Journal of Hydrology 538 (2016) 551–562

Contents lists available at ScienceDirect

Journal of Hydrology

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

Technical Note

Tools for investigating the prior distribution in Bayesian hydrology

Yating Tang [a], Lucy Marshall [a,*], Ashish Sharma [a], Tyler Smith [b]

[a] UNSW Water Research Centre, School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW 2052, Australia
[b] Department of Civil and Environmental Engineering, Clarkson University, Potsdam, NY 13699, USA

ARTICLE INFO

Article history:
Received 28 October 2015
Received in revised form 17 March 2016
Accepted 16 April 2016
Available online 24 April 2016
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Hamid Moradkhani, Associate Editor

Keywords:
Bayesian hydrological modeling
Prior distribution
Parameter sensitivity
Kullback-Leibler Divergence
Elasticity

SUMMARY

Bayesian inference is one of the most popular tools for uncertainty analysis in hydrological modeling. While much emphasis has been placed on the selection of appropriate likelihood functions within Bayesian hydrology, few researchers have evaluated the importance of the prior distribution in deriving appropriate posterior distributions. This paper describes tools for the evaluation of parameter sensitivity to the prior distribution to provide guidelines for defining meaningful priors. The tools described here consist of two measurements, the Kullback-Leibler Divergence (KLD) and the prior information elasticity. The Kullback-Leibler Divergence (KLD) is applied to calculate differences between the prior and posterior distributions for different cases. The prior information elasticity is then used to quantify the responsiveness of the KLD values to the change of prior distributions and length of available data. The tools are demonstrated via a Bayesian framework using an MCMC algorithm for a conceptual hydrologic model with both synthetic and real cases. The results of the application of this toolkit suggest the prior distribution can have a significant impact on the posterior distribution and should be more routinely assessed in hydrologic studies.

1. Introduction

Bayesian inference is one of the most popular tools for quantifying uncertainties in hydrological modeling. With the development of Markov chain Monte Carlo (MCMC) techniques and increased computer power, Bayesian inference is being used increasingly for hydrological model specification (e.g. Bates and Campbell, 2001; Kuczera, 1983; Marshall et al., 2004; Smith and Marshall, 2008; Vrugt et al., 2003; Jeremiah et al., 2011).

In Bayesian theory, the posterior distribution is estimated as a combination of our existing knowledge about the model parameters prior to observing the data (the prior distribution) and a likelihood function derived from a probability model of the data being observed. Although much research has been conducted on the investigation of various likelihood functions in hydrologic applications (Kuczera, 1983; Bates and Campbell, 2001; Schoups and Vrugt, 2010; Evin et al., 2013; Smith et al., 2010, 2015), limited work has been done on defining meaningful prior distributions for different models. The only journal paper to our knowledge that attempts to include truly informative priors for rainfall–runoff model parameters is that by Bates and Campbell (2001). A fully Bayesian approach to parameter estimation was provided in that study for a conceptual rainfall–runoff model, with prior distributions defined according to previous similar modeling studies and existing knowledge of the model parameters and their likely ranges.

Despite this, in recent years the importance of prior information and the influence of expert knowledge on model specification have gained increasing recognition. Rojas et al. (2009) presented an assessment of prior knowledge and a sensitivity analysis of the prior in groundwater modeling, emphasizing the importance of selecting proper prior probabilities. Renard et al. (2010) discussed the challenge of identifying input and structural errors in hydrological modeling, concluding that the success of uncertainty analysis is largely determined by prior hypotheses describing the distributional properties of rainfall and runoff errors. Smith et al. (2014) introduced a hierarchical Bayesian statistical approach to the PUB (Predictions in Ungauged Basins) literature, where global prior distributions were derived as a way to quantify the predictive uncertainty for predictions made at the ungauged catchments. Gharari et al. (2014) assessed the effect of imposing semi-quantitative, relational expert knowledge for model development and parameter selection. Based on comparisons with different models, the results show that a prior-constrained but uncalibrated semi-distributed model can predict with similar performance to a calibrated lumped model, and also reduce uncertainty in predictions. Finally, Hrachowitz et al. (2014) tested the value of a

* Corresponding author.
E-mail address: lucy.marshall@unsw.edu.au (L. Marshall).

http://dx.doi.org/10.1016/j.jhydrol.2016.04.032
0022-1694/

systematic use of hydrological signatures and expert knowledge for increasing performance of a model and its skill to reproduce hydrological signatures. These studies indicate the potential importance of prior information in classical Bayesian hydrology and suggest that the sensitivity of model results to assumed prior distributions should be more routinely assessed in hydrologic studies.

One of the key questions that have troubled practicing hydrologists is whether anything more than a uniform prior is needed in modeling studies. Note that a uniform prior is not necessarily a 'non-informative' prior (Kass and Wasserman, 1996). Other questions that are related to this and remain unaddressed are: how much data is needed for priors to become irrelevant, and how would the outcome be impacted if the assumed prior mean or variance is biased with respect to its optimal value? While answers to these questions depend on a range of factors (i.e., model type, complexity, uncertainties, and catchment characteristics), we aim to provide here general tools to measure the impact of the prior on different model parameters and quantify its impact for different prior distributions and lengths of data. Our expectation is this added information will lead to focused efforts on making the prior more informative in situations where its importance is high, and allowing use of less informed diffused priors when this is not the case.

To achieve this, a Bayesian framework for assessing the prior distribution is implemented using a MCMC algorithm and an established conceptual rainfall–runoff model. A synthetic data set with known parameter values is used to evaluate parameter sensitivity to the prior. The Kullback-Leibler Divergence (KLD) is applied to calculate the differences between the prior and posterior distributions. The concept of prior information elasticity is introduced to analyze the degree of responsiveness of the posterior distribution to the change in prior distribution. A real case study using the Bass River catchment (VIC, Australia) flow dataset is implemented, demonstrating the assessment of prior sensitivity to a range of factors through the application of this toolkit.

2. Methodology

The procedure we present here has three main steps: (1) measuring the differences between the prior and posterior distributions for model parameters using the Kullback-Leibler Divergence (KLD); (2) calculating the responsiveness of the change in KLD values to the change in the prior and the available data using prior information elasticity; and (3) determining the appropriate informative level of the prior distributions for each parameter by checking parameter sensitivity to the prior and the data according to the results from (1) and (2). We expand on the methods used in the first two steps below.

2.1. Kullback-Leibler Divergence (KLD) as a measure of prior information

The Kullback-Leibler Divergence (also called information divergence or relative entropy), first introduced by Kullback and Leibler (1951), is a non-symmetric measure of the difference between two probability distributions. In information theory, the KLD is a measure of the information lost when one probability distribution is used to approximate another.

If we let $P(x)$ and $Q(x)$ be two continuous probability density functions (pdf), the Kullback-Leibler Divergence between P and Q can be defined as:

$$D_{KL}(P||Q) = \int P(x)\log\frac{P(x)}{Q(x)}dx \tag{1}$$

The KLD has three properties:

(1) The KLD is always greater than or equal to zero, $D_{KL}(P||Q) \geqslant 0$.
(2) The KLD is non-symmetric, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.
(3) $D_{KL}(P||Q) = 0$ only if $P = Q$.

For continuous and/or non-parametric distributions (such as the posterior distribution of hydrologic model parameters estimated via MCMC), the KL divergence can be approximated via Monte Carlo (MC) estimation (Hershey and Olsen, 2007). Let X be a (multivariate) random variable, with pdf P. Then, by definition:

$$\mathrm{KLD}(P||Q) = E\left[\log\left(\frac{P(X)}{Q(X)}\right)\right] \tag{2}$$

The MC methodology can therefore be applied to estimate such expectations, by the following steps:

(1) Draw n independent samples x_i from the pdf P.
(2) Compute $\mathrm{KLD}(P||Q) = \frac{1}{n}\sum_i \log\left(\frac{P(x_i)}{Q(x_i)}\right)$ when $n \to \infty$.

When using an MC approach to estimate KLD in Bayesian inference, P is the posterior distribution. When the samples x_i are drawn directly from the posterior via MCMC, the pdf is estimated through a kernel smoothing function estimation based on these samples. Q is the prior distribution and typically has a parametric form that may be evaluated for the sample values.

The KLD is from a family of divergences known as the alpha-divergence (Rrnyi, 1961; Minka, 2005). In this family a symmetric measurement which is called Hellinger Distance (HD) also can be calculated easily from the MCMC outputs and then used to calculate the prior information elasticity. In our study the KLD is selected because the HD is less sensitive to the movement of the posterior than the KLD especially when the prior and the posterior are very different.

2.2. Elasticity as a measure of relative importance of the assumed prior

The term 'elasticity' was originally used in economics, commonly referring to the price elasticity of supply and demand (Marshall, 1890, 2009). It is a measurement of how responsive an economic variable is to a change in another. It is a popular tool due to the independence of units and simplification of data analysis. Thus, it is widely adopted by many studies as a measurement to summarize responsiveness. This concept was introduced in hydrologic analysis (Schaake and Waggoner, 1990) for evaluating the sensitivity of streamflow to changes in climate. For example, precipitation elasticity ε_p of streamflow Q can be defined by the proportional change in precipitation P (Sankarasubramanian et al., 2001):

$$\varepsilon_p(P,Q) = \frac{dQ/Q}{dP/P} = \frac{dQ}{dP}\frac{P}{Q} \tag{3}$$

One frequently used robust estimator of elasticity is (Sankarasubramanian et al., 2001):

$$\varepsilon_p = \mathrm{median}\left(\frac{Q_t - \overline{Q}}{P_t - \overline{P}}\frac{\overline{P}}{\overline{Q}}\right) \tag{4}$$

where $\overline{Q}$ and $\overline{P}$ are the long-term sample means.

In our study, we use the elasticity concept to measure the responsiveness of the differences between the prior and posterior distributions (KLD) to the proportional change of the prior distribution and the length of data. By testing prior distributions with