Journal of Hydrology 530 (2015) 361–371

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Determining probability distributions of parameter performances for time-series model calibration: A river system trial

Shaun Sang Ho Kim *, Justin Douglas Hughes, Jie Chen, Dushmanta Dutta, Jai Vaze

CSIRO Land and Water Flagship, Clunies Ross Street, Acton 2601, Australia

ARTICLE INFO

Article history:
Received 14 September 2015
Accepted 28 September 2015
Available online 9 October 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Sub-period
Consistency
Calibration
Prediction
Validation
Time-series modelling

SUMMARY

A calibration method is presented that uses a sub-period resampling method to estimate probability distributions of performance for different parameter sets. Where conventional calibration methods implicitly identify the best performing parameterisations on average, the new method looks at the consistency of performance during sub-periods. The method is implemented with the conceptual river reach algorithms within the Australian Water Resources Assessments River (AWRA-R) model in the Murray–Darling Basin, Australia. The new method is tested for 192 reaches in a cross-validation scheme and results are compared to a traditional split-sample calibration–validation implementation. This is done to evaluate the new technique's ability to predict daily streamflow outside the calibration period. The new calibration method produced parameterisations that performed better in validation periods than optimum calibration parameter sets for 103 reaches and produced the same parameterisations for 35 reaches. The method showed a statistically significant improvement to predictive performance and potentially provides more rational flux terms over traditional split-sample calibration methods. Particular strengths of the proposed calibration method is that it avoids extra weighting towards rare periods of good agreement and also prevents compensating biases through time. The method can be used as a diagnostic tool to evaluate stochasticity of modelled systems and used to determine suitable model structures of different time-series models. Although the method is demonstrated using a hydrological model, the method is not limited to the field of hydrology and could be adopted for many different time-series modelling applications.

1. Introduction

Model calibration is the process of adjusting model parameter values to get the best estimate of the observations (see e.g. Vaze et al. (2011) and references therein). Virtually all hydrological models must be calibrated to produce reliable simulations because there has been little evidence of strong links between physical characteristics of catchments and the parameters of hydrological models (Beven, 1989). Hydrological models are usually calibrated to observed flow data to demonstrate that the model is able to reproduce observed time-series within an acceptable level of accuracy. The success of this depends on the chosen fit statistics, which are based on the intended purpose of the model. The ability to reproduce observed data is referred to as 'performance'.

The confidence in hydrological models and calibrated parameters largely relies on whether there is good performance during extrapolation (Klemeš, 1986). 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 confidence that the selected models and parameters are appropriate for use for impact assessments, design, water management and forecasting purposes (Andréassian et al., 2009; Coron et al., 2012; Vaze et al., 2010).

It has been found that often very different parameter sets and model structures are equally acceptable system representations (Wagener, 2003). Therefore, the validation (or predictive) performance is often recognised as a more robust measure of success than calibration performance. There are multiple popular methods developed to estimate hydrological model parameters and their distributions, such as single or multi-objective calibration (Yapo et al., 1998), and dynamic identifiability analysis (Wagener et al., 2003). Techniques in defining the uncertainty of hydrological model parameters have been a growing research focus, particularly using Bayesian approaches (Beven and Binley, 1992; Wood and Rodríguez-Iturbe, 1975).

* Corresponding author. Tel.: +61 2 6246 5648.
E-mail addresses: shaun.kim@csiro.au (S.S.H. Kim), justin.hughes@csiro.au (J.D. Hughes), jie.chen@unisa.edu.au (J. Chen), dushmanta.dutta@csiro.au (D. Dutta), jai.vaze@csiro.au (J. Vaze).

http://dx.doi.org/10.1016/j.jhydrol.2015.09.073
0022-1694/

Model validation techniques have been developed for assessing how results will generalise to an independent data set (Klemeš, 1986; Mroczkowski et al., 1997). A common practice amongst hydrological modellers is the calibration–validation approach or sometimes referred to as the split-sample test (Andréassian et al., 2009; Klemeš, 1986). The model is first calibrated for a certain time period and then the calibrated parameters are transposed and evaluated in a different time period to determine its predictive performance. More complicated techniques like the k-fold cross-validation, partitions the data set into k random sub-samples. Calibration is performed on each sub-sample, and validation on the remaining data (Geisser, 1993). Another known method is to repeatedly random-sample both calibration and validation periods from the full data set (Kohavi, 1995).

Achieving acceptable validation is a major challenge in hydrological modelling. The parameters that produce the best performance during a calibration period may often yield poor results outside of that period (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.

Wagener et al. (2003) investigated the identifiability of model parameters in conceptual rainfall-runoff models by calculating parameter performances across different 'windows' of time. This method showed that there can be substantial uncertainty of model parameters through time. Coron et al. (2012) performed a similar approach whereby a sliding sub-period window calibration–validation approach was used to assess the transposability of parameters over time under various climatic conditions. The study found lower validation performances for periods that had larger differences in mean rainfall.

An approach developed by Gharari et al. (2013) provides the most consistently performing parameter sets for different periods by selecting parameter sets that are as close as possible to the optimum of each individual sub-period. Although using parameter sets with time consistent performance potentially reduced performance during the calibration period, they actually performed consistently better in the validation period. An important notion is that sub-optimal calibration period parameterisations, may perform better in validation periods than optimal ones. Often, there are multiple (sometimes very unique) parameterisations that are equally or more valid than the optima (Andréassian et al., 2012; Beven, 2006).

Given that there is ability to identify a range of potentially valid parameterisations during calibration periods, there may be undiscovered methods that utilise more information to allow further identification of which parameterisations might provide improved validation results. In statistics, resampling methods have been developed to estimate the precision of sample statistics. A simple method called jack-knifing uses subsets of data to infer the distribution of statistics of the population, assuming that the samples are suitable representations of the population (Quenouille, 1949; Quenouille, 1956). Bootstrapping is similar but draws randomly from a set of data points with replacement (Efron, 1979). Bayesian bootstrapping is an analogous method that provides the statistics' posterior distributions rather than their sampling distributions (Rubin, 1981).

Bootstrap resampling methods have been adapted to deal with time-series data. These methods have typically been used to determine distributions and confidence intervals of population statistics. These have also been useful as non-parametric time-series models since they do not assume Gaussian distributions of the data, which is an important and limiting assumption of parametric models (Bühlmann, 2002; Härdle et al., 2001; Vogel and Shallcross, 1996).

More complicated bootstrapping techniques are often required when considering time-series data since time-series data are often not independent and identically distributed (IID). A common technique is to sample blocks of data from the record with replacement, referred to as block bootstrapping (Berkowitz and Kilian, 2000). The block bootstrap replicates are synthetic ensemble time-series that are created to resemble the original data series. The resampling of blocks instead of single data points solves the problem for non-IID data since local structure is preserved within the blocks. Another solution is to bootstrap the series of residual errors rather than the data itself since these can be IID (Albeanu et al., 2008). Further variations and applications of bootstrapping time-series techniques have been explored in the literature, for example, overlapping/non-overlapping sample blocks, fixed/varying block sizes, maximum entropy, parametric/non-parametric methods, etc. (Albeanu et al., 2008; Berkowitz and Kilian, 2000; Bühlmann, 2002; Härdle et al., 2001; Politis and Romano, 1994; Politis and White, 2004; Ruiz and Pascual, 2002; Vinod, 2013; Vinod and López-de-Lacalle, 2009; Vogel and Shallcross, 1996).

While resampling methods have been used to provide an idea of the uncertainty of population statistics, using this information to improve the predictive performance of models is a novel concept. The aim of the current study is to develop a generic method that can determine model parameterisations that will give improved performance for prediction over basic calibration techniques. This is done by using a sub-period resampling method to estimate performance consistency. The new calibration method is designed to be used for time-series model calibrations to determine the most consistently performing parameters. These are, therefore, expected to be more robust than parameters that perform the best overall. The study uses a river system model to test the method in the Murray–Darling Basin in Australia. A river model, a genre of hydrological model used for water resources planning and management of a river system, combines rainfall and routing components that simulate natural hydrological processes with components representing anthropogenic water use (Dutta et al., 2014; Welsh et al., 2013).

The methods section first describes the new calibration method, followed by descriptions of the hydrological model and the study area used for the new method's evaluation. The use of the new method and the strategy of evaluating it against the traditional split-sample calibration is then given. A comparison of the two calibration methods is provided in the results section including a statistical test. This is followed by an analysis of the new calibration method in the discussion.

2. Methods

2.1. Sub-period consistency calibration

The sub-period consistency (SPC) calibration procedure is shown in Fig. 1. The new method uses a sub-period resampling method to determine the probability distributions of obtaining specified performances for sampled sub-periods. In the current