21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Why do sub-period consistency calibrations outperform traditional optimisations in streamflow prediction?

S.S.H. Kim [a], J.D. Hughes [a], D. Dutta [a] and J. Vaze [a]

[a] CSIRO Land and Water
Email: shaun.kim@csiro.au

Abstract: A previous study showed that a calibration method that utilises the distribution of sub-period performances routinely performs better for prediction than traditional optimisations. Kim et al. (Determining probability distributions of parameter performances for time-series model calibration: a river system trial, *Journal of Hydrology*, in press) describes a new method that uses sub-period resampling to estimate probability distributions of performance for different parameter sets (Figure 1). The method is designed to identify more time-consistent (and therefore more robust) parameterisations than the traditional split-sampling optimal parameterisations. However, the underlying reasons for the superior performance are not fully understood. Several hypotheses have been proposed but as yet none have been properly verified. There are two key steps of the sub-period consistency calibration which are thought to be important for better predictions: (1) the sampling of sub-period performances; and (2) using the distribution of performances to formulate a weighted score of predictive capability. It is assumed that the sub-period consistency calibration returns parameterisations that spread consistent-sized errors throughout the calibration period. The calibrations are therefore not over-influenced by rare periods of good fit. Subsequently, they do not contain as many periods of poor fit and are less susceptible to over-fitting. Also, consistently performing parameterisations might fit to more frequent flow regimes, which might help gain better validation performances if calibration and validation periods have similar hydrological characteristics. This paper aims to investigate which factor is more important for improved predictions. Different sub-period sampling lengths are tested to see whether this makes an impact on the performance of the method and whether this is related to cyclical patterns in the residual error time-series examined in Fourier transform analyses. Three hydrological models are used in the study: AWRA-R, GR4J and Sacramento. Each model is tested using data from about 80 river gauging sites across Australia.

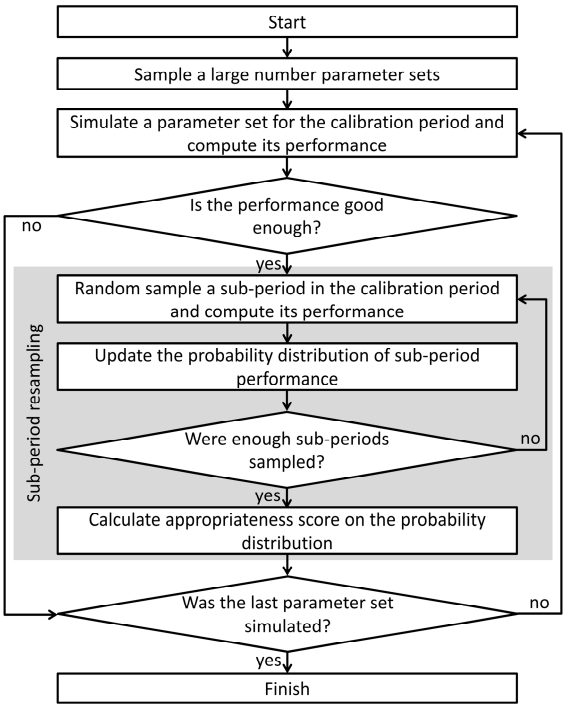

Figure 1. Flowchart of the sub-period consistency calibration method.

Keywords: *Sub-period, consistency, calibration, prediction, validation, time-series modelling*

1. INTRODUCTION

Conceptual models are commonly used in hydrology due to their practicality and fast simulation times. However, calibration of such models can be challenging particularly when parameters are difficult or impossible to measure in the field or laboratory. In hydrology, observed streamflow is commonly used to calibrate against. Calibration generally requires one or more objective functions, which are user-defined metrics that indicate how well the agreement is between the simulation and the observed response data (from here on referred to as 'performance').

Model validation is the process of using the calibrated model parameters to simulate the variable of interest over an independent period and calculating its performance. Good validation results provide some confidence that the selected models and parameters are appropriate for use for impact assessments, design, water management and forecasting purposes.

It has been known for a long time that often very different parameter sets and model structures can seemingly be equally acceptable 'real' system representations (Wagener, 2003). Also, while it might be easy to obtain acceptable performance during calibration, achieving decent validation performance is much more challenging (Beven, 2006; Kavetski et al., 2002). It is generally accepted that time periods with similar characteristics should perform similarly (Seibert, 2003). Gharari et al. (2013) explain that calibration parameters are inherently linked to the calibration time period and therefore may be inadequate to represent other periods. They provide a summary of the many causes for the decrease of model performance in non-calibration periods. Non-stationarity (i.e. differences in climate and/or changes in land-use) (Koutsoyiannis et al., 2009; Kuczera et al., 1993), poor model conceptual structure (Beven, 2006), over-fitting (Jakeman and Hornberger, 1993), and poor data quality (Bárdossy and Singh, 2008) are commonly blamed. Stochasticity is a particularly notable cause since it is characterised by significant variability in observed responses (Freer et al., 2003; Koutsoyiannis et al., 2009), and implies statistical approaches are required to deal with these systems.

In an attempt to utilise more information in calibration periods, a method called sub-period consistency (SPC) calibration has been developed to identify paramaterisations that give better validation performance than optimal ones (Kim et al., in press). An 'optimal parameterisation', in the current context, refers to the best objective function scoring parameter set during calibration. Using this will also be referred to as the 'traditional' approach in the current study. SPC calibration explores parameterisations to determine those that will perform better during validation. For each parameter set, the method resamples time blocks (sub-periods) from the simulation with replacement and computes the objective function for each block. Then the distribution of sub-period performances are used to formulate a weighted score of predictive capability, called appropriateness. The best appropriateness scoring parameterisation is tested in a validation period set aside earlier.

Kim et al. (in press) used the Australian Water Resource Assessment River (AWRA-R) model to show validation scores could be significantly improved using SPC calibration compared to traditional optimisation. However, the underlying reasons for this are not fully understood. One hypothesis is that the SPC calibration returns parameterisations that are not over-influenced by rare periods of good fit, and subsequently are less susceptible to over-fitting. Related to this is that perhaps SPC parameterisations will have more weighting towards fitting at more frequent flow magnitudes rather than rare events, which could be beneficial if the calibration period has similar hydrological characteristics to the validation period. In addition, it is thought that different objective functions and sub-period lengths are likely to influence results. The aim of the current study is to explore and examine some of the possible reasons for SPC calibration's success by testing different sub-periods and multiple hydrological models.

2. METHODS

2.1. Sub-period Consistency (SPC) Calibration

The sub-period consistency calibration procedure is shown in Figure 1. The objective function is given by:

$$\Omega = -1 - \frac{\sum_{i=1}^{N^Q}\left(\sqrt{Q_i^{sim}} - \sqrt{Q_i^{obs}}\right)^2}{\sum_{i=1}^{N^Q}\left(\sqrt{Q_i^{obs}} - \overline{\sqrt{Q^{obs}}}\right)^2} \quad (1)$$

where N^Q is the total number of observed flow data points, Q^{sim} and Q^{obs} are the simulated and observed flow, respectively. The function ranges from -infinity to -1, with -1 being perfect. The objective function used to determine performance is similar to the Nash-Sutcliffe Efficiency applied to square-root transformed flow (NSE^{root}), which has been found to provide an appropriate balance between low- and high-flow agreement (Oudin et al., 2006). The SPC calibration uses a sub-period resampling method to determine the probability distributions of obtaining specified performances for sampled sub-periods. For a parameter set, a single probability distribution is created of sub-period objective scores.

For parameter sampling, the Sobol sampling method was used (with 10,000 samples) to broadly and evenly cover the parameter space. For each sample parameter set, a simulation was performed for the calibration period and was deemed suitable for use in the sub-period resampling if it exceeded a performance threshold. Equation 1 was used to calculate performances. An acceptance threshold criterion of > -2 was used to determine which parameter sets were suitable for sub-period resampling. In previous trials, this threshold showed a good balance between large parameter coverage and short run time.

If a parameter set was deemed adequate, sub-periods were randomly sampled (with replacement) from the simulated outflow time-series within the calibration period. Performances were computed of the sub-period time-series against corresponding observed data using Equation 1 and used to update the probability distribution. 10,000 sub-period resamples were used for each parameter set's probability distribution since early trials found it took about this many to converge to a stable distribution.

After the probability distribution for a parameter set is complete, the appropriateness of the parameterisation for prediction was computed. The appropriateness function considers the densities at all performance ranges but puts more weight towards the better performance ranges in the probability distribution. This is designed to provide an indication of predictive performance. Appropriateness was calculated by:

$$A_n = \frac{\sum_{j=1}^{N^d} d_j \cdot w \cdot j}{N^d} \tag{2}$$

where j denotes the index of the performance range (or performance bin), d_j represents the density for performance range j, w is the width of the performance ranges (constant), and N^d is the total number of performance ranges. Higher j values correspond to better goodness of fit, so these performance bins will be favoured by the appropriateness function, subject to bin density. Figure 2 shows an example probability distribution and displays the variables of the appropriateness function. The final SPC calibration parameter set is the one that obtains the highest appropriateness value.

2.2. Investigation of performance

River gauges across Australia were used as the study sites. The gauge sites are spread across the Murray-Darling Basin, as well as the Flinders and Gilbert catchments in North Queensland. Streamflow records cover between 1970 and 2012.

Three specific hypotheses about the SPC calibration are investigated:

hypothesis 1. Sub-period sampling is crucial for favourable validation results.

hypothesis 2. The range of performance throughout the calibration period is reduced when using the SPC calibration.

hypothesis 3. The SPC calibration returns parameterisations that fit to more frequent flow magnitudes rather than rare events.

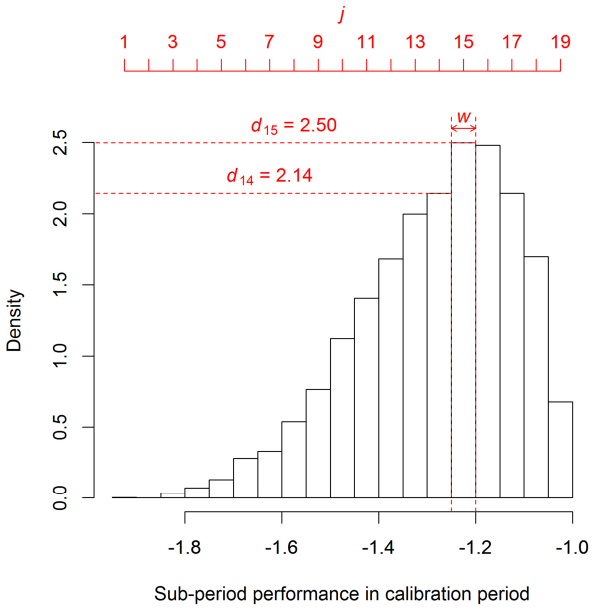

Figure 2. Illustration of the variables (shown in red) used to calculate appropriateness (as in Equation 2).

For each site, the first halves of the observed flow time-series were used for the calibration and the second halves for the validation. The SPC calibration was used with varying sub-period lengths. These included 1825 days (SPC-5yr), 730 days (SPC-2yr), 365 days (SPC-1yr), 183 (SPC-6mth) and 30 days (SPC-1mth).

To test hypothesis 1, calibrations were run that did not sample sub-periods, but instead calculated appropriateness on the distribution of each day's relative square error on square-root transformed flow (no sub-period). That is, the probability distribution was created using the series of values given by:

$$\mathbf{e} = \left(e_1, e_2, e_3, \ldots, e_{N^Q-1}, e_{N^Q}\right) \tag{3}$$

$$e_i = \frac{\left(\sqrt{Q_i^{sim}} - \sqrt{Q_i^{obs}}\right)^2}{Q_i^{obs} + 10^{-8}} \tag{4}$$

where e_i is the relative square error on square-root transformed flow for day, i, and $\mathbf{e}$ represents the complete series of e_i. The SPC calibration was performed for all sub-period length trials (including no sub-period) for the daily hydrological models, AWRA-R (Vaze et al., 2013), GR4J (Perrin et al., 2003) and Sacramento (Burnash et al., 1973). The optimal parameter sets (OPT) were taken as the best objective function scoring Sobol samples from the SPC calibration. The performances of the SPC calibrated parameters were compared to that of the OPT parameters for both calibration and validation periods for 78 gauge sites for AWRA-R, and 79 gauge sites for GR4J and Sacramento. The Wilcoxon signed-rank test was used to determine whether differences between SPC and OPT validation results were significant.

Fourier transform analyses were also performed on residual error time-series to examine any cyclical patterns in the error. This was to test the importance of time-scale, and whether there is a link between error periodicity and particular sub-period sample lengths in the SPC calibration. A significance test that assumes a simple univariate lag-1 autoregressive model for background noise (otherwise known as red noise) was used to determine significant spectral energy peaks (Zhang and Moore, 2011). To test the difference between energy peaks between SPC and OPT, the two-sided Wilcoxon rank-sum test was used. This was performed for each SPC sub-period length.

To test hypothesis 2, a statistic that assesses the consistency of agreement over specified time scales was used. This statistic requires that the data is divided into specified equal length blocks. NSE^{root} is calculated on each of the blocks:

$$\Lambda_b = 1 - \frac{\sum_{i=(b-1)\cdot L+1}^{b\cdot L}\left(\sqrt{Q_i^{sim}} - \sqrt{Q_i^{obs}}\right)^2}{\sum_{i=(b-1)\cdot L+1}^{b\cdot N^B}\left(\sqrt{Q_i^{obs}} - \overline{\sqrt{Q^{obs}}}\right)^2} \tag{5}$$

where b denotes the block index, and L is the length of the block. The largest b value (total number of blocks) is approximately N^Q/L. The statistic is finally the standard deviation of Λ, which is calculated for each site for both calibration and validation periods. Also, to test hypothesis 3, each block's mean flow was computed to determine whether the best agreement corresponded to the most frequent mean flow. For each site, a histogram of block mean flows was created. For each bin of the histogram, the mean NSE^{root} for blocks within the bin was computed. If the highest frequency bin is also the best mean NSE^{root} bin for SPC calibrations, this would provide some evidence that SPC calibrations fit to more frequent flow regimes.

3. RESULTS

The validation results are summarised in Table 1. SPC calibrated parameters performed significantly better in validation than optimal parameters for sub-period sample lengths of 2 and 5 years for both GR4J and Sacramento. Sacramento SPC-1yr validation scores also showed significant improvement compared to OPT results. For AWRA-R, it couldn't be shown that there was a statistically significant difference between OPT validation and SPC validation scores for any of the sub-period lengths chosen. Out of the tested sub-period sample lengths, the 2 year sample length showed the best validation for all hydrological models.

As the sub-period sample length increased, the SPC parameterisations more resembled those of OPT. For example, for AWRA-R's SPC-5yr, 45 out of 78 sites obtained identical parameterisations, where only 40 were identical for SPC-2yr and 26 for SPC-1yr.

The Fourier transform analysis was performed on the calibration period on all sites for each model. From basic observations, strong spectral energies were found around periods of 1 year and occasionally periods longer than 1 year (Figure 3). Figure 3 also shows the residual error time-series which also displays an annual periodicity of the error. Residuals from SPC calibrated parameters appeared to be more negative compared to OPT residuals but this wasn't examined closely in the current study.

Table 1. Validation performances for SPC using different sub-period lengths and hydrological models. Below are *p*-values from Wilcoxon signed-rank tests, where the null hypothesis was that the shift of the distribution of paired differences between SPC performances and traditionally optimised performances (i.e. $F(x-y)$) is less than or equal to zero (lower the *p*-value, the better the SPC performs compared to OPT).

	no sub-period vs OPT	SPC-1mth vs OPT	SPC-6mth vs OPT	SPC-1yr vs OPT	SPC-2yr vs OPT	SPC-5yr vs OPT
GR4J	1.000	1.000	1.000	0.889	0.00169	0.0127
Sacramento	1.000	1.000	0.607	0.0186	0.000234	0.0151
AWRA-R	1.000	1.000	0.996	0.200	0.0639	0.102

The significant spectral energy periods (e.g. significant peaks seen in Figure 3) were collected for each simulation to obtain distributions of significant periods for each combination of model and calibration method. The SPC distributions (for each sub-period sample length) were compared to the OPT distributions to test for any significant differences. For this, the two-sided Wilcoxon rank-sum test was used. No significant differences were found between OPT distributions and SPC distributions for any of the sub-period sample lengths. However, Table 2 shows that when the significant periods are limited to periods longer than 20 days, then Sacramento using SPC-1yr and SPC-2yr, and AWRA-R using SPC-6mth were significantly different to corresponding OPT distributions.

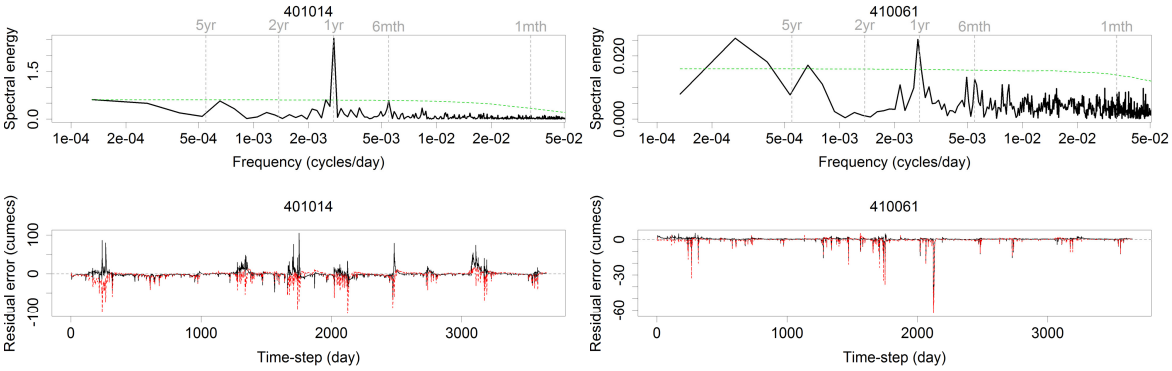

Figure 3. Top: Fourier transform for two gauging stations simulated for the calibration period using SPC-5yr parameters with GR4J. The green dashed line represents the 95th percentile of the background noise model. Any peaks that exceed this line are deemed to be significant signals. Bottom: Residual error time-series ($\mathbf{q}^{sim} - \mathbf{q}^{obs}$) for two gauging stations simulated for the calibration period using SPC-5yr parameters with GR4J. The black and red lines represent simulations with OPT and SPC parameters, respectively.

Table 2. Two-sided Wilcoxon rank-sum test *p*-values between OPT and SPC Fourier transform significant periods (performed on the calibration period). Note that significant periods are determined by the 95th percentile red noise model. Periods less than or equal to 20 days were omitted from the test. *p*-value < 0.05 indicates a statistically significant difference.

	no sub-period vs OPT	SPC-1mth vs OPT	SPC-6mth vs OPT	SPC-1yr vs OPT	SPC-2yr vs OPT	SPC-5yr vs OPT
GR4J	0.341	0.636	0.278	0.716	0.581	0.706
Sacramento	0.455	0.0816	0.566	0.00408	0.00385	0.107
AWRA-R	0.586	0.155	0.0194	0.973	0.989	0.734

Figure 4 displays standard deviations of NSE^{root} for validation sub-periods of different lengths for Sacramento. The standard deviations are noticeably less between sub-periods for SPC compared to OPT. As expected, the "no sub-period" calibration and SPC-1mth obtained higher standard deviations of NSE^{root} than OPT for longer block lengths, where SPC-6mth and SPC-1yr had relatively low standard deviations compared to OPT for shorter block lengths. The analysis was performed on the other models and for the

calibration period, which showed similar results to Figure 4. All SPC calibrations, for all models, obtained lower median standard deviations than OPT for block lengths that corresponded to the SPC sub-period sample length. This was seen in both calibration and validation periods.

Results of the mean flow frequency analysis are shown in Table 3. For GR4J and Sacramento calibration periods, there may have been a slight increase of number of sites where the most frequent mean flow sub-periods also produced the best mean NSE^{root}. These seemingly had an impact to the validation period as well, however, a more pronounced result would be expected if this was a major factor.

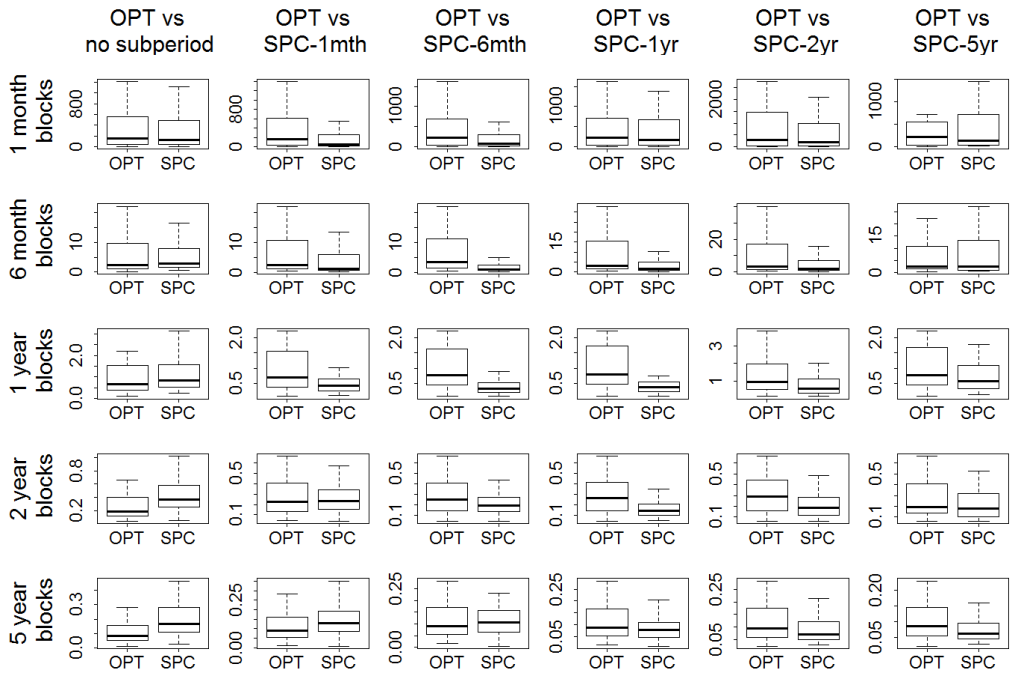

Figure 4. Boxplots of Sacramento's standard deviations of NSE^{root} calculated on various block lengths in the validation period. Only sites where OPT and SPC parameterisations differed were used in sub-plots. Note outliers were omitted to improve clarity.

Table 3. Number of sites where the most frequent mean flow blocks also produced the best mean NSE^{root}.

	GR4J						Sacramento						AWRA-R					
	no sub-period (1 month blocks)	SPC-1mth (1 month blocks)	SPC-6mth (6 month blocks)	SPC-1yr (1 year blocks)	SPC-2yr (2 year blocks)	SPC-5yr (5 year blocks)	no sub-period (1 month blocks)	SPC-1mth (1 month blocks)	SPC-6mth (6 month blocks)	SPC-1yr (1 year blocks)	SPC-2yr (2 year blocks)	SPC-5yr (5 year blocks)	no sub-period (1 month blocks)	SPC-1mth (1 month blocks)	SPC-6mth (6 month blocks)	SPC-1yr (1 year blocks)	SPC-2yr (2 year blocks)	SPC-5yr (5 year blocks)
OPT cal.	0	0	0	1	2	8	0	0	0	0	11	22	0	0	0	4	3	19
SPC cal.	2	1	3	5	2	11	0	0	2	5	11	19	0	0	0	3	3	19
OPT val.	0	0	0	0	3	6	0	0	0	0	3	20	0	0	0	1	5	22
SPC val.	0	0	1	1	3	7	0	0	0	0	6	19	0	0	0	1	6	22

4. DISCUSSION AND CONCLUSIONS

The Fourier transform analysis on residual errors produced largely inconclusive results due to confounding spectral energies at short periods (< 20 days). The short periods may be actual significant peaks or they may have been due to deficiency in the red background noise model that was used to determine significant peaks. In any case, generally there were significant spectral energies for 1 year periods and, when eliminating short periods, SPC showed significantly different spectral signals to OPT in some (albeit not all) cases (Table 2).

The main finding of the study was that the range of performance throughout the calibration period and validation period is reduced when using the SPC calibration compared to OPT (Figure 4). Also, consistency

during short time-scales seems to not be as important to validation as longer time-scales. However, as sub-period sample lengths increase, the SPC calibrated parameters more resembles those of OPT.

The mean flow frequency analysis showed that SPC calibrations might return parameterisations that fit to more frequent flow magnitudes rather than rare events. However, there was not substantial evidence to account this as a major factor.

The effectiveness of SPC calibration is likely to depend on the objective function chosen. This should be explored in future trials. Another future study could include determining if there is indeed a tendency for SPC calibration to produce negative residuals, which would also have implications on predictive performance.

REFERENCES

Bárdossy, A., Singh, S.K. (2008). Robust estimation oh hydrological model parameters. *Hydrol. Earth Syst. Sci.*, 12, 1273-1283.

Beven, K. (2006). A manifesto for the equifinality thesis. *J. Hydrol.*, 320, 18-36.

Burnash, R.J.C., Ferral, R.L., McGuire, R.A. (1973). A generalised stream-flow simulation system – conceptual modelling for digital computers. Tech. report, Joint Federal and State River Forecast Center, Sacramento.

Freer, J., Beven, K., Peters, N. (2003). Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure, *Water Science and Application*, 6, 69–87.

Gharari, S., Hrachowitz1, M., Fenicia1, F., Savenije, H.H.G. (2013). An approach to identify time consistent model parameters: sub-period calibration. *Hydrol. Earth Sys. Sci.*, 17, 149-161.

Jakeman, A.J., Hornberger, G.M. (1993). How much complexity is warranted in a rainfall-runoff model? *Water Resour. Res.*, 29, 2637-2649.

Kavetski, D., Franks, S.W., Kuczera, G. (2002). Confronting input uncertainty in environmental modelling. *Water Science and Application*, *6*, 49-68.

Kim, S.S.H., Hughes, J.D., Chen, J., Dutta, D., Vaze, J. (in press). Determining probability distributions of parameter performances for time-series model calibration: a river system trial. *J. Hydrol*..

Koutsoyiannis, D., Makropoulos, C., Langousis, A., Baki, S., Efstratiadis, A., Christofides, A., Karavokiros, G., and Mamassis, N. (2009). HESS Opinions: "Climate, hydrology, energy, water: recognizing uncertainty and seeking sustainability". *Hydrol. Earth Sys. Sci.*, 13, 247-257.

Kuczera, G., Raper, G.P., Brah, N.S., Jayasuriya, M.D. (1993). Modelling yield changes after strip thinning in a mountain ash catchment: an exercise in catchment model validation. *J. Hydrol.*, 150, 433-457.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., Michel, C. (2006). Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. *Water Resour. Res.*, 42, W07410.

Perrin, C., Michel, C. and V. Andreassian, (2003), Improvement of a parsimonious model for streamflow simulations. *J. Hydrol.*, 279, 275–289.

Seibert, J. (2003). Reliability of model predictions outside calibration conditions. *Nord. Hydrol.*, 34, 477-492.

Vaze, J., Viney, N., Stenson, M., Renzullo, L., Van Dijk, A., Dutta, D., Crosbie, R., Lerat, J., Penton, D., Vleeshouwer, J., Peeters, L., Teng, J., Kim, S., Hughes, J., Dawes, W., Zhang, Y., Leighton, B., Perraud, J-M., Joehnk, K., Yang, A., Wang, B., Frost, A., Elmahdi, A., Smith, A., Daamen, C. (2013). The Australian Water Resource Assessment Modelling System (AWRA). 20th International Congress on Modelling and Simulation, Adelaide, Australia, 1-6 December 2013.

Wagener, T. (2003). Evaluation of catchment models. *Hydrol. Process.*, 17, 3375-3378.

Zhang, Z., Moore, J. (2011). New significance test methods for Fourier analysis of geophysical time series. *Non-linear Processes in Geophysics*, 18, 643-652.