18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Comparison of multi-model and multi-donor ensembles for regionalisation of runoff generation using five lumped rainfall-runoff models

Viney, N.R. [1], **J. Vaze** [1], **F.H.S. Chiew** [1], **J. Perraud** [1], **D.A. Post** [1] **and J. Teng** [1]

[1] *CSIRO Water for a Healthy Country National Research Flagship,CSIRO Land and Water, Canberra, ACT, Australia*
Email: neil.viney@csiro.au

Abstract: Five lumped, conceptual rainfall-runoff models are calibrated for 240 gauged catchments in southeastern Australia. Climate input to the models is distributed at ~25 km^2 grid cells and the catchments range in size from 50 to 2000 km^2. Each of the models is calibrated on each of the 240 catchments. Each catchment is then simulated using parameters sets calibrated for the nearest neighbouring catchment. Model predictions are assessed using the daily Nash-Sutcliffe efficiency and the volume bias.

The results demonstrate that whilst an increasing number of optimisable parameters leads to increased calibration performance (when assessed using metrics based on the sum of squared residuals), the reverse is true for a large proportion of catchments in cross-verification using parameters from one donor catchment. This reversal, however, does not persist for the multi-donor averages, where the more highly parameterised models typically have the best performance.

A weighted average of the five models (weighted by calibration performance) is shown to yield better calibration predictions than an unweighted average, but in cross-verification there is little difference between the two. This suggests that the relative calibration performances of different models in a donor catchment are not necessarily good indicators of how well the models will contribute to prediction in a neighbouring catchment.

Five-member multi-donor ensembles of each individual model, weighted by distance, are superior to using a raw average, and both unweighted and weighted multi-donor ensembles are superior to the respective single-donor models. This indicates that while there is useful information delivered to an ensemble by the fifth nearest catchment, the value of this information is not as significant as the information from the nearest catchment.

Further investigation using a multi-donor ensemble approach indicates that the optimum number of catchments to include in a spatial ensemble is five or six, and that such an ensemble can lead to considerable improvement in runoff predictions in ungauged catchments. We show that for the set of models and catchments used, the multi-donor approach using a single rainfall-runoff model is superior to the multi-model approach, but that a combination of the two approaches yields the best overall predictions.

A five-member unweighted multi-model ensemble is shown to give regionalised predictions that are commensurate with the typical five-member unweighted multi-donor ensemble, but when the multi-model ensemble is weighted by donor calibration performance, its predictions are poorer than each of the multi-donor models weighted by distance from the target catchment. Nonetheless, the best predictions assessed in this study are those of a multi-model multi-donor ensemble that combines the weighted averaging methods of both combinations.

Keywords: *Rainfall-runoff modelling, ensemble modelling, model averaging, regionalisation, ungauged basins*

1. INTRODUCTION

Regionalisation, the process of transferring catchment modelling information and predictions to ungauged catchments, is becoming an increasingly prominent topic in catchment modelling. Some of the approaches to regionalisation are discussed and assessed by Oudin et al. (2008) and Zhang and Chiew (2009).

One method is to transfer calibrated model parameters from a nearby gauged catchment (e.g., Oudin et al., 2008; Chiew et al., 2009). The key assumption implicit in this approach is that catchments in close proximity are likely to share similar soils, topography, land cover and climate and that they therefore have similar hydrological response characteristics. Model parameters calibrated for one catchment are therefore likely to predict streamflows reasonably well on the second.

A complementary modelling approach that has potential to reduce uncertainty in ungauged catchment predictions is the use of ensemble techniques, whereby predictions from different sources are pooled to produce a consensus prediction (e.g., Ajami et al., 2006; Viney et al., 2009a). Ensembles may be constructed from different realisations of the same rainfall-runoff model (a single-model ensemble) or from several different models (a multi-model ensemble). Several researchers have reported that the optimal number of members for both multi-model and multi-donor ensembles is about five (e.g., Ajami et al., 2006; Reichl et al., 2007; Viney et al., 2008; Viney et al., 2009a), although Zhang and Chiew (2009) suggest eight to ten.

In this paper, we compare predictions from five rainfall-runoff models. We then assess the effectiveness of multi-model ensembles (MMEs) constructed from the predictions of the five rainfall-runoff models. Finally, we examine the potential for combining cross-verification predictions from several catchments to give multi-donor (or multi-catchment) ensembles (MDEs), which are a form of single-model ensemble. Given the optimal number of ensemble members appears to be about five, one of the aims of the paper is to assess whether a five-member MME performs better in cross-verification than a five-member MDE.

2. STUDY AREA AND DATA

This study uses observed streamflow data from 240 gauged catchments in southeastern Australia (Figure 1). Most are located on the southern or eastern edges of the Murray-Darling Basin or in adjacent coastal regions. The catchments have areas ranging between 50 km^2 and 2000 km^2 and their streamflows are not affected by impoundment or significant irrigation withdrawal. All catchments have streamflow records that are at least 75 % complete during the period 1975 to 2006. All available non-nested catchments in the study area that meet these criteria have been chosen. Within the study region, mean annual precipitation varies from less than 300 mm in the west to more than 1500 mm in the southeast, and is summer-dominated in the north and winter-dominated in the south (Chiew *et al.*, 2008). For the 240 study catchments, mean annual streamflow varies from less than 2 mm to more than 1400 mm, and runoff coefficients range from less than 1 % to more than 90 %.

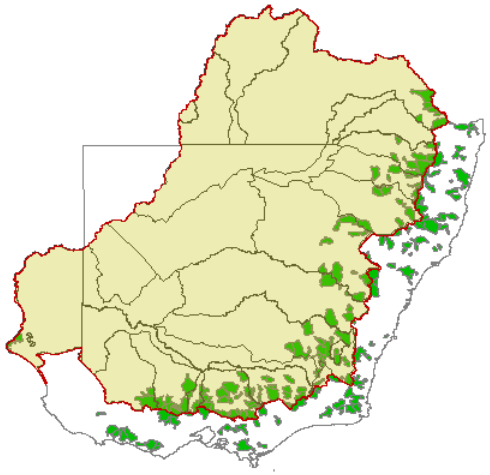

Figure 1. Location on the 240 study catchments (green) and the Murray-Darling Basin (red).

Daily rainfall input data is obtained from the Silo Data Drill (Jeffrey et al., 2001), a data set gridded at a 0.05° (~5 km) spacing. The Data Drill rainfall data is interpolated from point observations of daily rainfall. Areal potential evaporation data is also derived from the Data Drill.

3. METHODS

Five lumped, conceptual, daily rainfall-runoff models are calibrated separately on each of the 240 catchments: AWBM (Boughton, 2004), IHACRES (Croke et al., 2006), Sacramento (Burnash et al., 1973), Simhyd (Chiew et al., 2002) and SMAR-G (Goswami et al., 2002). All models have previously been applied widely in runoff modelling. In this study, six model parameters are optimised for Simhyd, including one parameter in a Muskingum routing algorithm (Tan et al., 2005). For the implementation of the remaining models, we optimise six parameters for AWBM, seven for IHACRES, 13 for Sacramento and eight for SMAR-G.

Each model is operated using the gridded rainfall and potential evaporation data in 0.05° x 0.05° grid cells across each catchment. For calibration, the observed runoff at the catchment outlet is compared with a spatial average of the modelled runoff in each grid cell within the catchment.

Calibration is achieved through a sequential combination of the shuffled complex evolution algorithm (~10000 model runs) and Rosenbrock (~300 model runs) methods. Tests of this procedure have shown it to provide reproducible results for the five models. The objective function is a weighted combination of the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) and a logarithmic bias constraint (Viney et al., 2009b) and is given by

$$F = E - 5 \left| \ln(1 + B) \right|^{2.5}$$

where E is the daily Nash-Sutcliffe efficiency and B is the bias (total prediction error divided by total of observations). The coefficients (5 and 2.5) are chosen from trial and error to provide a reasonable weighting between efficiency and bias for calibration. The value of F ranges from 1.0 (a perfect fit) to minus infinity.

Cross-verification is achieved by modelling each catchment using its local climate data but with parameters taken from the nearest neighbouring gauged catchment. The distance to the nearest neighbour (measured from centroid to centroid) ranges from 7 km to 65 km. This cross-verification procedure thus gives an indication of the likely quality of ungauged basin predictions that could be achieved if proximity was used as the sole regionalisation criterion.

Two different multi-model ensembles are assessed. In the first one, we construct a time series of streamflows from the mean of the five daily model predictions. In the second, we apply a weighting that ensures that the best calibrated model has greater weight in the five-model average. For each model, the weighting is proportional to $1/(1 - F)^2$. Here, calibration quality is assessed in terms of the same objective function that is used to calibrate the models. Where a catchment is assessed in cross-verification, we use the model weightings of the donor catchment, not the target catchment.

Finally, we assess the usefulness of multi-donor ensembles. Here, instead of cross-verifying using parameters from the nearest neighbour, we use parameters from the nearest five neighbours and either average the resulting five streamflow time series, or weight them by the inverse of the square of their distance from the target catchment.

4. RESULTS

4.1. Model calibration

Calibration efficiencies and biases for the five models are shown in Figures 2 and 3, respectively. For 77 % of the 240 catchments, the Sacramento model has the best efficiencies. In the main, these tend to be the catchments that are calibrated well by all models. However, for the catchments that are more difficult to model (typically those depicted on the left of Figure 2), Sacramento's calibration performance degrades noticeably with respect to the other four models. For the poorest 10 % of catchments, IHACRES has the best efficiencies. Despite failing to sustain its relative dominance to the right side of Figure 2, IHACRES has by far the lowest absolute biases (Figure 3). For 76 % of catchments, IHACRES' absolute bias is less than 1 % and all its absolute biases are less than 6 %. Sacramento's absolute biases are also low, except for its worst 10 % of catchments, where it has a strong tendency towards underprediction.

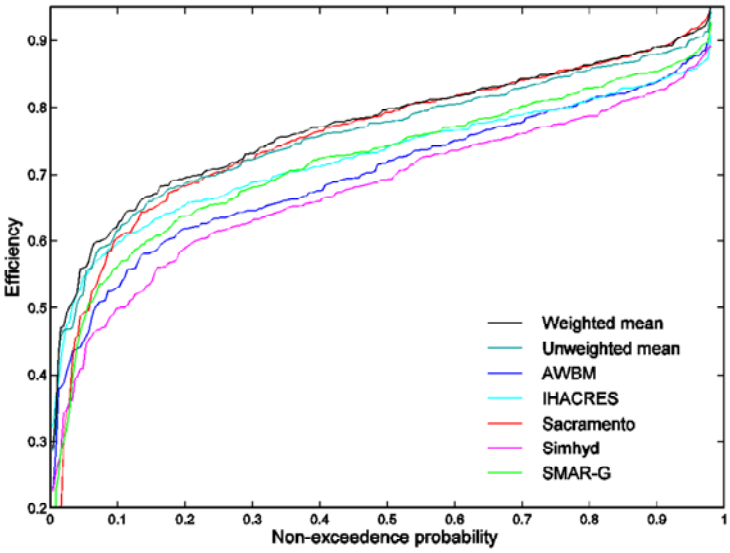

Figure 2. Cumulative distribution of calibration efficiency for the five models and two multi-model averages.

Also shown in Figures 2 and 3 are the performances of two multi-model ensembles. The unweighted mean MME is constructed by using the raw

mean of the five models daily predictions. Its efficiencies are as good as or better than the best of the individual models for the worst 25 % of catchments, but thereafter it is slightly poorer than Sacramento, although significantly better than the remaining models. Its biases are, by definition, the mean of the biases of the five models. As such they are poorer than those of IHACRES and Sacramento, but better than the other three models for the best 85 % of catchments, but thereafter, under the dominance of Sacramento's extremely poor predictions they are worse than most of the other models at the right of Figure 3. The weighted mean MME has efficiencies that are as good as or better than those of the best model (IHACRES or Sacramento), and always better than the unweighted mean. Its biases are also better than those of the unweighted mean, but remain significantly poorer than IHACRES and (for most catchments) Sacramento.

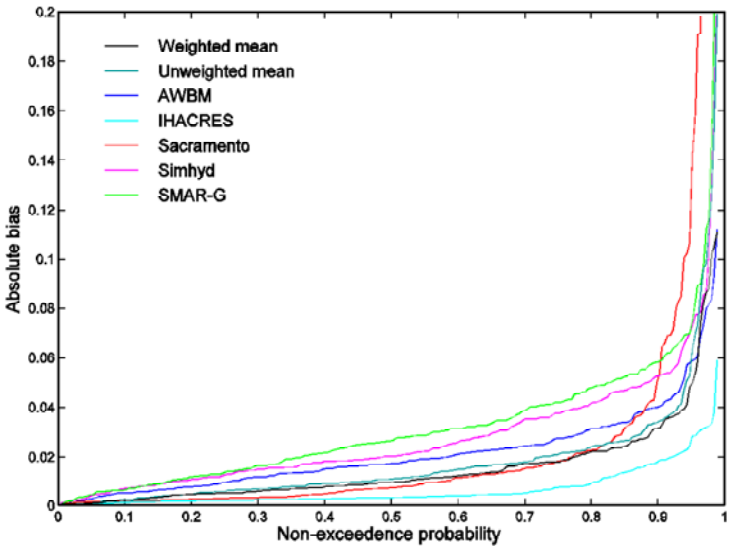

Figure 3. Cumulative distribution of absolute calibration bias for the five models and two multi-model averages.

4.2. Cross-verification with nearest neighbour

We assess cross-verification performance with the objective function, F. Although F has no formal function in cross-verification, it nonetheless provides a suitable measure for integrating the effects of efficiency and bias. Values of F for cross-verification using parameters from the nearest neighbour are shown in Figure 4. For the best 50 % of catchments, Sacramento has the best cross-verifications, while elsewhere, AWBM and IHACRES are best. A weighted mean MME (weighted by $1/(1 - F)^2$) made up of the daily cross-verification predictions of all five models generally performs better than the best individual model. The cumulative probability curve for this weighted mean MME is virtually indistinguishable from the corresponding unweighted MME (not shown).

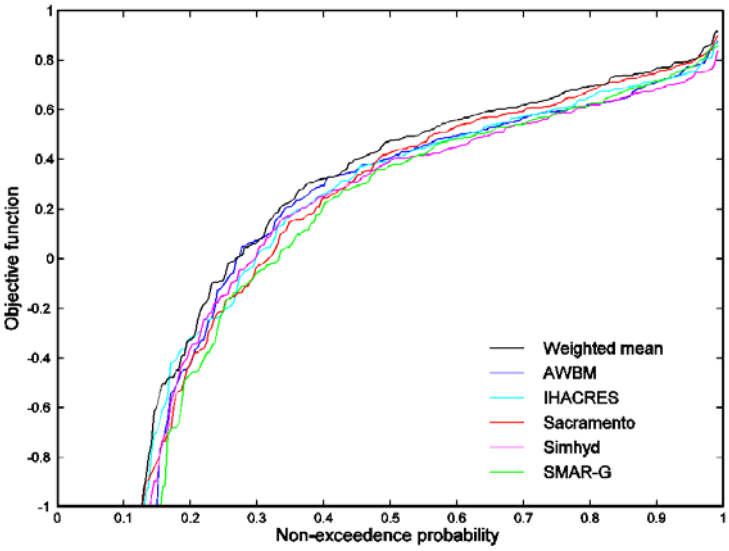

Figure 4. Cumulative distribution of cross-verification objective function for the five models and a multi-model average.

4.3. Multi-donor ensembles

All the MDEs assessed here include predictions using the calibrated parameters from the five nearest neighbouring catchments. Figure 5 shows values of F when the raw mean of the predictions from the five catchments are used for each model. Also shown, to enable comparison with Figure 4, is the weighted mean MME for the nearest neighbour (the black line in both figures). For all models, F is improved (by about 0.07 at the median) when five neighbours are used. Whereas the nearest neighbour weighted MME is better than all the individual models in Figure 4, it is commensurate with the median of the five models' F values in Figure 5 (it exceeds the median in 52 % of catchments).

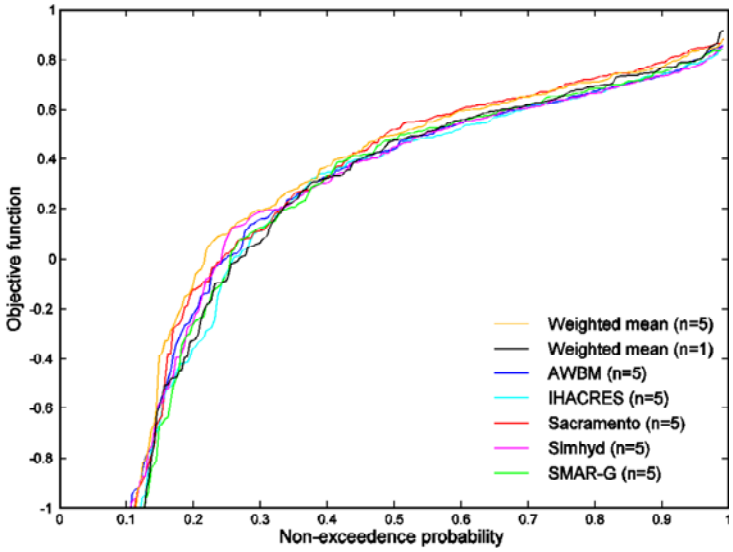

Figure 5. Cumulative distribution of objective function for unweighted multi-donor ensembles of the five models, a single-donor multi-model average and a multi-donor multi-model average, where *n* is the number of donors.

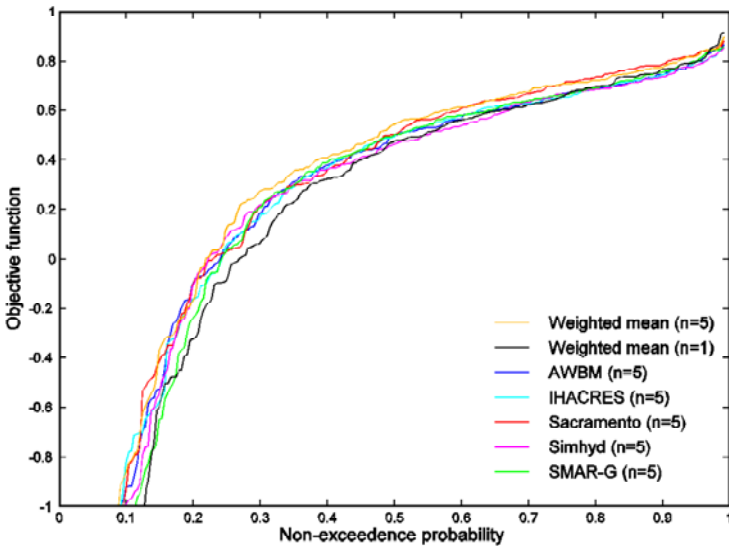

Figure 6. Cumulative distribution of objective function for weighted multi-donor ensembles of the five models, a single-donor multi-model average and a multi-donor multi-model average, where *n* is the number of donors.

Figure 6 shows *F* values for distance-weighted five-donor MDEs for each model. The weighting that each donor contributes to the MDE is inversely proportional to the square of its distance from the target catchment. Again, the weighted mean MME is shown for comparison, and is it is seen that distance weighting leads to further slight improvement in *F* (by about 0.02 at the median) for the MDEs. The MDEs of the individual models are now generally better than the single-donor weighted MME. Overall, the best predictions are for a multi-model multi-donor ensemble, where both models and donors are weighted.

5. DISCUSSION

In general, the Sacramento model has better calibration performance than the other models. In large part, this is no doubt due to the larger number of parameters available for optimisation. Indeed the calibration performance for efficiency shown in Figure 2 correlates well with the number of optimisable parameters in each model, with the six-parameter models AWBM and Simhyd having generally poorer efficiencies than the seven- and eight-parameter models, IHACRES and SMAR-G, respectively. IHACRES has the best bias characteristics in calibration, with no catchment having an absolute bias of more than 6 %. This is most likely due to the presence in IHACRES of a parameter that effectively scales rainfall, thus enabling it to be calibrated with ease to almost any data set.

The advantages of having a larger number of optimisable parameters do not appear to translate so well to cross-verification (Figure 4). Although Sacramento generally has the best predictions for the best 50 % of catchments, and the ordering of model performance is approximately similar to that of calibration, for the rest of the catchments the situation is reversed. Here the less parameterised AWBM and Simhyd are among the best, while the highly parameterised Sacramento and SMAR-G are among the worst. A possible reason is that with a larger number of parameters, there is a greater possibility that some parameters become too site specific. This reversal, however, does not persist for the multi-donor ensembles, where Sacramento typically has the best performance. The averaging of output from several donors appears to also average out the site-specificity with single donor predictions.

In calibration, an unweighted MME has efficiencies that are typically slightly lower than those of Sacramento, but greater than those of its other constituent models. Its absolute biases, though, are worse than those of IHACRES and mostly worse than those of Sacramento. When this MME is weighted by the calibration objective function, its efficiencies improve to the extent that they generally outperform the best individual model and while its absolute bias improves upon the unweighted MME, it remains inferior to IHACRES and Sacramento for the majority of catchments. This suggest that there may be slight advantages

to averaging output from the five models in calibration (especially if the MME is weighted) over the alternative of just choosing the best-calibrated individual model.

In contrast, the option of choosing the best individual model is not available in ungauged basin studies. The nearest feasible single-model option is to choose the model that calibrates best on the donor catchments. As we have seen above, this will usually be Sacramento. However, as we have also seen, Sacramento becomes one of the worst models for 50 % of the catchments in cross-verification where parameters from the single nearest neighbour are used. The MMEs (both unweighted and weighted) now have a clear advantage over any individual model, especially the model that may have been chosen *a priori* (i.e., Sacramento). Interestingly, while weighting of the MME improves predictions in calibration, there is little discernible difference between the unweighted and weighted MMEs in cross-verification, with the easier-to-implement unweighted MME performing just as well as the weighted MME. This result contradicts the observations of Viney et al. (2008) for a two-member MME, and suggests that the relative calibration performances of different models in a donor catchment are not necessarily good indicators of how well the models will contribute to prediction in a neighbouring catchment.

The use of donor parameters from five nearest neighbours in single-model MDEs improves the predictions of all models. However, given the apparent narrowing of the range of model performances at each quantile, multi-donor averaging appears to improve the predictions of the poorer models more than it improves the predictions of the better models. The reasons for this are not clear. Weighting of the MDEs by distance improves predictions even further for all models. This contrasts with the observation that weighting of MMEs in cross-verification fails to improve on the unweighted MMEs.

In determining which is better in cross-verification—a five-member MME or a five-member MDE—we can make two comparisons. The first is for unweighted averages. We can compare the unweighted MME (which is essentially the same as the weighted MME with $n = 1$ in Figure 5) with the five unweighted model MDEs in Figure 5. The MME is close to the median. This indicates that its predictions are similar to those that might be expected from a single unweighted MDE chosen at random from the five models. The second comparison is for weighted averages. In the case of the MME, weighting is by donor calibration performance, while for the MDEs weighting is by distance from the target catchment. This comparison is between the weighted MME (with $n = 1$) and the weighted MDEs in Figure 6. Here it becomes apparent that throughout most of the range of objective functions the MME is inferior to all the MDEs. This suggests that there may be a greater diversity of information content when five donor catchments are used with a single rainfall-runoff model—regardless of which model that is—than when five models are used from a single donor catchment.

However, despite the foregoing, the best predictions are for a weighted multi-model multi-donor ensemble. This ensemble is constructed using as many as 25 parameter sets, which is vastly more than the optimum ensemble size reported in other studies.

6. CONCLUSIONS

Calibration and cross-verification of five lumped rainfall runoff models on 240 catchments in southeastern Australia have shown that the relative calibration performance of the five models does not necessarily persist in regionalisation where calibrated parameters from the nearest catchment are used. Whilst an increasing number of optimisable parameters leads to increased calibration performance, the reverse is true for a large proportion of catchments in cross-verification when parameters from one donor catchment are used. Averaging of donors, however, overcomes this reversal.

A weighted average of the five models (weighted by calibration performance) is shown to yield better calibration predictions than an unweighted average, but in cross-verification there is little difference between the two. This suggests that the relative calibration performances of different models in a donor catchment are not necessarily good indicators of how well the models will contribute to ensemble prediction in a neighbouring catchment.

For five-member multi-donor ensembles of each individual model, weighting by distance is superior to using a raw average, and both are superior to the respective single-donor models. This indicates that while there is useful information delivered to an ensemble by the fifth nearest catchment, the value of this information is not as significant as that from the nearest catchment.

A five-member unweighted multi-model ensemble is shown to give regionalised predictions that are commensurate with the typical five-member unweighted multi-donor ensemble, but when the multi-model ensemble is weighted by donor calibration performance, its predictions are poorer than each of the multi-

donor models weighted by distance from the target catchment. Nonetheless, the best predictions assessed in this study are those of a multi-model multi-donor ensemble that combines the weighted averaging methods of both combinations.

ACKNOWLEDGMENTS

This work is part of the water information research and development alliance between CSIRO's Water for a Healthy Country Flagship and the Australian Bureau of Meteorology.

REFERENCES

Ajami, N.K., Q. Duan, X. Gao and S. Sorooshian (2006). Multimodel combination techniques for analysis of hydrological simulations: application to Distributed Model Intercomparison Project results. *Journal of Hydrometeorology*, 7, 755–768.

Boughton, W.C., (2004). The Australian water balance model, *Environmental Modelling and Software*, 19, 943–956.

Burnash, R.J.C., R.L. Ferral and R.A. McGuire (1973), A generalized streamflow simulation system—conceptual modeling for digital computers. Tech. Rep., Joint Federal and State River Forecast Center, Sacramento, 204pp.

Chiew, F.H.S., M.C. Peel, and A.W. Western (2002), Application and testing of the simple rainfall-runoff model SIMHYD. In Singh, V.P. and Frevert, D.K. (eds.), Mathematical models of small watershed hydrology and applications, Water Resources Publications, Littleton, USA, pp. 335–367.

Chiew, F.H.S., J. Teng, J. Vaze, D.A. Post, J. Perraud, D. Kirono and N.R. Viney (2009). Estimating climate change impact on runoff across south-east Australia: method, results and implications of modelling method. *Water Resources Research* (in press).

Croke, B.F.W., F. Andrews, A.J. Jakeman, S.M. Cuddy and A. Luddy (2006). IHACRES Classic Plus: A redesign of the IHACRES rainfall-runoff model. *Environmental Modelling and Software*, 21, 426–427.

Goswami, M., K.M. O'Connor and A.Y. Shamseldin (2002). Structures and performances of five rainfall-runoff models for continuous river-flow simulation. Proceedings of 1st Biennial Meeting of International Environmental Modeling and Software Society, Lugano, Switzerland, 1, 476–481.

Jeffrey, S.J., J.O. Carter, K.B. Moodie and A.R. Beswick (2001), Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling and Software*, 16, 309–330.

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

Oudin, L., V. Andréassian, C. Perrin, C. Michel and N. Le Moine (2008). Spatial proximity, physical similarity and ungaged catchments: a comparison based on 913 French catchments. *Water Resources Research*, 44, W03413.

Reichl, J.P.C., A.W. Western and F.H.S. Chiew (2007). Developing similarity measures for predicting ungauged streamflow within a model averaging framework. Congress on Modelling and Simulation (MODSIM 2007), Christchurch, New Zealand, pp. 2480–2486.

Tan, K.S., F.H.S. Chiew, R.B. Grayson, P.J. Scanlon and L. Siriwardena (2005), Calibration of a daily rainfall-runoff model to estimate high daily flows. Congress on Modelling and Simulation (MODSIM 2005), Melbourne, Australia, pp. 2960–2966.

Viney, N.R., J. Vaze, F.H.S. Chiew and J. Perraud (2008). Regionalisation of runoff generation across the Murray-Darling Basin using an ensemble of two rainfall-runoff models. Water DownUnder 2008 conference, Adelaide, Australia, 1700–1711.

Viney, N.R., H. Bormann, L. Breuer, A. Bronstert, B.F.W. Croke, H. Frede, T. Gräff, L. Hubrechts, J.A. Huisman, A.J. Jakeman, G.W. Kite, J. Lanini, G. Leavesley, D.P. Lettenmaier, G. Lindström, J. Seibert, M. Sivapalan and P. Willems (2009a). Assessing the impact of land use change on hydrology by ensemble prediction (LUCHEM). II: Ensemble combinations and predictions. *Advances in Water Resources*, 32, 147–158.

Viney, N.R., J. Perraud, J. Vaze F.H.S. Chiew, D.A. Post and A. Yang (2009b). The usefulness of bias constraints in model calibration for regionalisation to ungauged catchments. Proceedings, MODSIM 2009 (this volume).

Zhang, Y. and F.H.S. Chiew (2009). Relative merits of different methods for runoff predictions in ungauged catchments. *Water Resources Research* (in press).