Journal of Hydrology 529 (2015) 754–767

Contents lists available at ScienceDirect

Journal of Hydrology

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

A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation

Richard Arsenault *, Philippe Gatien, Benoit Renaud, François Brissette, Jean-Luc Martel

École de technologie supérieure, 1100 rue Notre-Dame Ouest, Montréal, QC H3C 1K3, Canada

ARTICLE INFO

Article history:
Received 16 February 2015
Received in revised form 31 August 2015
Accepted 1 September 2015
Available online 8 September 2015
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Vazken Andréassian, Associate Editor

Keywords:
Model averaging
Objective functions
Averaging method comparison
Model error reduction

SUMMARY

This study aims to test whether a weighted combination of several hydrological models can simulate flows more accurately than the models taken individually. In addition, the project attempts to identify the most efficient model averaging method and the optimal number of models to include in the weighting scheme. In order to address the first objective, streamflow was simulated using four lumped hydrological models (HSAMI, HMETS, MOHYSE and GR4J-6), each of which were calibrated with three different objective functions on 429 watersheds. The resulting 12 hydrographs (4 models $\times$ 3 metrics) were weighted and combined with the help of 9 averaging methods which are the simple arithmetic mean (SAM), Akaike information criterion (AICA), Bates–Granger (BGA), Bayes information criterion (BICA), Bayesian model averaging (BMA), Granger–Ramanathan average variant A, B and C (GRA, GRB and GRC) and the average by SCE-UA optimization (SCA). The same weights were then applied to the hydrographs in validation mode, and the Nash–Sutcliffe Efficiency metric was measured between the averaged and observed hydrographs. Statistical analyses were performed to compare the accuracy of weighted methods to that of individual models. A Kruskal–Wallis test and a multi-objective optimization algorithm were then used to identify the most efficient weighted method and the optimal number of models to integrate. Results suggest that the GRA, GRB, GRC and SCA weighted methods perform better than the individual members. Model averaging from these four methods were superior to the best of the individual members in 76% of the cases. Optimal combinations on all watersheds included at least one of each of the four hydrological models. None of the optimal combinations included all members of the ensemble of 12 hydrographs. The Granger–Ramanathan average variant C (GRC) is recommended as the best compromise between accuracy, speed of execution, and simplicity.

1. Introduction

Many aspects of daily operations in water resources management require an ability to predict future streamflows with the best possible accuracy. Over the years, numerous hydrological models have been proposed, each with its strengths and weaknesses. All adequate hydrological models have the capacity to predict streamflows, but none is able to consistently outperform the others for all basin characteristics and heterogeneous climatologies (i.e. the best all-around model). Quite a few studies on limited numbers of catchments have shown that weighted averages of multiple model simulations are more robust and more efficient than their individual members. Cavadias and Morin (1986) introduced the concept of weighted multi-model averaging for streamflow determination using the Granger and Newbold method (Granger and Newbold, 1977). Shamseldin et al. (1997) then showed that multi-model averaging improved performance over individual model simulations using three averaging techniques: simple arithmetic mean, constrained ordinary least-squares weighting and a neural network averaging method. Shamseldin et al. (2007) compared three types of neural networks (Simple Neural Network, Radial Basis Function Neural Network and Multi-Layer Perceptron Neural Network) in a flow averaging study. They found that the neural networks outperform the models taken independently. However, neural networks are time-consuming to conduct and are prone to over-fitting. Other weighting schemes have been put forth which can combine streamflows in various manners to improve the averaged hydrograph. One such method, the Bayesian model averaging method (BMA), computes weights based on the probability density function of the ensemble (Hoeting et al., 1999; Raftery, 1993; Raftery and Zheng, 2003; Raftery et al., 2005). While weighted averaging was devised to incorporate the advantages of each individual member, it was shown that BMA is not appropriate if

* Corresponding author. Tel.: +1 514 396 8800x7690.
E-mail addresses: richard.arsenault.1@ens.etsmtl.ca (R. Arsenault), philippe.gatien@gmail.com (P. Gatien), benoit.renaud1@hotmail.com (B. Renaud), francois.brissette@etsmtl.ca (F. Brissette), martel.jean.luc@gmail.com (J.-L. Martel).

http://dx.doi.org/10.1016/j.jhydrol.2015.09.001
0022-1694/

too many members are used (Neuman, 2003). BMA should therefore be limited to fewer and relatively similar member ensembles (Jefferys and Berger, 1992).

The seminal paper by Diks and Vrugt (2010) compared 7 model averaging methods: Equal Weights Averaging (EWA), Akaike/Bayes Information Criterion Averaging (AICA/BICA), Bates and Granger Averaging (BGA), Granger–Ramanathan-A Averaging (GRA), Bayesian model averaging (BMA) and Mallows Model Averaging (MMA). They conclude that the unconstrained methods (weights are not constrained to sum to unity) perform better than the constrained methods, and that the GRA method is the best overall since it is much faster and quicker to implement than MMA and BMA while offering the same performance.

Another study by Ajami et al. (2006) compared the EWA and constrained Ordinary-Least-Squares methods to the Multi-Model Super Ensemble (MMSE) and Modified MMSE (M3SE) methods using the Distributed Model Intercomparison Project Results (Smith et al., 2004). MMSE is used mostly in climate and weather forecasting but was applied to hydrological time series. M3SE is a frequency-based bias-corrected averaging method. These methods include bias correction and variance reduction to further improve simulation quality. The authors showed that the M3SE and MMSE methods are better than individual models, as previous studies have shown. They also showed that MMSE can sometimes produce unrealistic results (such as negative flows) because of the bias correction method implemented in the method.

Applications of multi-model flow prediction have been studied for over a decade. See and Openshaw (2000) proposed a probabilistic switching mechanism where the output from a single member was used at each time step, switching the donor member as hydrological conditions evolve. Hu et al. (2001) proposed a similar concept except model switching occurred based on discharge levels. Abrahart and See (2002) compared six flow amalgamation strategies (both switching and averaging) on two catchments. They determined that in flow forecasting, neural network methods improve predictive skill compared to the individual models if the flow regime is stable, whereas in volatile environments, a fuzzified probabilistic mechanism was the best tool. These applications are different from the simulation framework considered in this study as the averaging and prediction is balanced at each time step with the newly acquired information.

Other comparative studies have been published in the last few years on the subject of multi-model averaging (Bowler et al., 2008; Cavadias and Morin, 1986; Mylne et al., 2002; Raftery and Zheng, 2003; Raftery et al., 2005), especially in the hydrology and weather/climate prediction research fields. However, most of these use either a limited set of basins, of models or of model averaging methods (or some combination thereof). In this paper, we compare 9 model averaging techniques on 429 catchments from the MOPEX database using 4 hydrological models calibrated with 3 objective functions. The 3 objective functions are used to produce different parameterizations of the models. This allows diversifying the models' ability to target different parts of the hydrograph. Oudin et al. (2006) noted that models calibrated with two different objective functions produced flows that improved the overall simulation performance when combined adequately. Consequently 12-member ensembles are available for the model averaging methods. This large sample size will allow a better understanding of which methods are to be used in future applications.

2. Data, models and multi-model averaging methods

2.1. Basins, hydrometric and climate data

The hydrometric and climate data were collected from the MOPEX (Model Parameter Estimation Experiment) database (Duan et al., 2006) for 429 catchments ranging in size from 66 to 10,324 km^2. The dataset covers years 1949–2003, but many of these years are incomplete or missing. All available data was used for each of the catchments. The MOPEX database was designed to have a minimal density of stations per catchment, ensuring a certain level of quality in the dataset. Even years were used for the calibration period and validation was carried out on the odd years in the available time series. The opposite (calibration on odd years and validation on even years) was also tested but the results were practically identical, and are thus not presented here. In all cases, the first year in calibration and in validation was sacrificed for model warm-up.

The geographical extents of the catchments as well as their mean annual precipitation (mm) are shown in Fig. 1.

It can be seen that the average annual precipitation varies greatly depending on the region, with clear gradients across the US. Some catchments in the west coast receive more than 2000 mm of precipitation, while arid regions in south-central US receive less than 300 mm. The east–west gradient is also clear, with increasing precipitation values toward the east coast. Another lesser gradient is also observed in the north–south direction east of 95°W longitude. This information will be relevant for later analysis.

An overview of the hydrometeorological characteristics of the catchments in this study is presented in Table 1.

The potential evapotranspiration (PET) is taken directly from the MOPEX database and is based on the based NOAA Freewater Evaporation Atlas. Different PET estimation methods would also impact the aridity index, which is the ratio of potential evapotranspiration to total precipitation.

2.2. Hydrological models

Since the project required calibrating a large number of hydrological model/objective function combinations on an even larger set of basins, distributed models were not considered for this study, and five lumped models were retained. The five models are presented here.

2.2.1. GR4J-6

The GR4J model (Perrin et al., 2003) is an empirical and lumped, reservoir-based model. It was developed by the research group at CEMAGREF (now IRSTEA). It was conceived for water resources management and spring flood prediction for hydrologic applications. Initially, this model was parsimonious with only 4 parameters, with most secondary processes being represented by empirical constants. Since GR4J does not simulate snow accumulation or melt processes, a snow module (CEMANEIGE) was added to the basic model (Valéry, 2010; Valéry et al., 2014a,b) to make it applicable in northern basins. The GR4J model with the snow model has 2 more calibrated parameters, for a total of 6. This is the GR4J-6 model. Evapotranspiration must be fed to the model, as it does not estimate it itself. The Oudin formulation (Oudin et al., 2005) was used to pre-process the evapotranspiration data for the GR4J model variants.

2.2.2. HSAMI

The HSAMI model (Fortin, 2000) has been used by *Hydro-Quebec*, Quebec's hydroelectric company, for over three decades to forecast daily flows on more than one hundred basins in the province. It is also used in research applications such as climate change impact studies (Poulin et al., 2011; Arsenault et al., 2013), streamflow prediction at ungauged sites (Arsenault and Brissette, 2014) and water resources management (Minville et al., 2008, 2009, 2010). It simulates the entire hydrological cycle with a strong snow accumulation and melt model. Potential evapotranspiration is estimated using a confidential formulation