamea, 2007; Wang et al., 2009]. The equation for calculation of the CRPS for a specified case t is,

$$CRPS^t = \int_{-\infty}^{\infty} \left[F_f^t(y^t) - F_{obs}^t(y^t) \right]^2 dy^t \tag{2}$$

where F_f^t is the forecast probability *cdf* for the forecast case at t, and F_{obs}^t is the observation expressed as a cdf. If the observation is of a specific value, then the corresponding *cdf* is a single step-function with the step from 0 to 1 at the observed value of the variable. Therefore, the expression of CRPS can be rewritten as,

$$CRPS^t = \int_{-\infty}^{\infty} \left[F_f^t(y^t) - H(y^t - y_{obs}^t) \right]^2 dy^t \tag{3}$$

where

$$H(y^t - y_{obs}^t) = \begin{cases} 0, & y^t < y_{obs}^t \\ 1, & y^t \geq y_{obs}^t \end{cases} \tag{4}$$

with H( ) is the Heaviside function. CRPS can be regarded as the total area between the *cdf* of the probabilistic forecasts and the *cdf* of the observation. The minimum CRPS value of zero is achieved only in the case of a perfect single value forecast. A skill score based on CRPS is then defined as [Wang et al., 2009]

$$SS_{CRPS} = 1 - \overline{CRPS^t} / \overline{CRPS_{ref}^t} \tag{5}$$

where $\overline{CRPS^t}$ is the average CRPS of all forecast cases, $\overline{CRPS^t_{ref}}$ is the average of CRPS when a reference probabilistic forecast $y^t \sim F^t_{ref}(y^t)$ is used. It is obvious that the skill score for reference probabilistic forecasts is zero, while SSCRPS reaches to 1 for perfect forecast.

The probability integral transform (PIT) is often used to evaluate the reliability of the ensemble forecasts. The PIT plot is useful for indicating whether the forecast probability distributions are predicted too high or too low, or too wide or narrow [Laio and Tamea, 2007; Thyer et al., 2009]. Given a probabilistic forecast $y^t \sim F^t_f(y^t)$, a probability integral transform can be applied to the observed value yo to give

$$\pi_t = F^t_f(y_{o,t}) \tag{6}$$

where the function F^t_f is derived from all ensemble forecast members at case *t*. For an ideal forecast, π_t should be uniformly distributed. The uniformity can be checked by pooling together π_t values for all the forecast cases t = 1,2, . . ., n, and displaying the ranked π_t values in a uniform probability plot with the Kolmogorov-Smirnov confidence bands. For the convenience of comparison, the significance level of the band is accepted herein as an indicator for the uniformity test, which is given as:

$$Alpha = 1 - P_k(x < \sqrt{n} D_{\max}) \tag{7}$$

where *x* is a random variable of Kolmogorov distribution (P_k), n is the sample size of π_t, and D_{max} is the maximum distance between the π_t values and the diagonal line in a PIT plot. If π'_i is the series ranked ascending of π_t, and the empirical probability is i/n, then

$$D_{\max} = \sup\left|\pi^t_i - i/n\right|, i = 1,2,...,n \tag{8}$$

According to the definition, a smaller D_{max} means that π_t is closer to uniform distribution at a higher significance level Alpha and therefore the ensemble forecasts are more reliable. Alpha varies in (0, 1). For significance level Alpha=0.05, the critical value $D_c = 1.36/\sqrt{n}$, which implicates that PIT plot is within the band of width ±Dc with probability of 95%.

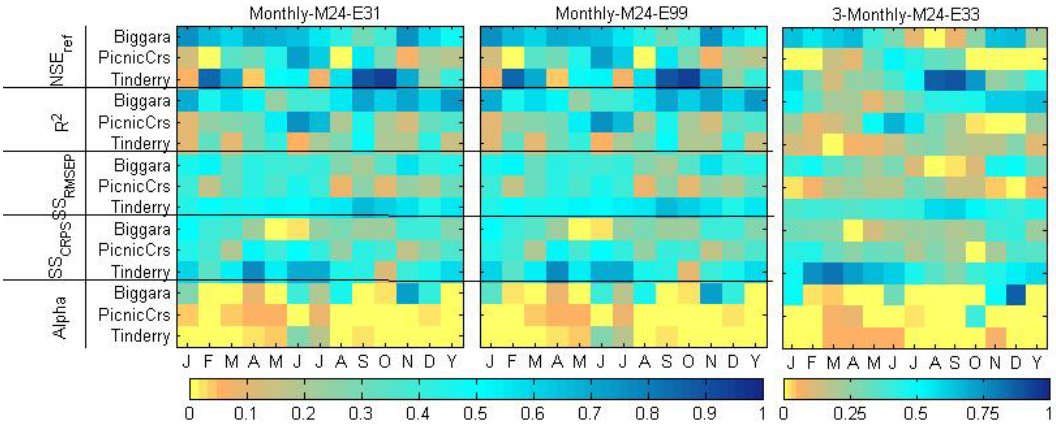

Figure 2 Forecasting skills of monthly and 3-monthly streamflow. M24-E33 and M24-E99 stand for POAMA M24 with 33 and 99 ensemble forcings respectively.

4. RESULTS

4.1. Skill scores of monthly forecasts

Fig. 2 shows the forecasting skills of the monthly streamflows generated using downscaled POAMA M24 as ensemble forcings for the GR4J model. For the forecasting scheme M24-E33 and M24-E99, no significant

difference of skill score is found, indicating that the additional forcings with lead time 1 and 2 month do not improve the forecasting skill. The NSE_{ref} of most months and catchments (94.4%) are larger than 0, meaning that the forecasts are better than reference predictions by climatology mean. The proportion of months and catchments with useful skills of forecasts ($NSE_{ref} \geq 0.2$) is around 80.6%. The mean value of those months×catchments with $NSE_{ref} > 0$ reaches to 0.48. In term of correlation coefficient (R^2) larger than 0.355 (significance level of 0.05), the proportion of months×catchments is around 50%. The mean SSRMSEP of all months×catchments reaches to 0.40. For most months of the three catchments, the SSRMSEP values are all larger than 0, again indicating a better performance than climatology mean. This is further proved in SSCRPS shown at Fig.2, where the mean SSCRPS of all months×catchments reaches to 0.38. The reliability of the ensemble forecasts in term of the Alpha shows that the distribution of the forecasts is reasonable only in some months of the catchments (27.8%). Most of the months and catchments have an Alpha value less than 0.05, meaning that the PIT values do not lie within the Kolmogorov 5% significance bands, and the probability distributions of the forecasts may be biased or with a spread that is too high or too low, or too wide or too narrow. Fig.3 provides a comparison of forecast median and [0.05, 0.95] quantile range with observed value for individual cases. It appears that the forecast median is consistent with the observed value for Biggara at SMD region. For Picnic Crossing at QLD and Tinderry at SEC, however, the observed values show to be out of the range when the observed monthly streamflow is higher than a threshold, indicating a forecast bias in high flow months.

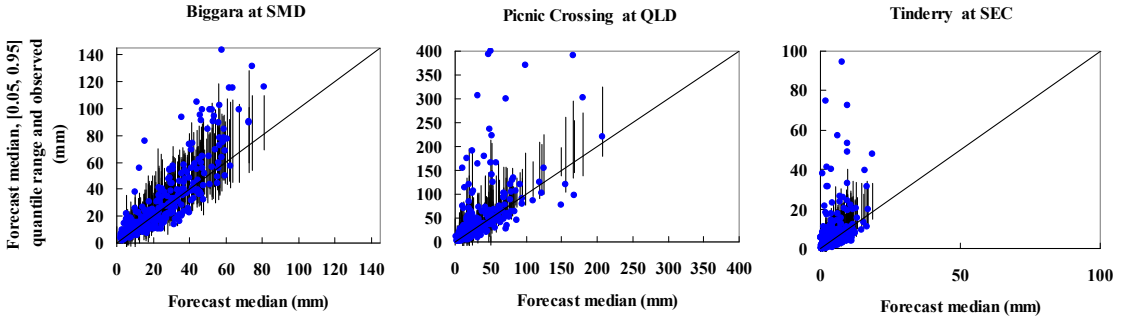

Figure 3 Monthly streamflow forecast quantiles and observed value plotted according to forecast median (M24-E33)

It can also be seen at Fig.2, the skill scores are both catchment and season dependent. Among the three catchments, Biggara at the SMD region shows the highest skill scores for monthly streamflow forecasts, while Picnic Crossing at QLD region is the catchment with the lowest skill scores. In Biggara, both the NSE_{ref} and SSRMSEP for Aug. to Oct. are lower than the other months. However, the months with smaller R^2 and SSCRPS are May and June. In term of reliability based on Alpha, the scores of Jan., Jun., Aug., Nov. and Dec. are all higher than 0.05. For Picnic Crossing at QLD, Mar. to Jul. has higher skill scores than other months, but reliable forecasts are only in Apr., May and Jul. For Tinderry at the SEC region, the median of the ensemble forecasts show significant correlation with the observed monthly streamflow in May, Sept. and Dec., where R^2 are all higher than 0.355. However, NSE_{ref}, SSRMSEPS and SSCRPS of Tinderry in all months are all positive, indicating that the ensemble forecast is much better than the climatology mean. It is obvious that the ensemble forecast is more reliable for June and July, where a higher Alpha value is found.

4.2. Skill scores of 3-monthly forecasts

Fig. 2 shows also the skill scores of the forecasts of three-monthly streamflow totals using downscaled POAMA M24. The proportions of seasons×catchments with $NSE_{ref} > 0$, $NSE_{ref} \geq 0.2$ and $R^2 > 0.355$ are 83.3%, 69.4% and 25.0% respectively, while the means of $NSE_{ref} > 0$, SSRMSEP and SSCRPS are 0.35, 0.28 and 0.35 respectively. The ratio of Alpha≥0.05 is around 33.3%. It can be seen clearly that the skill scores of three-monthly streamflow forecasts are all lower than that of monthly streamlfow forecast except for the reliability based on Alpha, indicating that three-monthly forecasts have lower accuracy but higher reliability than monthly streamflow forecasts. However, as shown in Fig.4, it is similar to monthly streamflow forecast that the forecast median appears to be consistent with the observed value for Biggara, but the forecasts tend to be bias for wet seasons in Picnic Crossing and Tinderry.

In Fig.2, we can also find the difference of skill scores of 3-monthly streamflow forecasts among catchments and seasons. For the three catchments, Biggara has the higher R^2 value than the other two catchments, but the ensemble forecasts of Tinderry perform best in terms of NSE_{ref}, SSRMSEP and SSCRPS, followed by that of

Biggara. The improvements of skill scores in Tinderry by using ESP over climatology mean are larger than the other two catchments. For Biggara at the SMD region, the mean NSE_{ref}, R^2, SSRMSEP and SSCRPS of all seasons are 0.408, 0.373, 0.268 and 0.215 respectively, but the season ASO has negative NSE_{ref} and SSRMSEP. In term of Alpha, the seasons with higher reliability of forecast at Biggara are JFM, JJA, NDJ and DJF. For Picnic Crossing at QLD region (Fig.2), the three-monthly forecasts in Oct. to Apr. all have negative NSE_{ref} values and relative lower value of R^2, SSRMSEP and SSCRPS, indicating a poorer forecast than climatology mean. The seasons of JJA and JAS have rather high scores of NSE_{ref}, R^2, SSRMSPES and SSCRPS, but not for the reliability indicator Alpha. Those seasons with reliable forecast are MAM, AMJ and OND. For Tinderry at SEC region, R^2 of all seasons are lower than 0.355, implying that the correlation between forecasts and observed is not significant. However, NSE_{ref}, SSRMSEP and SSCRPS of most months are positive, meaning that the ensemble forecast median is better than climatology mean. The reliable three-monthly streamflow forecasts in Tinderry are that of Mar. to Aug. and NDJ.

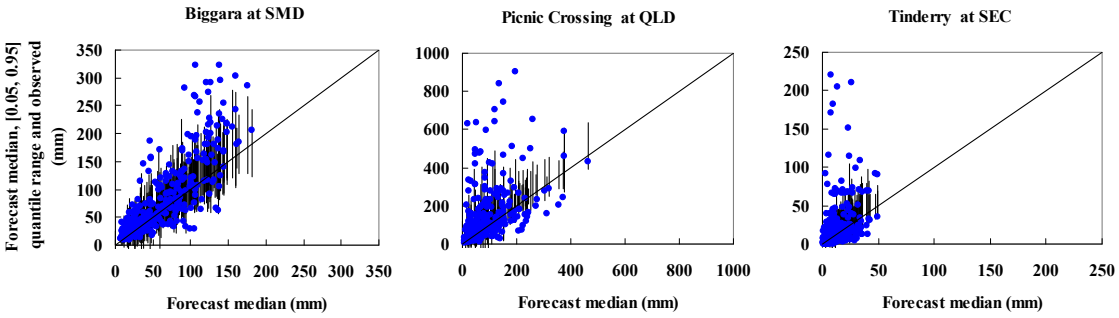

Figure 4 Three-monthly streamflow forecast quantiles and observed value plotted according to forecast median (M24-E33)

5. CONCLUSIONS

In this paper, the downscaled POAMA M24 rainfall is used as input for the GR4J model to forecast the monthly and three-monthly total streamflows in three selected catchments covering different hydrological regions. The model is run on the period 1980 to 2008 in the mode of cross-validation with 200 parameter sets derived through BATEA framework. The accuracy and reliability of the forecast are evaluated based on skill scores including NSE_{ref}, R^2, SSRMSEP, SSCRPS and reliability in Alpha.

The results show that ensemble streamflow forecasting approach based on conceptual rainfall-runoff model with forcings from downscaled POAMA M24 provides a potential way for monthly and three-monthly streamflow forecasting. The skill scores of the ensemble forecasting are better than forecast based on climatology mean. However, the forecasting skills are both catchment and month/season dependent. It is also found that the skill scores of monthly steramflow forecasts are higher than that of three-monthly forecasts except for the reliability of the forecasts. The comparison of forecast scheme using different downscaled POAMA M24 ensemble numbers shows that additional forcings with lead time 1 and 2 month do not improve the forecasting skill of monthly streamflow.

ACKNOWLEDGEMENT

We gratefully acknowledge the funding support from the Water Information Research and Development Alliance (WIRADA), which is a strategic investment between CSIRO and the Australian Bureau of Meteorology.

REFERENCES

Alves, O. and Hendon, H. (2003), POAMA – Seasonal prediction including ACCESS plans http://cawcr.gov.au/bmrc/basic/wksp18/papers/Alves.pdf

Bouttier, F. (1994), Sur la pre´vision de la qualite´ des pre´visions me´te´orologiques, Ph.D. thesis, 240 pp., Univ. Paul Sabatier, Toulouse, France.

Brown, T. A. (1974), Admissible scoring systems for continuous distributions, Manuscr. P-5235, 22 pp., RAND Corp., Santa Monica, Calif.

Day, G.N. (1985), Extended streamflow forecasting using NWSRFS. J. Waer Resour. Plann. Manage. Div.

Am. Soc. Civ. Eng,, 111 (2), 157-170.

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007), Probabilistic forecasts, calibration and sharpness, J. R. Stat.Soc.,Ser.B,69,243–268, doi:10.1111/j.1467-9868.2007.00587.x.

Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15(5), 559–570, doi:10.1175/1520-0434 (2000)015<0559:DOTCRP>2.0.CO;2.

Laio, F., and Tamea, S.(2007), Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11(4), 1267–1277.

Matheson, J.E., and Winkler R.L. (1976), Scoring rules for continuous probability distributions, Manage. Sci., 22, 1087– 1095, doi:10.1287/ mnsc.22.10.1087.

Perrin, C., Michel, C. and Andréassian, V. (2003), Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology 279(1-4): 275-289

Raupach, M.R., Briggs,P.R. Haverd,V., King,E.A., Paget, M. and Trudinger, C.M. (2009), Australian Water Availability Project (AWAP): CSIRO Marine and Atmospheric Research Component: Final Report for Phase 3, CAWCR Technical Report No. 013, Bureau of Meteorology.

Shao, Q. and Li, M. (2013), Assessment and Development of Bias correction in GCM Downscaling Procedure. Stochastic Environmental Research and Risk Assessment (accepted).

Thyer, M., Renard,B., Kavetski,D., Kuczera,G., Franks, S.W. and Srikanthan, S. (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.

Timbal B, Li Z, Fernadez E (2008) The Bureau of Meteorology Statistical Downscaling Model Graphical User Interface: user manual and software documentation. CAWCR Technical Report No. 004. Bureau of Meteorology, Australia.

Tuteja, N.K., Shin, DH, Laugesen, R., Khan, U., MacDonald, A., Le, B., Chia, T., Shao, Q., Wang, E., Li, M., Zheng, H., Kuczera, G., Kavetski, D., Evin, G., and Thyer, M., (2011), Experimental Evaluation of the Dynamic Seasonal Streamflow Forecasting Approach

Wang, E., Zhang, Y., Luo,J. Chiew, F.H.S. and Wand, Q.J. (2011a), Monthly and seasonal streamflow forecasts using Rainfall-runoff modeling and historical weather data, Water Resour. Res., 47, W05516, doi:10.1029/2010WR009922.

Wang, E., Zheng H, Shao Q, Wang Q.J. (2011b), Skill improvement through conditional model parameterisation and bias correction in seasonal streamflow forecasting. WIRADA Symposium

Wang, Q.J., Robertson, D.E. and Chiew, F.H.S. (2009), A Bayesian Joint Probability Modelling Approach for Seasonal Forecasting of Streamflows at Multiple Sites, Water Resources Research, 45, W05407, doi:10.1029/2008WR007355

Wood, A.W., Kumar, A. and Lettenmaier, D.P. (2005), A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States, Journal of Geophysical Research, 110, D04105, doi:10.1029/2004JD004508.