Hydrol. Earth Syst. Sci. Discuss., 6, C83–C91, 2009
www.hydrol-earth-syst-sci-discuss.net/6/C83/2009/

Hydrology and Earth System Sciences Discussions

Interactive comment on "A novel approach to parameter uncertainty analysis of hydrological models using neural networks" *by* D. L. Shrestha et al.

D. L. Shrestha et al.

d.shrestha@unesco-ihe.org

Received and published: 18 March 2009

We are very grateful to Dr. J. Vrugt for valuable comments. We carefully studied the comments and the followings are our responses (bold type) to them. We will revise our manuscript in accordance with his comments as far as possible.

In this paper the authors use Artificial Neural Networks (ANN) to help perform Monte Carlo analysis, and predict the ranges of the model output in terms of 5 and 95The approach is illustrated using discharge data from the Brue Catchment in the UK using the HBV conceptual watershed model. This model has 9 parameters and has been extensively used and discussed in the literature.

Comments: 1. P1678 - 1679: L24 - 4: The authors highlight the GLUE concept of Beven and coworkers as widely used approach that utilizes Monte Carlo simulations to estimate parameter uncertainty. Other approaches that deserve to be highlighted in this context include Markov Chain Monte Carlo (MCMC) simulation approaches that use iterative sampling to estimate nonlinear parameter uncertainty intervals. Examples of this include the Random Walk Metropolis (RWM), Delayed Rejection Adaptive Metropolis (DRAM: Haario et al. 2006), SCEM-UA (Vrugt et al. 2003), and DREAM (Vrugt et al. 2008) methods that have found application in the field of hydrology. Given the increasing interest in Bayesian sampling, and uncertainty quantification these methods are increasingly likely to be used to estimate parameter and model prediction uncertainty.

As a MC method we used the widely accepted GLUE method - also because we wanted to have some sort of a comparison base with many other studies where GLUE was used as the MC method. However we agree that the other, more efficient, methods have to be mentioned as well and this will be done in the revised version.

2. P1679: 5 - 7: If replaced with MCMC then comparison of statistics of multiple sample paths in parallel would provide a more formal solution to assessing how many model evaluations are required to assess convergence and obtain stable statistics of model output and parameters.

We agree that for MCMC there exist formal approaches (e.g. Gelman and Rubin (1992) to assess the convergence of the simulations. It will be mentioned in the revised version.

3. P1679: 19 - 22: Dekker and Bouten used ANN approaches to analyze the mismatch between model predictions and observations through a hierarchical analysis approach. This method was able to correct the response functions of a Jarvis type of forest transpiration model, and provide important insights into model structural error. This work

deserves to be highlighted within the present context.

There were many papers where the use of ANN for building error model was reported - but the size of the references list is limited so one has to make choices. The present paper is devoted to a different subject so we thought to have the two references to ANN error modelling is enough. If the reviewer insists, we could add the reference to Dekker, Bouten and Schaap (2001) as well.

4. P1687 - 1688: The authors consider 74,467 different MC runs to define parameter uncertainty with the HBV model. This is a significant number of model runs considering that only 9 different parameters are tuned. Adaptive MCMC simulation with multiple chains in parallel such as utilized in SCEM-UA or DREAM would have generated these results much more efficiently, and would have also produced a calibrated model. My experience suggests that about 25,000 HBV model evaluations are required at most to derive the optimum of the HBV model, and a sample set defining parameter uncertainty. This approach is much more efficient than the MC method considered herein, and provides a statistically exact estimate of the model prediction bounds.

MC simulation are performed to ensure dense coverage of 9-parameter space, so using this logic one may say 74467 is not a really big number (note that for a 9-dimensional grid with only 4 values on each axis the number of grid points is $4^9 = 262144$). In this study a high number (74,467) MC simulations were required since many of them would be rejected as non-behavioural (in our case we have left 25000 behavioral samples with the Nash-Sutcliff coefficient of efficiency of 0 and higher). However, the convergence of the MC simulations is achieved much earlier around 10,000 simulations (see figure 4). We have also analyzed the number of simulations required to obtain 10,000 behavioral with respect to different threshold values. The following figure (Fig. 1) shows percentage of the behavioral samples retained out of total simulations in the left axis and the number of simulation required to obtain 10,000 behavioral simulations in the right axis. This result was however not shown in the paper due to space limitations.

In the future we plan to repeat the experiments with the other MC methods as well.

5. P1687 - 1688: L24 - 17: The authors use the Coefficient of Variation to determine how many samples are sufficient to obtain reliable and stable MC statistics. Why did the author select this criterion, and how would the MC results look if another metric was used to characterize the distance between model predictions and observations? Nowadays many hydrologic studies utilize multiple performance functions to measure the quality of a model. In principle, the ME and SDE statistics can be computed with any objective function.

Since we have used CE as objective function to calibrate the model and CE was suggested to be used as the likelihood measure in GLUE, we used the same metric also to determine the convergence of the MC simulations. In fact, we have also carried out experiments with other 9 performance functions (RMSE, Absolute error, Percentage bias and others). The results are consistent with CE. Furthermore we have also analyzed the convergence of the runoff predictions at some arbitrary points (e.g. in peak flow, medium flow and base flow), and the results are quite consistence with the previous performance functions. Unfortunately due to space limitations all this was not reported.

6. P1688 - 1689: Histograms of parameters: MCMC simulation with adaptive proposal updating could have created these results much more efficiently. Also, this approach has a more formal way of testing convergence by comparing the evolution of different sampling paths. This is somewhat similar to the statistics utilized herein, but then applied to a multi-chain method using the R-statistic of Gelman and Rubin (1992). See previous comment related to the introduction.

We agree. Sampling in GLUE is done randomly, that is why we have only one third of total simulations are behavioral wrt threshold 0. However in MCMC, samples are drawn from a random walk which adapts to the true posterior distribu-

tion. See also our answer to comment 1.

7. ANN emulating prediction uncertainty bounds: Does this ANN model consider uncertainty in its predictions? That would be appropriate within the current context. Bayesian calibration could in principle be used to derive the posterior distribution of the weights and biases of the network. Also, how does the performance of the ANN depend on the selection of the complexity of the network? Some insights into this might be helpful and appreciated by the readers.

We did not consider uncertainty in ANN model explicitly - it is a purely deterministic model. However we have done a number of steps to reduce the uncertainty of the predictions by ANN. For example, we select the number of neurons optimally (see pp 1691 of the original paper). We have repeated experiments with different initializations of weight and bias of the networks. In principle one could also use Bayesian neural networks, but this was not done. In the revised paper we will mention this possibility.

8. P1692, Figure 7: The scatter plots demonstrate considerable autocorrelation between the residuals. This suggests either error in the input data or structural inadequacies in the model or a combination of both. The analysis presented in this paper would be more completely if it includes explicit recognition of model and/or rainfall (forcing) error.

It is observed that residuals are autocorrelated and hetroscedastic (see, e.g. Shrestha and Solomatine (2008)). The percentage coverage of the observed data by the estimated 90(about 77parameter, input data or model structure uncertainties or combination of them. We have pointed out model structure uncertainty in the paper (p 1692, L 4-5). In the revised paper we will highlight other sources of uncertainty. However, this paper is not intended to find sources of uncertainty, but to approximate the uncertainty estimated by any time consuming uncertainty method (GLUE in this study). In doing so, we used GLUE method and consider

only parameter uncertainty for sake of the simplicity. In the near future we are exploring other sources of uncertainty and different uncertainty methods.

9. P1692. Figure 9: Model prediction uncertainty caused by parameter uncertainty is rather large. Obviously the approach utilized herein to distinguish between behavioral and non-behavioral parameter combinations is not formally Bayesian. This needs to be highlighted.

We agree. Formally, GLUE is not Bayesian, as was discussed in a number of recent publications and discussions (e.g. Vrugt et al., Stoch Envir Res and Risk Assess, doi:10.1007/s00477-008-0274-y). (However we never claim we are Bayesian.) We will reflect this in the paper accordingly.

10. P1692: What is the computational time required to run / implement the proposed approach, and how does this compare to using standard MC runs with the HBV model? My experience with the HBV model suggests that this model can be calibrated within just a few minutes, including proper assessment of parameter uncertainty. So, there is no need to adopt the approach presented herein for the current problem.

The standard MC runs of 25000 simulations with the HBV model for hourly resolution data of one year takes 9 hours (Pentium Dual core 1.8 GHz, 1.8GHz, 2GB RAM) (there is the exchange of data between the MC code and the model via a file). This prohibits the use of MC in many real-time applications, especially for computationally intensive models. Training of ANN model in this study took about 30 minutes of CPU time using Levenberg-Marquardt optimization algorithm including a couple of iterations to get good results. Time required for choice of variables and the data preparation (see response below about the data processing work) varies person to person and depends on the person experience. Generally it may take a few hours to a day, but this is done once. Afterwards a fast ANN is used in operation. Note that the purpose of this paper is to demonstrate the methodology and we consider HBV model as an example.

11. How would the ANN model perform during extrapolation? The way the authors have used the ANN in the present context is essentially for interpolation. However, in many practical situations involving computationally demanding forward models it is conceivable that prediction limits are required for yet unobserved events whose magnitude and conditions are quite different than any observed during the calibration period. The ANN model will exhibit severe difficulties producing reliable prediction limits.

It is well recognized that ANN and other data-driven models may be not very strong in extrapolation. This means the results are reliable only within the boundaries of the domain where the training data are given. Hence an attempt should be made to ensure that the training data includes all possible combinations of the events including the extremes - as far as it is possible. In the revised paper the problem of extrapolation will be mentioned.

12. The impact of the presented work would be greatly improved if the authors provide a software package that others can download and use in their own research. Would it be possible to make available such a package? If not available, I wonder whether the approach developed herein will actually find practical implementation.

There are basically three steps involved in the presented work: a) running MC simulations to generate data for ANN model, b) preprocessing the data to select the input variables for ANN model and c) building and training ANN model. In principle, it is possible to integrate all these three steps and make a software package available for downloading. However, in ANN and other machine learning, the most important issue is to select the relevant and causal input data from the available hydrometrological data and possible state variables (if any). The selection of the input data consists of preprocessing of the data such as correlation or average mutual information analysis and considering the expert knowledge of the system under consideration (e.g. in this study hydrological knowledge). This preprocessing of the input data is rather manual work and needs human interface. The first and third steps in the presented methodology are straight forward

to implement or conduct by anyone to repeat the experiment. We deliberately did not show the flowchart of the method so as to economize the length of the paper as mentioned before. In the revised version we will explicitly mentioned the steps required to conduct the experiment.

Of course we agree with Dr Vrugt that it would be great to have usable software, and we are considering to make the tools available that can be used by others with minimum effort to his/her own application.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 6, 1677, 2009.

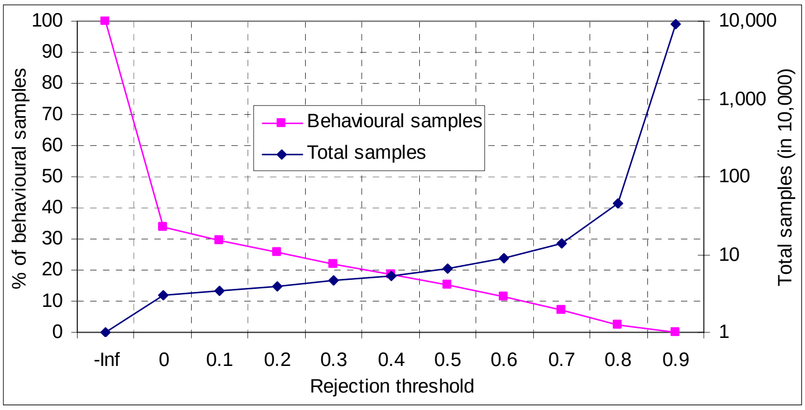

Figure 3-31 Percentage of behavioural samples and total number of sample required for 10,000 behavioural samples with different value of rejection threshold as measured by coefficient of model efficiency, CE

Fig. 1. Percentage of behavioural samples and total number of sample required for 10,000 behavioural samples with different value of rejection threshold as measured by coefficient of model efficiency, CE