Journal of Hydrology 511 (2014) 72–81

Contents lists available at ScienceDirect

Journal of Hydrology

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

Monthly streamflow forecasting using Gaussian Process Regression

Alexander Y. Sun [a,*], Dingbao Wang [b], Xianli Xu [c,d]

[a] Bureau of Economic Geology, Jackson School of Geosciences, University of Texas Austin, Austin, TX 78713, United States
[b] Department of Civil, Environmental, and Construction Engineering, University of Central Florida, Orlando, FL 32816, United States
[c] Key Laboratory for Agro-Ecological Processes in Subtropical Region, Institute of Subtropical, Agriculture, Chinese Academy of Sciences, Changsha, China
[d] Huanjiang Observation and Research Station for Karst Ecosystem, Chinese Academy of Sciences, Guangxi, China

ARTICLE INFO

Article history:
Received 18 September 2013
Received in revised form 9 January 2014
Accepted 11 January 2014
Available online 20 January 2014
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Gaussian Process Regression
Machine learning theory
Water/energy interactions
Probabilistic streamflow forecasting
Hydrologic similarity

SUMMARY

Streamflow forecasting plays a critical role in nearly all aspects of water resources planning and management. In this work, Gaussian Process Regression (GPR), an effective kernel-based machine learning algorithm, is applied to probabilistic streamflow forecasting. GPR is built on Gaussian process, which is a stochastic process that generalizes multivariate Gaussian distribution to infinite-dimensional space such that distributions over function values can be defined. The GPR algorithm provides a tractable and flexible hierarchical Bayesian framework for inferring the posterior distribution of streamflows. The prediction skill of the algorithm is tested for one-month-ahead prediction using the MOPEX database, which includes long-term hydrometeorological time series collected from 438 basins across the U.S. from 1948 to 2003. Comparisons with linear regression and artificial neural network models indicate that GPR outperforms both regression methods in most cases. The GPR prediction of MOPEX basins is further examined using the Budyko framework, which helps to reveal the close relationships among water-energy partitions, hydrologic similarity, and predictability. Flow regime modification and the resulting loss of predictability have been a major concern in recent years because of climate change and anthropogenic activities. The persistence of streamflow predictability is thus examined by extending the original MOPEX data records to 2012. Results indicate relatively strong persistence of streamflow predictability in the extended period, although the low-predictability basins tend to show more variations. Because many low-predictability basins are located in regions experiencing fast growth of human activities, the significance of sustainable development and water resources management can be even greater for those regions.

1. Introduction

Streamflow forecasting plays a pivotal role in water resources planning and management. The capability to provide accurate and reliable streamflow forecasts over a flow regime has a direct impact on not only water allocation policies, but also sustainable economic development in the area. A major challenge of streamflow prediction stems from the fact that streamflow is a temporally lagged, spatial integral of runoff over a river basin (Milly et al., 2005) and, thus, can exhibit strong nonlinear dependency on hydrometeorological and anthropogenic factors. Existing methods for streamflow forecasting fall into three broad categories: physics-based methods, time series methods, and machine learning methods (Bourdin et al., 2012). Physics-based models are mathematical abstractions of physical processes that govern the water movement and storage in watersheds. These models typically require quantification and calibration of one or more conceptual models with uncertain physical parameters, leading to the challenge of equifinality (Beven and Freer, 2001). In addition, the theoretical foundation of many physics-basedmodels is small-scale physics, the application of which to larger watersheds is difficult due to "the effects of spatial heterogeneity in landscape properties, the inherent nonlinearity of many hydrological processes, and the process interactions at all scales" (Kirchner, 2006; McDonnell et al., 2007). Conventional time series methods are linear regression models that are best suited for short-term forecasting based on daily or weekly timescales, but not for long-term forecasting at seasonal and annual timescales, neither can they handle nonlinearity exhibited by rainfall-runoff models well (Hsu et al., 1995; Vogel et al., 1999; Zealand et al., 1999). These and other challenges/deficiencies associated with the traditional rainfall-runoff models and time series analyses partly explain the continued interest of the hydrologic community in machine learning methods.

Machine learning methods and, in particular, supervised learning methods, refer broadly to statistical techniques for developing predictive models using training data. Unlike physics-based models, machine learning methods are data-driven and rely almost

* Corresponding author.
E-mail address: alex.sun@beg.utexas.edu (A.Y. Sun).

0022-1694/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2014.01.023

exclusively on information embedded in training datasets. Artificial neural network (ANN) is one of the earliest machine learning methods adopted by the hydrologic community. Despite its popularity in streamflow forecasting (e.g., Chang and Chen, 2001; Hsu et al., 1995; Tokar and Markus, 2000), main issues of ANN include its tendency to overfit training data and instability with short training data records (Hsieh and Tang, 1998). An ultimate concern of all supervised machine learning algorithms is related to their generalization capability, which refers the capability of a trained model to deliver similar predictive performance on data not seen during training. Poor generalization may result from either overfitting or underfitting.

Recent decades have seen a surge of interest in the development of kernel-based machine learning methods. In particular, the support vector machine (SVM) algorithm (Vapnik, 1995) was introduced to address two challenges alluded in the above, namely, (a) how to establish a relationship between the size of training data and generalization performance of a trained model and (b) how to incorporate such knowledge in the training process to prevent overfitting. SVM projects the input data into a high or even infinite-dimensional space, such that the projected training data exhibit linearity and linear regression methods can be applied. An elegant feature of SVM is that the actual form of nonlinear mapping does not need to be known, and only their inner products (i.e., the so-called kernel function) are required to train an SVM model. This is known as the "kernel trick" in machine learning, which has served as a building block in all kernel-based methods (Bishop, 2006).

Both the SVM and ANN are deterministic algorithms per se and do not provide a direct quantification of prediction uncertainty. For the latter purpose, a common strategy is to create an ensemble of SVM or ANN models through certain resampling (e.g., bootstrapping and boosting) or random initialization techniques, and then use statistics of the ensemble models to quantify prediction performance (Sun, 2013; Zhou, 2012). Although ensemble methods can improve predictability of single models, they inevitably incur significant computational overhead. Alternatively, the regression problem may be cast into a probabilistic setting such that prediction uncertainty can be assessed directly. The relevance vector machine (RVM), originally improvised by Tipping (2001), represents a significant stride toward such direction.

RVM was designed to improve several deficiencies of the original SVM, including (a) predictions are not probabilistic, (b) the SVM solutions are not sparse enough, and (c) ad hoc procedures are needed for selection of hyperparameters in the SVM (note: in the current context, hyperparameters refer to parameters of the kernel or covariance functions). Like the SVM, RVM is a kernel method that parameterizes the unknown function as a weighted sum of nonlinear basis functions in the feature space. Unlike the SVM, RVM assumes that the weights are random variables and uses a Bayesian framework to estimate the posterior distribution of weights using data. So far, applications of the RVM in hydrological forecasting have been relatively limited. A notable work is the use of RVM in statistical downscaling of climate model outputs for predicting streamflow of several Indian river basins (Ghosh and Mujumdar, 2008).

A main limitation of the RVM is that it can yield unreliable results when a test data point is located far from the relevance vectors (i.e., the solution of RVM), in which case the predictive distribution will be a Gaussian with mean close to zero and variance also close to zero (Rasmussen and Williams, 2006). To mitigate the aforementioned issue of RVM, the Gaussian Process Regression (GPR) was introduced. The GPR is a full Bayesian learning algorithm that has received significant attention in the machine learning community for applications such as model approximation, multivariate regression, and experiment design (Girard et al., 2003; Quiñonero-Candela and Rasmussen, 2005; Rasmussen and Williams, 2006).

Gaussian processes (GP) assume that the joint probability distribution of model outputs is Gaussian. The notion of GP is not new in the hydrological literature. In fact, GP is underlying the kriging algorithm in classical geostatistics, the autoregressive moving average models (ARMA), Kalman filters, geostatistical inversion methods (Kitanidis, 1995), and radial basis function networks (Bishop, 2006). The ensemble Kalman filter (Evensen, 2003) and Gaussian particle filter (Kotecha and Djuric, 2003) may also be regarded as sequential versions of GP-based learning algorithms. Nevertheless, the GPR, which was originally formulated by Rasmussen and his coworkers, provides a "principled, practical, and probabilistic approach to learning in kernel machines" (Rasmussen, 1996; Rasmussen and Williams, 2006). The advantage of GPR over many other machine learning methods lies in its seamless integration of several machine learning tasks, including hyperparameter estimation, model training, and uncertainty estimation; thereby, the regression process is streamlined significantly and the results are less affected by subjectivity and more interpretable. Importantly, a suite of GPR tools are now available in the public domain for various applications (Rasmussen and Nickisch, 2010). In comparison, similar methods mentioned in the above usually only address certain aspects of the regression/prediction problem.

GPR can be considered a type of multivariate regression techniques. In this sense, GPR is closely related to generalized least squares, which has been used extensively in the so-called regional regression analysis in hydrology (e.g., Reis et al., 2005; Stedinger and Tasker, 1985; Vogel et al., 1999). However, most existing studies parameterize the predictand as a linear combination of (transformed) predictors and then estimate the linear coefficients. In contrast, GPR expresses the unknown as a linear combination of nonlinear basis functions, as we shall see in Section 2. The application of GPR in streamflow forecasting has been rather limited. The Bayesian joint probability method proposed recently by Wang and his coworkers (Robertson and Wang, 2012; Wang et al., 2009; Wang and Robertson, 2011) used Bayesian inference to predict seasonal streamflow. However, the authors mainly focused on learning parameters of an enhanced Box-Cox transform using Monte Carlo Markov chain sampling and did not adopt a kernel-based machine learning approach in their work.

The main objective of this work is twofold. First, the efficacy of GPR is demonstrated using data collected as part of the Model Parameter Estimation project (MOPEX), which includes long-term hydrometeorological time series from a large number of unregulated basins located in different climatic regions across the U.S. (Duan et al., 2006; Schaake et al., 2000). We show that a relatively simple and fixed group of predictors can already give satisfactory streamflow prediction over the majority of MOPEX basins at the monthly scale. The performance of GPR is then compared to two streamflow forecasting algorithms, autoregressive moving average with exogenous variables (ARMAX) and multilayer perceptron (MLP) neural network model. The former is a widely used linear regression algorithm and the latter is a type of ANN algorithm. For completeness, brief summaries of ARMAX and MLP algorithms are provided in Appendices A and B, respectively. More details of the two algorithms can be readily found in many textbooks (e.g., Haykin, 1994; Loucks et al., 1981).

The second purpose of this work is to offer a systematic analysis of factors that can potentially affect basin streamflow predictability, which has been the subject of immense interest in recent years under topics such as hydrologic similarity (e.g., Berger and Entekhabi, 2001; Blöschl and Sivapalan, 1995; Olden et al., 2012; Oudin et al., 2010; Wagener et al., 2007), catchment-scale water and energy partition (Sankarasubramanian et al., 2001; Zhang et al., 2001), prediction at ungauged basins (Li et al., 2011; Patil and