Journal of Hydrology 527 (2015) 641–655

Contents lists available at ScienceDirect

Journal of Hydrology

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

Do higher data frequency and Bayesian auto-calibration lead to better model calibration? Insights from an application of INCA-P, a process-based river phosphorus model

Leah A. Jackson-Blake [a,*], Jostein Starrfelt [b,c]

[a] The James Hutton Institute, Craigiebuckler, Aberdeen AB15 8QH, UK
[b] NIVA, Gaustadalléen 21, NO-0349 Oslo, Norway
[c] Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, P.O. Box 1066 Blindern, NO-0316 Oslo, Norway

ARTICLE INFO

Article history:
Received 23 June 2014
Received in revised form 10 February 2015
Accepted 1 May 2015
Available online 9 May 2015
This manuscript was handled by Peter K. Kitanidis, Editor-in-Chief, with the assistance of Wolfgang Nowak, Associate Editor

Keywords:
Catchment modelling
Phosphorus
INCA
MCMC DREAM
Data frequency

SUMMARY

We use Bayesian auto-calibration to explore how observed data frequency affects the performance and uncertainty of INCA-P, a process-based catchment phosphorus model. A fortnightly dataset of total dissolved phosphorus (TDP) concentration was derived from 18 months of daily data from a small (51 km^2) rural catchment in northeast Scotland. We then use the DiffeRential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo (MCMC) algorithm to calibrate 29 of the >127 model parameters using the daily and the fortnightly observed datasets. Using daily rather than fortnightly data for model calibration resulted in a large reduction in parameter-related uncertainty in model output and lower risk of obtaining unrealistic results. However, peaks in TDP concentration were as well simulated as when fortnightly data were used. A manual model calibration did a better job of simulating the magnitude of TDP peaks and baseflow concentrations, suggesting that alternative measures of model performance may be needed in the auto-calibration. Results suggest that higher frequency sampling, perhaps for just a short period, can greatly increase the confidence that can be placed in model output. In addition, we highlight the many subjective elements involved in auto-calibration, in an attempt to temper a common perception that auto-calibration is an objective and rigorous alternative to manual calibration. Finally, we suggest practical improvements that could make models such as INCA-P more suited to auto-calibration and uncertainty analyses, a key requirement being model simplification.

1. Introduction

Nutrient enrichment of aquatic ecosystems can result in the excessive growth of algae and changes in species composition. Nitrogen and phosphorus (P) are the two most important nutrients contributing to eutrophication, and P is generally of most concern in freshwaters (Correll, 1998). Reducing P fluxes to surface waters has therefore become a top conservation priority. In-stream P concentration is controlled by a variety of input fluxes and processes, many of which are highly variable spatially and temporally (e.g. House, 2003; Edwards and Withers, 2007; Withers and Jarvie, 2008). Many of these fluxes can be modified by land management, but there is often no straightforward or immediate link between implementing a measure to reduce P inputs and observable in-stream effects (Meals et al., 2010; Jarvie et al., 2013; Sharpley et al., 2013). This makes it difficult for policy makers and water managers to decide on appropriate measures, or to know when improvements in water quality might be expected after measures have been implemented. Given this complexity, it can be useful to use process-based, integrated catchment models, which provide a means of formalising current knowledge of a catchment system. These models can be used to help set appropriate water quality and load reduction goals, to advise on the best means of achieving those goals, to predict time lags in the system and to explore potential system responses to future environmental change.

For such models to be useful they need to be well calibrated and tested and be shown to reproduce key catchment processes. Many model parameter values may need to be determined through calibration, either because they are not directly measurable, there is a lack of data to constrain values, or measurements apply at different spatial and/or temporal scales to those at which the model operates (Oreskes et al., 1994). In general, the more parameters a model has, the more observations are required for those parameters to be

* Corresponding author. Tel.: +44 (0)1224 395161.
E-mail addresses: leah.jackson-blake@hutton.ac.uk (L.A. Jackson-Blake), jostein.starrfelt@niva.no (J. Starrfelt).

http://dx.doi.org/10.1016/j.jhydrol.2015.05.001
0022-1694/

identifiable (Kuczera and Mroczkowski, 1998). Calibration can therefore be challenging for process-based catchment models, which tend to be complex and highly parameterised. Despite this, most catchment P model applications are calibrated using relatively sparse fortnightly or monthly observed data (reviewed in Jackson-Blake et al., 2015). We might expect daily data, in which short-lived flow events are better captured, to result in more realistic model calibrations. McIntyre and Wheater (2004) obtained better model calibrations using higher frequency synthetic datasets, but few studies have compared model calibrations using different frequencies of real data. A primary aim of this paper is therefore to investigate how increasing the sampling frequency of the observed data used in model calibration impacts on the quality and uncertainty of simulated dissolved P concentrations over the calibration period. We used the INtegrated CAtchment model of Phosphorus dynamics (INCA-P; Wade et al., 2002, 2007), which has been applied extensively to assess the potential effects of changing land management, land use and climate on water quality (e.g. Crossman et al., 2013; Farkas et al., 2013; Whitehead et al., 2013). To allow a robust comparison between model calibrations obtained using different observed data frequencies, we used an automated model calibration procedure.

In recent years, particular emphasis has been placed on the benefits of automated calibration techniques over manual calibration (e.g. Gupta et al., 2006). For most hydrological and process-based water quality models, the large number of non-linearly interacting parameters makes manual calibration a labour intensive and difficult process, requiring considerable experience (Wagener and Gupta, 2005). With automated techniques, algorithms search the parameter space to find an optimum parameter set or set of probability distributions, for a given set of observed data. Automated techniques may therefore allow a more thorough searching of the parameter space, and can often be used to investigate model sensitivity to the parameters and parameter-related uncertainty in model output, both key in improving confidence in models. Whilst the ideas behind auto-calibration are certainly good, auto-calibration may produce less realistic results than manual calibration (Boyle et al., 2000; Van Liew et al., 2005). In addition, despite the obvious benefits of auto-calibration, it is still a highly subjective procedure (e.g. Pappenberger et al., 2007). We explore these issues by comparing auto-calibration results with a manual calibration, and by highlighting some of the key subjective elements encountered during the analysis. Finally, automated techniques can only be applied well when model design is suited to such techniques. We point out practical issues in the design of INCA-P which reduce the ease with which it, and similar models, can be automatically calibrated, and suggest improvements that could be made.

A number of automated calibration techniques exist. Here, we use a Bayesian framework, which results in probability densities for calibrated parameters rather than single 'optimum' values, as in procedures such as PEST (Doherty et al., 1994). Probability densities allow equifinality to be taken into account (Beven and Freer, 2001), as well as allowing us to quantify uncertainty in model output due to parameter uncertainty. In recent years much effort has been put into developing algorithms which are able to generate samples representative of complex posteriors. We use the DiffeRential Evolution Adaptive Metropolis (DREAM) Markov Chain Monte Carlo (MCMC) algorithm developed by Vrugt et al. (2009b), and adapted for use with INCA-P by Starrfelt and Kaste (2014), which has been shown to perform well when used with complex hydrological models (e.g. Huisman et al., 2010).

In summary, this paper has the following primary aims: (1) Investigate how increasing the data frequency used in calibration affects the uncertainty and realism of model output; (2) Highlight the subjective elements involved in auto-calibration and uncertainty analysis; and (3) Suggest practical improvements that could be made to models such as INCA-P to make them better suited to auto-calibration.

2. Methods

2.1. Study catchment

The model applications were carried out for the Tarland Burn catchment (51 km^2), a rural catchment in northeast Scotland where diffuse agricultural sources provide the dominant phosphorus input. Land use is a mixture of arable (17%), improved grassland (30%), rough grazing (13%) and forestry and moorland (40%). Agricultural activities focus on beef cattle, sheep and spring barley. Humus iron podzols and brown forest soils are the dominant soils underlying all but semi-natural land, where peaty podzols are also important. Over the study period (2004–2005), mean annual rainfall was 960 mm and median discharge was 0.65 $m^3\,s^{-1}$, with 5th and 95th percentiles of 0.29 and 1.38 $m^3\,s^{-1}$. The village of Tarland has a small wastewater treatment works and septic tanks serve around half the catchment's residents. Water quality is of concern, primarily due to inputs of nutrients and sediments from agriculture.

2.2. Monitoring data used for calibration

INCA-P has been shown to produce adequate dissolved P simulations in the study catchment, whilst particulate P is currently more problematic (Jackson-Blake et al., 2015). The analysis therefore focuses on dissolved P, and measurements of discharge and total dissolved P (TDP) concentration were used for model calibration. Daily mean discharge was obtained from 15 minute data from a gauging station at Coull, at the catchment outflow (Fig. 1). Daily stream water samples were collected between April 2004 and June 2005, also from the catchment outflow. Water samples were analysed for total dissolved P (TDP), giving 15 months of daily data – full details of sampling protocol, auto-sampler set-up and analytical methods are given in Stutter et al. (2008). Additionally, lower frequency sampling (fortnightly or less) took place at three locations on the Tarland Burn upstream of Coull (Fig. 1), as well as between June and December 2005 at Coull. This was also used for model calibration to give two full years of data. Observed stream water TDP from September and October 2005 was excluded

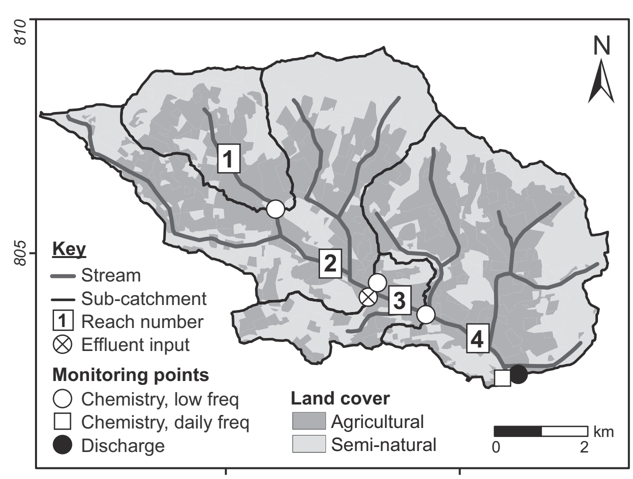

Fig. 1. Map of the Tarland catchment, together with the location of monitoring points and effluent input, simplified land use, and reaches and associated sub-catchments used in the INCA set up. Coull is at the bottom of reach 4. Eastings and northings (km) are relative to the British National Grid.