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

Stieglitz, 2012; Sivapalan, 2003; Sivapalan et al., 2003), and impact of climate change on streamflow (Sivapalan et al., 2011; Wang and Hejazi, 2011; Xu et al., 2013).

The rest of this paper is organized as follows. Section 2 presents data processing techniques and describes formulation of GPR. Section 3 provides results and discussion and, finally, Section 4 summarizes main findings.

2. Materials and methods

2.1. Data and data processing

The MOPEX database used in this study includes long-term hydrometeorological records from 438 basins across the U.S. (Fig. 1) and is downloable from ftp://hydrology.nws.noaa.gov. The original goal of the MOPEX project was to test techniques for estimating parameters used in land surface parameterization schemes of atmospheric models and in hydrological models. During the course of the project, the MOPEX team had developed a comprehensive database that can be used to test any streamflow forecasting framework. For each MOPEX basin, the available data consist of basin-averaged daily hydrometeorological data (streamflow, precipitation, minimum and maximum temperature, and potential evaporation) from January 1, 1948 to December 31, 2003, as well as basin characteristics data (e.g., topography, soil type, vegetation type, and land use/land cover types). The drainage areas of the basins range from 66 to 10,400 km^2. MOPEX streamflow data were acquired from the U.S. Geological Survey (USGS) stream gauge network, whereas precipitation and temperature data were compiled from U.S. National Climate Data Center (NCDC) resources. A detailed account of basin selection criteria and data preparation procedure was provided in Duan et al. (2006) and Schaake et al. (2000). During pre-processing, a small number of gauges were removed because of invalid streamflow records, leaving a total of 430 basins to be actually used in this study. Daily time series were then aggregated into monthly data. A large number of gauges (261) have continuous records throughout the MOPEX period, while the rest of the gauges have one or more missing records.

More than a decade has passed since the end of the original MOPEX data period. A salient question is whether changes in the last decade have affected predictability. Flow regime modification due to climate change and anthropogenic activities has received increased attention in recent years because of its adverse impact on streamflow predictability (Botter et al., 2013; Sivapalan et al., 2011). Climate change is manifested in redistribution of precipitation, as well as temperature change; whereas anthropogenic activities include energy production, irrigation, and others. Several studies confirmed that the annual minimum and median daily streamflow at many stream gauges in the U.S. have exhibited noticeable changes since 1970s (McCabe and Wolock, 2002; Villarini et al., 2009). To examine the persistence of predictability of MOPEX basins, we extended the MOPEX hydrometeorological records from January 2004 to December 2012 and performed additional testing of the trained GPR models using the extended time series. The streamflow data of the extended period were downloaded from the USGS stream gauge network (http://waterwatch.usgs.gov/). Precipitation and temperature data were extracted from the PRISM dataset (4-km resolution) (PRISM Climate Group, 2012). PRISM data were further aggregated and averaged for each basin by using the basin boundary files that were distributed with the original MOPEX database.

2.2. Gaussian Process Regression (GPR)

Let us consider a predictor group $\mathbf{x}$ consisting of d variables. The objective of a typical machine-learning method is to "learn" the functional relationship between the d-dimensional predictor $\mathbf{x} \in \mathbb{R}^d$ and the target variable y,

$$y = f(\mathbf{x}), \tag{1}$$

where f is the unknown function and $\mathbb{R}$ denotes real space. A discrete approximation of the unknown function f is given by the following linear combination of basis functions

$$\hat{f}(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}), \tag{2}$$

in which $\{\phi_j(\mathbf{x})\}_{j=1}^{M}$ is a set of basis functions that can be either linear or nonlinear; $\mathbf{w} = [w_1, \ldots, w_M]^T$ is the unknown weight vector; and M is the number of basis functions used to approximate f. Assuming additive model error, we have the following model

$$y = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}) + \varepsilon, \tag{3}$$

in which ε is the error term. The unknown weights $\mathbf{w}$ can be estimated by using a set of training data that include N predictor observations, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ (i.e., each row of the d-column matrix $\mathbf{X}$ is an observation of $\mathbf{x}$), and co-observed predictand values, $\mathbf{y} = [y_1, \ldots, y_N]^T$. Eq. (2) is general and covers a wide range of linear and nonlinear regression algorithms, including ARMAX, ANN, and kernel-based methods. In the latter case, the role of basis function $\phi(\mathbf{x})$ can be seen as a transformation that projects $\mathbf{x}$ from the original input space into a high-dimensional feature space. The actual form of basis function, however, is not needed, as it will be shown below.

As mentioned before, the building block of GPR is GP, which assumes Gaussian priors for (transformed) function values (Rasmussen and Williams, 2006). Thus, a GP is completely specified by its second-order statistics,

$$f(\mathbf{x}) \sim \mathrm{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{4}$$

where $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ are the mean and covariance function of f, respectively. By definition, any finite subset of a GP has a joint Gaussian distribution. Thus, if $\mathbf{f} = \{\hat{f}(\mathbf{x}_i, \mathbf{w})\}_{i=1}^{N}$ denotes the model outputs corresponding to the input dataset $\mathbf{X}$,

$$\hat{f}(\mathbf{x}_i, \mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j(\mathbf{x}_i), \quad i = 1, \ldots, N \tag{5}$$

or simply,

$$\mathbf{f} = \mathbf{\Phi w} \tag{6}$$

then the prior distribution of $\mathbf{f}$ is Gaussian

$$p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}) \tag{7}$$

In Eqs. (5)–(7), the $N \times M$ matrix $\mathbf{\Phi}$ is referred to as the design matrix, and each row of $\mathbf{\Phi}$ contains the outputs of basis functions corresponding to the input $\mathbf{x}_i$, viz.

$$\boldsymbol{\phi}_j = [\phi_1(\mathbf{x}_i), \phi_2(\mathbf{x}_i), \ldots, \phi_M(\mathbf{x}_i)], \quad j = 1, \ldots, N \tag{8}$$

The mean of $\mathbf{f}$ is assumed to be zero and the $N \times N$ matrix $\mathbf{K}$ is a covariance matrix of $\mathbf{f}$, with its hyperparameters denoted by $\boldsymbol{\theta}$. More specifically, from Eq. (6) the covariance matrix $\mathbf{K}$ can be written as

$$\mathbf{K} = \mathbf{\Phi} E(\mathbf{w}\mathbf{w}^T)\mathbf{\Phi}^T = \mathbf{\Phi}\mathbf{\Sigma}_w\mathbf{\Phi}^T \tag{9}$$

where the $M \times M$ matrix $\mathbf{\Sigma}_w$ is the covariance matrix of the weight vector $\mathbf{w}$, and the second equality sign indicates that $\mathbf{K}$ is an inner product with respect to $\mathbf{\Sigma}_w$. Hyperparameters $\boldsymbol{\theta}$ are specific to the actual covariance structure used for $k(\mathbf{x}, \mathbf{x}')$. Note that the zero-

mean prior of $\mathbf{f}$ used in Eq. (7) is more for convenience than a restricting assumption because (a) we can always normalize $\mathbf{f}$ by using appropriate scaling such that the mean of $\mathbf{f}$ becomes zero for fixed $\mathbf{x}$ and (b) the posterior mean is often nonzero. If supported by prior information, however, a nonzero mean term can be specified by adding extra terms in Eq. (2), after which the zero-mean GP is then applied to the difference between $f(\mathbf{x})$ and the constant mean or trend function. In practice, it is often the case that the unknown trend parameters are estimated simultaneously with other parameters.

If the model error in Eq. (3) is independent and identically Gaussian distributed, the likelihood function of the training target vector $\mathbf{y}$ also becomes Gaussian

$$p(\mathbf{y}|\mathbf{f},\sigma^2) \sim \mathcal{N}(\mathbf{f},\sigma^2\mathbf{I}) \tag{10}$$

where σ^2 is the variance of model error ε and $\mathbf{I}$ is the identity matrix. The posterior distribution of $\mathbf{f}$ can then be obtained by applying Bayes' rule

$$p(\mathbf{f}|\mathbf{y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\sigma}^2) = \frac{p(\mathbf{y}|\mathbf{f},\sigma^2)p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta},\sigma^2)} \tag{11}$$

In this case because both the prior and likelihood function are Gaussian, the posterior distribution of $\mathbf{f}$ is also Gaussian, and its mean and covariance are obtained by substituting (7) and (10) into (11) (Rasmussen and Williams, 2006)

$$\mu = \mathbf{K}^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y} \tag{12}$$

$$\boldsymbol{\Sigma} = \mathbf{K} - \mathbf{K}^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{K} \tag{13}$$

Note that neither $\phi(\mathbf{x})$ nor $\mathbf{w}$ appears explicitly in Eqs. (12) and (13). This is because both terms appear in an inner product term that defines the covariance matrix $\mathbf{K}$. Thus, the main effort in GP modeling is now shifted from determining the actual basis functions (including the dimension M) and their weights to determining the structure and hyperparameters of the covariance function.

The covariance function $k(\cdot,\cdot)$ is also referred to as the kernel function in machine learning. Some commonly used kernel functions in the GPR literature include the squared exponential or Gaussian kernel (Rasmussen and Williams, 2006),

$$k(\mathbf{x},\mathbf{x}'|\boldsymbol{\theta}) = \sigma_f^2 \exp\left(-\frac{1}{2}\frac{r^2}{l^2}\right), \quad \boldsymbol{\theta} = (\alpha, l, \sigma_f^2) \tag{14}$$

and the Matérn family of covariance functions

$$k(\mathbf{x},\mathbf{x}'|\boldsymbol{\theta}) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{l}\right), \quad \boldsymbol{\theta} = (\nu, l, \sigma_f^2) \tag{15}$$

In the above equations, $r = |\mathbf{x}-\mathbf{x}'|$ is Euclidean distance between two input points $\mathbf{x}$ and $\mathbf{x}'$; θ denotes the collection of hyperparameters associated with each covariance function; K_ν is modified Bessel function; Γ is gamma function; l is correlation length; σ_f^2 is variance; ν controls the degree of differentiability. The process noise variance σ^2 (see Eq. (10)) is an additional parameter that is determined during training. The marginal probability can be computed by integration over $\mathbf{f}$ (Rasmussen and Nickisch, 2010),

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f},\sigma^2)p(\mathbf{f}|\mathbf{X},\boldsymbol{\theta})d\mathbf{f} \tag{16}$$

from which the log marginal likelihood is obtained as

$$\log p(\mathbf{y}|\mathbf{X}) \propto -\frac{1}{2}\mathbf{y}^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y} - \frac{1}{2}\log|\mathbf{K}+\sigma^2\mathbf{I}| - \frac{N}{2}\log(2\pi) \tag{17}$$

The unknowns θ and σ^2 can then be estimated from Eq. (17) by using a gradient-based algorithm.

Having determined the posterior of $\mathbf{f}$ through training, we can evaluate the predictive distribution of any new test data (denoted by $\mathbf{x}_*$) conditioned on training results, namely,

$$p(f_*|\mathbf{x}_*,\mathbf{y},\mathbf{X},\theta,\sigma^2) \tag{18}$$

It can be shown that the predictive distribution Eq. (18) is Gaussian, with its mean, m, and variance, v^2, given by (Rasmussen and Williams, 2006):

$$m(\mathbf{x}_*) = \phi(\mathbf{x}_*)^T\mu = \mathbf{k}_*^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y} \tag{19}$$

$$v^2(\mathbf{x}_*) = \phi(\mathbf{x}_*)^T\boldsymbol{\Sigma}\phi(\mathbf{x}_*) = k_{**} - \mathbf{k}_*^T(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{k}_* \tag{20}$$

where $\mathbf{k}_* = [k(\mathbf{x}_*,\mathbf{x}_1),\ldots,k(\mathbf{x}_*,\mathbf{x}_N)]^T$, $k_{**} = k(\mathbf{x}_*,\mathbf{x}_*)$, and μ and $\boldsymbol{\Sigma}$ are the posterior mean and variance of $\mathbf{f}$ defined in (12), (13). The predictive Eqs. (19) and (20), constitute the main results of the GPR. As mentioned in the Introduction section, the GPR equations are the same as the kriging equations used in spatial statistics (Cressie, 1993). In fact, the predictor variables in the latter case are spatial coordinates.

We remark that (a) an advantage of the GP model is that it is more immune to missing input values because training data are "pooled" to estimate hyperparameters of covariance function; (b) in principle the appropriateness of a particular covariance function can be tested through cross-validation or other model selection techniques; in practice, however, it has been found that GP is not very sensitive to different choices of covariance functions for time series modeling (Shi et al., 2007), which is also our experience in this work; and (c) the use of covariance function can be seen as a regularization mechanism, making the performance of a GP model more robust than other machine learning methods such as ANN. In this work, the Matlab toolbox GPML developed by Rasmussen and Nickisch (2010) was used to develop and train GP models. The prior mean was assumed zero (on the normalized data) and the kernel function used is squared exponential (i.e., Eq. (14)).

2.3. Predictor selection

Predictor selection represents a key step in developing data-driven streamflow forecasting models. In general, predictors represent two sources of information that may contribute to streamflow predictability, the influence of initial catchment conditions and the effect of climate during the forecasting period (e.g., Piechota et al., 2001; Robertson and Wang, 2012). Antecedent streamflow, precipitation, and temperature are the most widely used predictors of the former group, while the climate group includes a large number of climate indices that may influence future precipitation conditions. The effect of climate indices on streamflow predictability are seasonal- and regional-dependent and have been the subject of numerous climate teleconnection studies (e.g., Barlow et al., 2001; Chiew and McMahon, 2002; Hidalgo and Dracup, 2003). A main purpose of this work is to demonstrate the efficacy of GPR for forecasting at the monthly level. Thus, we focused on predictors representing the initial catchment conditions.

A convenient way for characterizing the stations is to associate them with water resources regions, which have been delineated for the continental U.S. to assess the state of the water resources (see Fig. 1 and Table S1). Previous studies noted that the 18 water resources regions are either temperate or humid, with the exception of the midwestern and southwestern regions which are semiarid or arid (see regions 7–16 in Fig. 1 and Table S1) (Sankarasubramanian and Vogel, 2003; Sankarasubramanian et al., 2001). Within each of the water resources regions, Vogel et al. (1998) showed that the streamflow persistence is relatively homogeneous. In lieu of a region specific model selection process, we mainly tested two predictor groups. The first predictor group (denoted as PD-I hereafter)

mainly exploits persistence of predictors representing initial catchment conditions

$$Q_{t-1}, Q_{t-2}, P_{t-1}, T_{\max t-1}, T_{\max t-2}, \text{ and } T_{\min t-1} \quad (21)$$

in which Q represents streamflow; P is basin-averaged precipitation; $T_{\min}$ and $T_{\max}$ are basin-averaged maximum and minimum temperatures, respectively; the subscripts represent the lags in months and t is the month to be predicted. The particular predictor group in Eq. (21) was selected after testing the predictors at different lags.

Runoff generation theory dictates that in arid and semiarid regions, hydrograph is dominated by direct precipitation, return flow, and surface runoff, whereas in humid regions baseflow plays a more pronounced role (Dunne, 1983; Wagener et al., 2007; Wang and Wu, 2012). If the flow regimes are stationary, long-term averages of precipitation and temperature corresponding to the prediction month (i.e., t) will reflect expected precipitation for the catchment and may help improve prediction. Thus, in the second predictor group (denoted as PD-II hereafter) we included long-term monthly averages of all three types of predictors,

$$Q_{t-1}, Q_{t-2} P_{t-1}, T_{\max,t-1} T_{\max,t-2} T_{\min,t-1} \bar{P}_t \bar{T}_{\max,t} \bar{T}_{\min,t} \quad (22)$$

in which the overbar represents long-term averages for the predicting month. The performance of the two models is compared in Section 3.

2.4. Performance metrics

Goodness-of-fit is calculated on the testing data using two metrics. The first is the standard Nash–Sutcliffe efficiency (NSE) defined as

$$NSE = 1 - \frac{\sum_{i=1}^{n} (Q_i - Q_{o,i})^2}{\sum_{i=1}^{n} (Q_{o,i} - \bar{Q}_o)^2} \quad (23)$$

where $Q_{o,i}$ and Q_i are the observed and predicted streamflow, respectively, and $\bar{Q}_o$ represents mean observed value. NSE quantifies the skill of a model to explain streamflow variance, as compared to a reference model using $\bar{Q}_o$. The NSE is known to be sensitive to extreme values. The second metric is mean cumulative error of the model, or water balance (WB) error, which is defined as

$$WB = 1 - \left| 1 - \sum_{i=1}^{n} Q_i \Big/ \sum_{i=1}^{n} Q_{o,i} \right| \quad (24)$$

WB measures the ability of a model to correctly reproduce streamflow volumes over a testing period (Oudin et al., 2005). Both NSE and WB range from $-\infty$ to 1.

3. Results and discussion

3.1. GPR performance validation

A separate GP model was developed for each MOPEX basin. For each model, 70% of the predictor-predictand data pairs were used for training and the rest for testing. Before training, the streamflow data were normalized using Box-Cox transform and all variables were linearly scaled to the interval $[-1, 1]$. After testing, the results were transformed back to the original input space to calculate the performance metrics. Training of a GP model typically took less than 5 s on a PC equipped with Intel Core-i7 CPU.

We experimented with both predictor groups (i.e., PD-I and PD-II) that were described under Section 2. Overall, the inclusion of long-term averages in PD-II led to better performance for most of the water resources regions, although the improvement is marginal for some regions (see Fig. S1). We thus report results from PD-II for the rest of this study.

Fig. 1 shows a map of all MOPEX stations, which are classified into quartiles according to the NSE obtained. A histogram of NSE values is provided in Fig. S2. Results suggest that basins located in the Pacific Northwest and the eastern U.S. tend to exhibit better predictability than those located in the Midwest. This pattern is generally in line with previous studies. For example, from a hydrologic similarity perspective, Patil and Stieglitz (2012) found that the high predictability catchments are confined to the Appalachian Mountains in eastern US, the Rocky Mountains, and the Cascade Mountains in the Pacific Northwest, whereas low predictability catchments are located mostly in the drier regions west of Mississippi river.

To showcase the performance of GPR, we selected six out of a set of 12 MOPEX basins that have often been used in the literature to "form a hydroclimatic gradient" in the eastern U.S. (e.g., Herman et al., 2013; Nasonova et al., 2009). Table 1 lists detailed information pertaining to each basin. Climates of the six selected basins range from very wet (e.g., the French Broad basin in North Carolina) to very arid (e.g., the Guadalupe basin in Texas). The effect of anthropogenic activities on the six basins is considered insignificant (Herman et al., 2013).

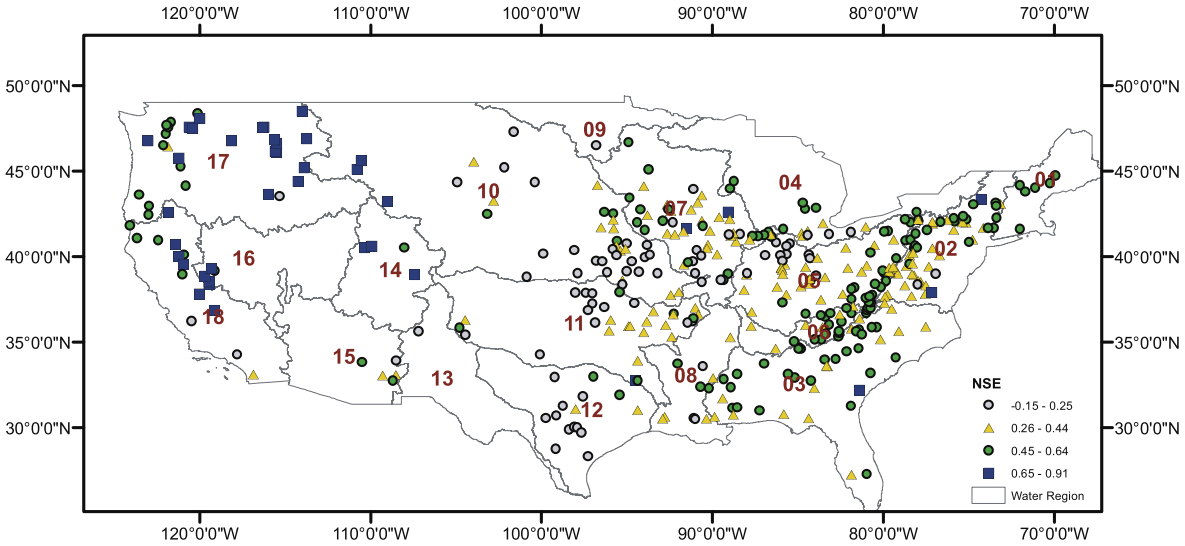

Fig. 1. Map of MOPEX stations, where each station is colored by the NSE value obtained by its one-month-ahead GP model for the testing data. Results are classified into quartiles. The background map shows 18 water resources regions covering the continental U.S.

Table 1
Gauge information for plots shown in Fig. 2.

Station ID	Lon. (deg)	Lat. (deg)	Area (km^2)	Annual Q/P	Annual PET/P	Gauge Info
03451500	−82.58	35.61	2445	0.5	0.54	French Broad River at Asheville, NC
03054500	−80.04	39.15	2361	0.56	0.54	Tygart Valley River at Philippi, WV
03179000	−81.01	37.54	1024	0.42	0.76	Bluestone River near Pipestem, WV
03364000	−85.93	39.20	4419	0.37	0.83	East Fork White River at Columbus, IN
07186000	−94.57	37.25	2999	0.26	1.01	Spring River near Waco, MO
08167500	−98.38	29.86	3457	0.13	1.98	Guadalupe River near Spring Branch, TX

The last two columns in Table 1 list the ratio between the mean annual Q and P (i.e., runoff ratio) and the ratio between mean annual potential evaporation to P (i.e., aridity index, PET/P), respectively. Budyko (1974) demonstrated that the evaporation ratio (i.e., 1 – runoff ratio) is primarily controlled by the aridity index. For basins with aridity index less than 1, the energy supply is the limiting factor for evaporation, whereas for basins with aridity index greater than 1, water supply is the limiting factor. Table 1 suggests that all but the last station have aridity index less than or equal to 1.

Fig. 2 shows the GP model prediction and the 95% confidence envelope for the six stations. The NSE tends to improve when moving from dry to humid regions, which are characterized by higher runoff ratios and lower aridity indices. From a different angle, we observe that the GP model tends to perform better on those stations exhibiting more persistent (e.g., the top row in Fig. 2) than erratic flow regimes (e.g., the last row in Fig. 2). Erratic regimes are expected in fast-responding catchments during seasons with sporadic rainfall events, or during hot humid seasons; conversely, persistent regimes are expected during humid, cold seasons in slow-responding basins (Botter et al., 2013). The variability of erratic regimes is much more significant than that of the persistent regimes and, thus, is less predictable. Many Midwest basins fall into the category of erratic regimes, and the runoff generation of which is dominated by direct precipitation. For most cases, the 95% confidence envelope obtained by GPR captures streamflow variations adequately during the testing period, except for very flashy flooding events (e.g., the Guadalupe basin in 2003). Overall, the pattern of the WB metrics is consistent with that of the NSE, which shows higher values for the first row and lower values for the last row in Fig. 2.

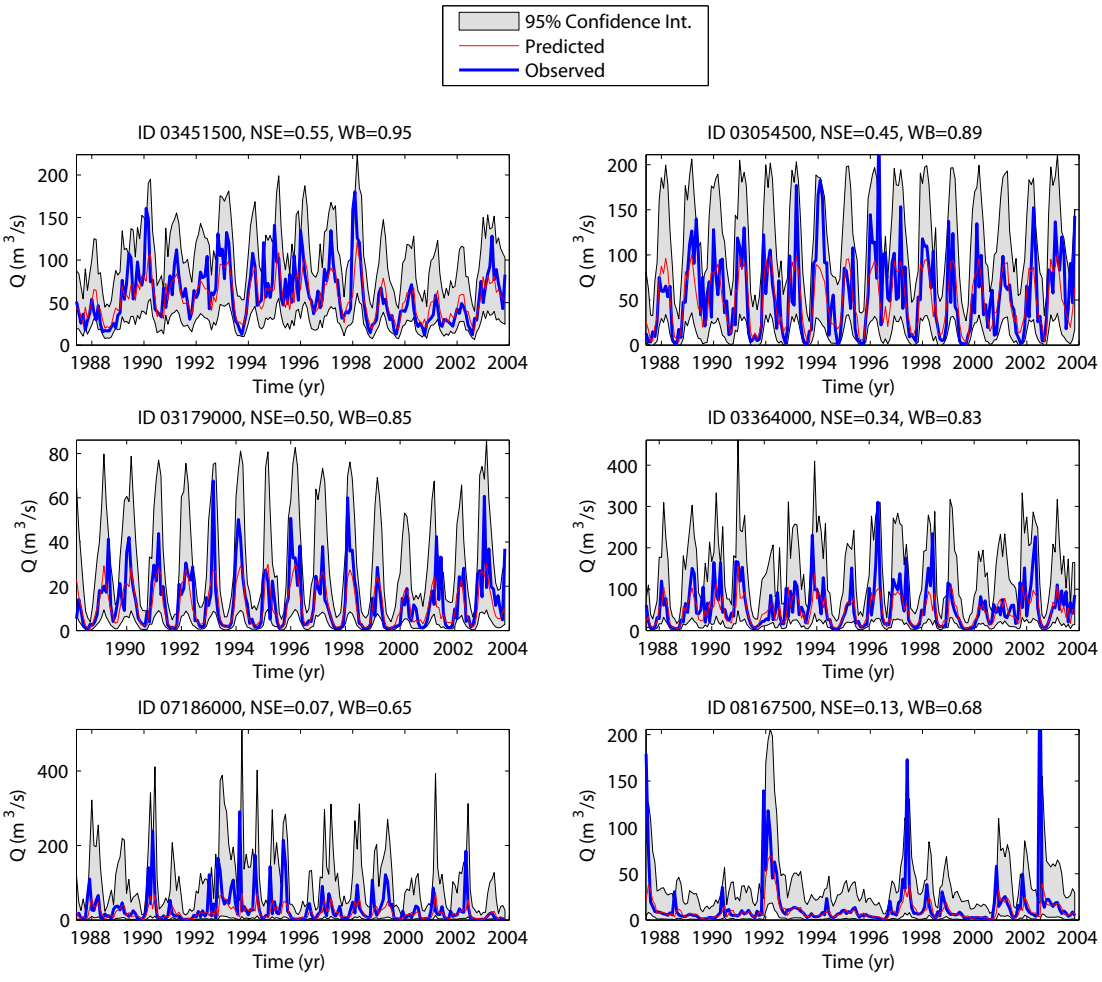

Fig. 2. Comparison between observed (thick blue line) and predicted (thin red line) streamflow for the testing period for six MOPEX stations located in different climatic regions (see Table 1 for details). Shaded areas correspond to 95% confidence intervals of GP models. Station ID, NSE, and WB are labeled on subplot titles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Next, the performance of GPR was compared to that of the ARMAX and MLP on all MOPEX stations possessing continuous observation records during the MOPEX period. The ARMAX models were trained using the Matlab function armax, whereas the MLP models were trained using the Matlab neural network toolbox (Demuth et al., 2008). All MLP models were assigned a single hidden layer with three hidden neurons, a structure that was determined based on the number of predictors used in the current problem and trial-and-error (Sun, 2013). Although only a single MLP model is used per station, the performance of MLP over many similar stations can indirectly reflect its generalization performance. Results show that GPR outperformed both ARMAX and MLP in most cases (Fig. 3). The four noticeable GPR underperformers shown in Fig. 3b are all located in the semiarid central Texas, including Station 08167500 that was shown in Fig. 2; these basins all belong to the erratic regimes mentioned previously and are less predictable. Overall, however, the superior performance of GPR over the other two models is encouraging, suggesting the merits of the underlying Bayesian approach.

The generalization capability of all trained GP models was further tested using data from extended period, which spans from the beginning of 2004 to the end of 2012. Ideally, if the flow regimes remain little changed and the trained models are not overfitted, the performance of the models should stay relatively unchanged. Fig. 4 shows a scatter plot that compares the performance of GP models during the extended period to that during the original testing period. Correlation coefficient between the two series is 0.76, which is relatively satisfactory. Nevertheless, we observe a significant scatter. The scatter pattern shown in Fig. 4 suggests that the higher the original NSE, the more likely the performance of a GP model will persist into the extended period. Conversely, the lower the original NSE, the more unstable the prediction capability tends to be. Thus, the pattern shows a cone shape along the 45° line, with the tip of cone pointing to the upper-right direction. This phenomenon may be attributed to several factors: (a) physically, a lower NSE indicates that a flow regime is more erratic and, thus, is less predictable in the first place; (b) basins having lower NSEs tend to locate in dryer areas where sporadic rainfalls create high flows that deviate significantly from the basins' nominal flows, making Gaussian distributions less suitable even after variable transformation; (c) semiarid areas tend to be more vulnerable to anthropogenic impacts, which may have altered watershed responses and flow regimes in recent years; and (d) data-driven models trained using historic observations may no longer serve as sufficient guides for future conditions, known as the effect of nonstationarity (Wagener et al., 2010). Recently, Wang and Hejazi (2011) found that the potential impact of human

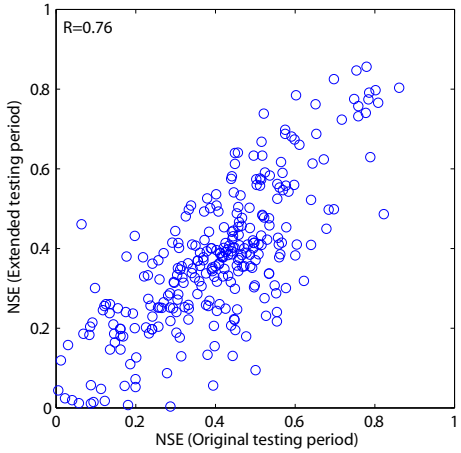

Fig. 4. Comparison of GP model performance on the original MOPEX testing data and those in the extended period (2004–2012).

activities on MOPEX streamflows is more significant than previously thought; even though the percentage of urbanization is small over most of the basins, irrigation and crop land management activities may cause shifts in mean annual streamflow. In reality, the change in predictability is likely to be a result of interactions among multiple factors identified herein. In the next subsection, we will take a closer look of basin characteristics that may influence predictability.

3.2. Factors affecting GPR predictability

The Budyko framework provides a simple and yet powerful model for studying long-term water-energy partition (i.e., partitioning of precipitation into evaporation and runoff) at the basin scale (Budyko, 1974). The dimensionless indices (i.e., evaporation ratio and aridity index calculated using annual averages) also provide a means for exploring hydrologic similarity among basins (Patil and Stieglitz, 2012; Wang and Wu, 2012). Fig. 5a shows a Budyko diagram of MOPEX basins, in which only NSEs that are less than the lower quartile (squares) and greater than the 3rd quartile (circles) are shown. The plot shows a rather distinctive pattern and echoes our previous findings based on Fig. 1, namely, basins exhibiting the best predictability tend to be energy-limited, while those exhibiting the worst predictability tend to be water supply-limited and are mostly located in arid and semiarid regions. Thus, the long-term water and energy partitions of a basin also shed light on its

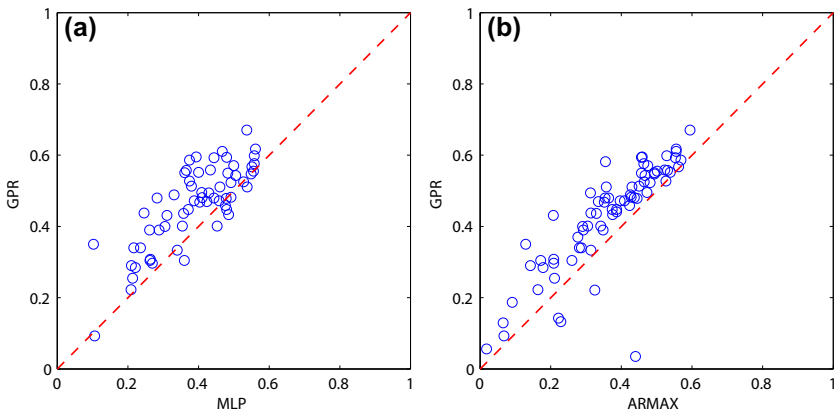

Fig. 3. Comparison of NSE obtained by (a) GPR and MLP and (b) GPR and ARMAX, for original MOPEX testing data.

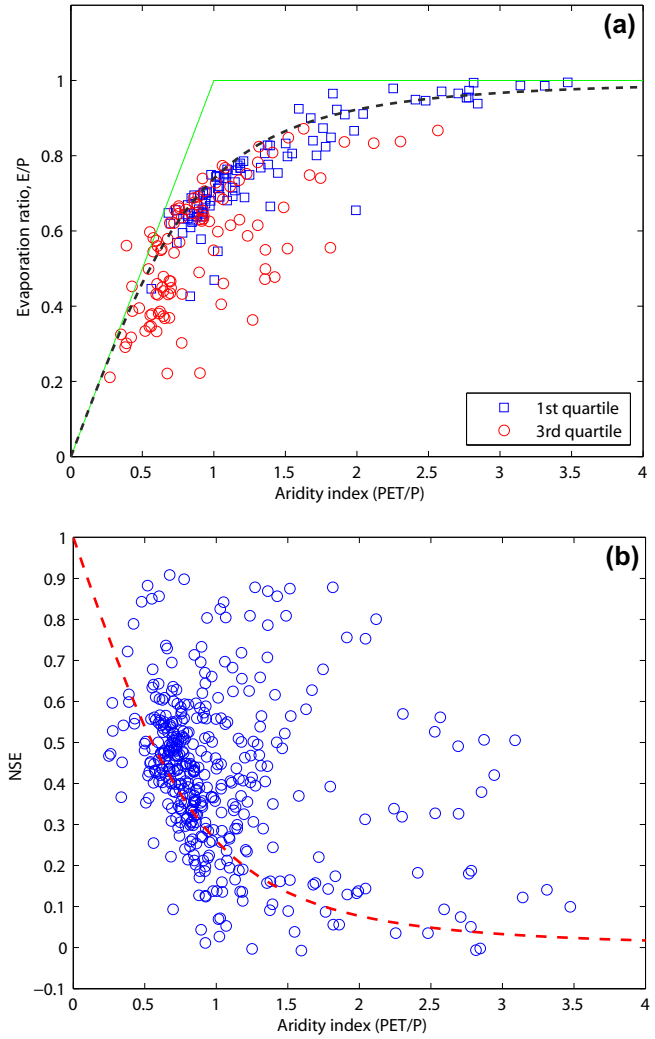

Fig. 5. (a) Budyko diagram for illustrating NSE similarity, where square and circle symbols correspond to NSE's <1st and >3rd quartiles, respectively. The horizontal axis is aridity index and vertical axis is evaporation ratio. Budyko curve is the gray dash line and the limit lines are in green; (b) NSE vs. aridity index.

predictability at the monthly scale. In Fig. 5a, the data were fitted to a Turc-Pike type of Budyko function (Pike, 1964)

$$\frac{E}{P} = \left[1 + \left(\frac{PET}{P}\right)^{-\lambda}\right]^{-1/\lambda} \tag{25}$$

where the estimated parameter λ is 2.3. Fig. 5b plots all NSE values as a function of the aridity index. The data were fitted by using the complementary of the above relationship in Eq. (25), following a similar approach taken by Wang and Wu (2012). The results show that NSE has a clear dependence on aridity index.

To examine the effect of land surface characteristics, we plotted NSE as a function of average greenness fraction, which is the average of monthly fractional vegetation coverage derived from NDVI (Gutman and Iganatov, 1998) (Fig. 6). Vegetation coverage reflects climate seasonality, with lower coverage commonly associated with "larger seasonal phase mismatch between precipitation and radiation" and reduced evapotranspiration; the effects of soil texture and topography, which regulate runoff generation and the water available for vegetation, are also reflected by vegetation coverage (Li et al., 2013; Williams et al., 2012). The plot in Fig. 6, which includes basins greater than 5000 km^2, suggests that a weak linear pattern (correlation coefficient 0.29 and p value of 0.01) exists

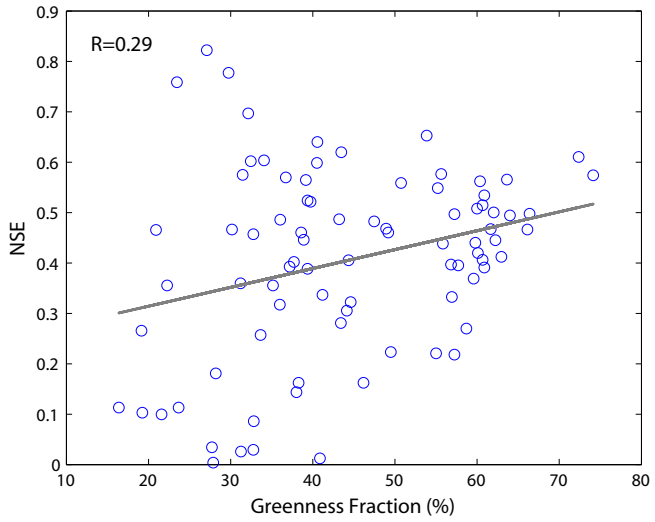

Fig. 6. NSE vs. greenness fraction, where a linear regression model is fitted to the data (slope 0.26 and intercept 0.0).

between the two variables. This may be attributed to the fact that vegetation control plays a major role only for large river basins (>300,000 km^2), while for small catchments the "key ecohydrological processes influencing the water and energy balances are more localized and diverse" (Li et al., 2013).

Fig. 7 shows the relationship between NSE and the dominant land use/land cover (LULC) class at each basin. The University of Maryland vegetation classification system was used (Schaake et al., 2000). An immediate observation from Fig. 7 is that basins with high NSE can almost belong to any of the LULC classes with significant samples. On average, however, the Evergreen Needleleaf Forest class (Category 1) achieved the best NSE, followed by the Woodland class (Category 6); in contrast, the Grassland (Category 10) and Cropland (Category 11) classes are the worst performers. The latter two LULC classes are typical of many basins located in southwestern and midwestern U.S. Thus, observations made here are consistent with the spatial NSE distribution plotted in Fig. 1 and the Budyko diagram presented in Fig. 5.

In addition to the basin characteristics mentioned in the above, we also examined the correlation between NSE and basin size. However, no clear relationship was identified. Basin size did not play a significant role in the extended period either. The improvement or degradation of NSE over the extended period occurred for basins of all sizes.

4. Summary and conclusion

A kernel-based machine learning algorithm, GPR, is applied to perform one-month-ahead streamflow forecast. Given a set of training data, GP provides a flexible Bayesian framework for identifying nonlinear relationship between predictors and predictand. GP models are not only conceptually easier to understand, but also give estimates of prediction uncertainty.

To demonstrate the efficacy of GPR, we developed GP models for more than 400 MOPEX basins across the U.S. Most of the basins studied have more than 40 years of hydrometeorological observations. Results indicate that GPR outperformed both ARMAX and MLP in most cases. The spatial pattern of NSE reflects hydroclimatic and vegetation controls, with basins located in the Pacific Northwest and eastern U.S. generally having higher predictability than those located in the Midwest. The pattern of NSE can also be well explained by the Budyko diagram, which is a generic framework

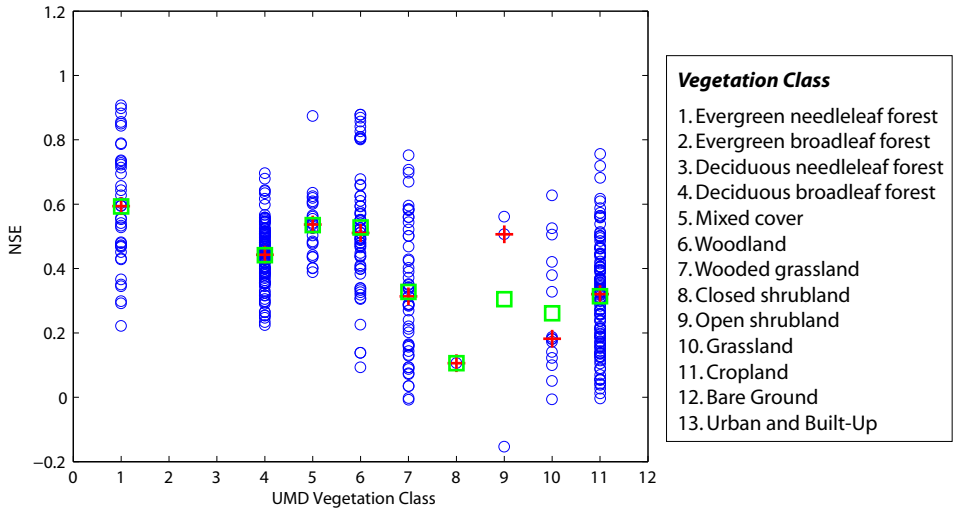

Fig. 7. NSE (open circles) for different vegetation classes listed under the University of Maryland classification system (see legend), where mean and median of NSE's are labeled by square and plus symbols, respectively.

for modeling long-term water/energy partition at a watershed. Basins having the best predictability tend to be energy-limited (i.e., aridity index less than 1), while those exhibiting the worst predictability tend to be water supply-limited.

The changing environment and its impact on streamflow predictability have received considerable attention in recent years. A basin can potentially move in all directions on the Budyko diagram, driven by climate change and human impacts. We extended the original MOPEX database from 2003 to the end of 2012, and performed additional testing using the GP models. Results indicate basins originally exhibiting higher predictability tend to persist into the extended period, indicating a certain degree of stationarity or resilience to changes. Nevertheless, from the water availability perspective, basins located in the Midwest and southern U.S. are more of concern to water managers because of the predicted population growth and increasing energy production activities in those regions. If predictability is low, the risks of not meeting water demands will be high. The implication is that more prudent water planning and conservation measures need to be instituted to reduce water shortage risks and minimize socioeconomic disruptions caused by unforeseen climate events.

Acknowledgments

We are grateful to the associate editor, Prof. Attilio Castellarin, and two anonymous reviewers for their constructive comments.

Appendix A. Summary of ARMAX

ARMAX is a commonly used multivariate linear regression method (Box et al., 2008). An ARMAX model consists of three groups of terms, namely, autoregressive (AR) terms, moving average (MA) terms, and exogenous inputs. For the predictors considered in this paper, the ARMAX model for predicting streamflow Q_t can be written as

$$Q_t = \underbrace{\sum_{i=1}^{p} a_i Q_{t-i\text{AR}}}_{} + \underbrace{\sum_{k=1}^{K}\sum_{l=1}^{n_k^b} b_{kl} V_{k,t-n_k^d-l}}_{\text{Exogenous inputs}} + \underbrace{\sum_{j=1}^{q} c_j \varepsilon_{t-j\text{MA}}}_{} + \varepsilon_t, \tag{A.1}$$

where p is the order of AR terms; V_k are exogenous inputs ($k = 1,\ldots,K$), and n_k^b and n_k^d are the order and delay of the k-th exogenous input; q is the number of MA terms; and $\varepsilon_t, \varepsilon_{t-1}, \ldots$ are white noise terms. In the case of second predictor group (PD-II), $p = 2$, $K = 6$, $q = 0$, and no delays are applied. The unknown coefficients a_i, b_{kl}, and c_j are estimated as part of the training process.

Appendix B. Summary of MLP

MLP is a type of feedforward ANN that consists of an input layer, one or more hidden layers, and an output layer, with each layer consisting of one of more neurons. Development of an MLP model proceeds by connecting layers in a forward manner using weighted sum of basis functions,

$$\begin{cases} a_j^{(l)} = \sum_{i=0}^{M_{l-1}} \mathbf{w}_{ij}^{(l-1)} y_i^{(l-1)} \\ y_j^{(l)} = \phi(a_j^{(l)}) \end{cases}, \quad (l = 1,2,\ldots,L+1) \tag{A.2}$$

where l is layer number, $w_{ij}^{(l-1)}$ are weights associated with the neurons in the previous layer $l-1$, and the basis function ϕ is better known as the activation function or transfer function in the literature. The weights of an MLP can be obtained through a backpropagation algorithm, in weights are updated in a backward manner from layer to layer to minimize the error function (Haykin, 1994). In this work, a single hidden layer is used in all MLP models.

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jhydrol.2014.01.023.

References

Barlow, M., Nigam, S., Berbery, E.H., 2001. ENSO, Pacific decadal variability, and US summertime precipitation, drought, and stream flow. J. Climate 14 (9), 2105–2128.
Berger, K.P., Entekhabi, D., 2001. Basin hydrologic response relations to distributed physiographic descriptors and climate. J. Hydrol. 247 (3), 169–182.
Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. 249 (1), 11–29.
Bishop, C.M., 2006. Pattern recognition and machine learning. Information Science and Statistics, vol. xx. Springer, New York, p. 38.
Blöschl, G., Sivapalan, M., 1995. Scale issues in hydrological modelling: a review. Hydrol. Process. 9 (3–4), 251–290.
Botter, G., Basso, S., Rodriguez-Iturbe, I., Rinaldo, A., 2013. Resilience of river flow regimes. Proc. Natl. Acad. Sci..

Bourdin, D.R., Fleming, S.W., Stull, R.B., 2012. Streamflow modelling: a primer on applications. Approaches and challenges. Atmos. Ocean 50 (4), 507–536.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 2008. Time series analysis: forecasting and control. In: Wiley Series in Probability and Statistics. vol. xxiv. John Wiley, Hoboken, NJ, p. 46.

Budyko, M.I., 1974. Climate and Life. Academic, San Diego, CA.

Chang, F.-J., Chen, Y.-C., 2001. A counterpropagation fuzzy-neural network modeling approach to real time streamflow prediction. J. Hydrol. 245 (1–4), 153–164.

Chiew, F.H.S., McMahon, T.A., 2002. Global ENSO-streamflow teleconnection, streamflow forecasting and interannual variability. Hydrol. Sci. J. 47 (3), 505–522.

Cressie, N.A.C., 1993. Statistics for spatial data. Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics, vol. xx. Wiley, New York, p. 900.

Demuth, H., Beale, M., Hagan, M., 2008. Neural Network Toolbox™ User's Guide. The MathWorks Inc., Natick, MA, pp. 907.

Duan, Q. et al., 2006. Model Parameter Estimation Experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. J. Hydrol. 320 (1), 3–17.

Dunne, T., 1983. Relation of field studies and modeling in the prediction of storm runoff. J. Hydrol. 65 (1), 25–48.

Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn. 53 (4), 343–367.

Ghosh, S., Mujumdar, P., 2008. Statistical downscaling of GCM simulations to streamflow using relevance vector machine. Adv. Water Resour. 31 (1), 132–146.

Girard, A., Rasmussen, C.E., Quiñonero-Candela, J., Murray-Smith, R., 2003. Gaussian Process priors with uncertain inputs – application to multiple-step ahead time series forecasting. In: Becker, S. et al. (Eds.), Advances in Neural Information Processing System 15. MIT Press, Cambridge, Mass, pp. 529–536.

Gutman, A., Iganatov, A., 1998. The derivation of the green vegetation fraction from NOAA/AVHRR data for use in numerical weather prediction models. Int. J. Remote Sens. 19 (8), 1533–1543.

Haykin, S.S., 1994. Neural Networks: A Comprehensive Foundation, vol. xix. Maxwell Macmillan, Canada, Toronto, Canada, pp. 696.

Herman, J., Reed, P., Wagener, T., 2013. Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. Water Resour Res..

Hidalgo, H.G., Dracup, J.A., 2003. ENSO and PDO effects on hydroclimatic variations of the Upper Colorado River basin. J. Hydrometeorol. 4 (1), 5–23.

Hsieh, W.W., Tang, B., 1998. Applying neural network models to prediction and data analysis in meteorology and oceanography. Br. Am. Meteorol. Soc. 79, 1855–1870.

Hsu, K.L., Gupta, H.V., Sorooshian, S., 1995. Artificial neural-network modeling of the rainfall-runoff process. Water Resour. Res. 31 (10), 2517–2530.

Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resour. Res. 42, W03S04, http://dx.doi.org/10.1029/2005WR004362.

Kitanidis, P.K., 1995. Quasi-linear geostatistical theory for inversing. Water Resour. Res. 31 (10), 2411–2419.

Kotecha, J.H., Djuric, P.M., 2003. Gaussian particle filtering. IEEE Trans. Signal Proc. 51 (10), 2592–2601.

Li, H. et al., 2011. Evaluating runoff simulations from the Community Land Model 4.0 using observations from flux towers and a mountainous watershed. J. Geophys. Res.: Atmos. (1984–2012) 116 (D24).

Li, D., Pan, M., Cong, Z., Zhang, L., Wood, E., 2013. Vegetation control on water and energy balance within the Budyko framework. Water Resour. Res..

Loucks, D.P., Stedinger, J.R., Haith, D.A., 1981. Water Resource Systems Planning and Analysis. Prentice-Hall.

McCabe, G.J., Wolock, D.M., 2002. A step increase in streamflow in the conterminous United States. Geophys. Res. Lett. 29 (24), 2185.

McDonnell, J.J. et al., 2007. Moving beyond heterogeneity and process complexity: a new vision for watershed hydrology. Water Resour. Res. 43, W07301. http://dx.doi.org/10.1029/2006WR005467.

Milly, P.C.D., Dunne, K.A., Vecchia, A.V., 2005. Global pattern of trends in streamflow and water availability in a changing climate. Nature 438 (17), 347–350.

Nasonova, O.N., Gusev, Y.M., Kovalev, Y.E., 2009. Investigating the ability of a land surface model to simulate streamflow with the accuracy of hydrological models: a case study using MOPEX materials. J. Hydrometeorol. 10 (5), 1128–1150.

Olden, J.D., Kennard, M.J., Pusey, B.J., 2012. A framework for hydrologic classification with a review of methodologies and applications in ecohydrology. Ecohydrology 5 (4), 503–518.

Oudin, L., Michel, C., Anctil, F., 2005. Which potential evapotranspiration input for a lumped rainfall-runoff model?: Part 1—can rainfall-runoff models effectively handle detailed potential evapotranspiration inputs? J. Hydrol. 303 (1), 275–289.

Oudin, L., Kay, A., Andréassian, V., Perrin, C., 2010. Are seemingly physically similar catchments truly hydrologically similar? Water Resour. Res. 46 (11).

Patil, S., Stieglitz, M., 2012. Controls on hydrologic similarity: role of nearby gauged catchments for prediction at an ungauged catchment. Hydrol. Earth Syst. Sci. 16 (2), 551–562.

Piechota, T.C., Chiew, F.H.S., Dracup, J.A., McMahon, T.A., 2001. Development of exceedance probability streamflow forecast. J. Hydrol. Eng. 6, 20–28.

Pike, J., 1964. The estimation of annual run-off from meteorological data in a tropical climate. J. Hydrol. 2 (2), 116–123.

PRISM Climate Group, 2012. PRISM, <http://prism.oregonstate.edu>. Last . Oregon State University, Corvallis, OR (accessed 10.08.12).

Quiñonero-Candela, J., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. 6, 1939–1959.

Rasmussen, C.E., 1996. Evaluation of Gaussian Processes and Other Methods for Non-linear Regression. University of Toronto, Toronto, Canada.

Rasmussen, C.E., Nickisch, H., 2010. Gaussian processes for machine learning (GPML) toolbox. J. Mach. Learn. Res. 11, 3011–3015.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning. Adaptive Computation and Machine Learning, vol. xviii. MIT Press, Cambridge, Mass., p. 48.

Reis, D.S., Stedinger, J.R., Martins, E.S., 2005. Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. Water Resour. Res. 41 (10), W10419.

Robertson, D.E., Wang, Q.J., 2012. A Bayesian approach to predictor selection for seasonal streamflow forecasting. J. Hydrometeorol. 13 (1), 155–171.

Sankarasubramanian, A., Vogel, R.M., 2003. Hydroclimatology of the continental United States. Geophys. Res. Lett. 30 (7).

Sankarasubramanian, A., Vogel, R.M., Limbrunner, J.F., 2001. Climate elasticity of streamflow in the United States. Water Resour. Res. 37 (6), 1771–1781.

Schaake, J.C., Duan, Q., Smith, M., Koren, V., 2000. Criteria to select basins for hydrologic model development and testing. In: 15th Conference on Hydrology, January 10–14, 2000, Amer. Meteor. Soc., Long Beach, CA.

Shi, J.Q., Wang, B., Murray-Smith, R., Titterington, D.M., 2007. Gaussian process functional regression modeling for batch data. Biometrics (63), 714–723.

Sivapalan, M., 2003. Prediction in ungauged basins: a grand challenge for theoretical hydrology. Hydrol. Process. 17 (15), 3163–3170.

Sivapalan, M. et al., 2003. IAHS decade on predictions in ungauged basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol. Sci. J. 48 (6), 857–880.

Sivapalan, M., Thompson, S., Harman, C., Basu, N., Kumar, P., 2011. Water cycle dynamics in a changing environment: Improving predictability through synthesis. Water Resour. Res. 47 (10).

Stedinger, J.R., Tasker, G.D., 1985. Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared. Water Resour. Res. 21 (9), 1421–1432.

Sun, A.Y., 2013. Predicting groundwater level changes using GRACE data. Water Resour. Res. 49 (9), 5900–5912.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Tokar, A.S., Markus, M., 2000. Precipitation-runoff modeling using artificial neural networks and conceptual models. J. Hydrol. Eng. 5 (2), 156–161.

Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer Verlag, New York, NY.

Villarini, G., Serinaldi, F., Smith, J.A., Krajewski, W.F., 2009. On the stationarity of annual flood peaks in the continental United States during the 20th century. Water Resour. Res. 45 (8), W08417.

Vogel, R.M., Tsai, Y., Limbrunner, J.F., 1998. The regional persistence and variability of annual streamflow in the United States. Water Resour. Res. 34 (12), 3445–3459.

Vogel, R.M., Wilson, I., Daly, C., 1999. Regional regression models of annual streamflow for the United States. J. Irrig. Drain. Eng. 125 (3), 148–157.

Wagener, T., Sivapalan, M., Troch, P., Woods, R., 2007. Catchment classification and hydrologic similarity. Geography Compass 1 (4), 901–931.

Wagener, T. et al., 2010. The future of hydrology: an evolving science for a changing world. Water Resour. Res. 46 (5), W05301.

Wang, D., Hejazi, M., 2011. Quantifying the relative contribution of the climate and direct human impacts on mean annual streamflow in the contiguous United States. Water Resour. Res. 47 (10), W00J12.

Wang, Q.J., Robertson, D.E., 2011. Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. Water Resour. Res., 47.

Wang, D., Wu, L., 2012. Similarity between runoff coefficient and perennial stream density in the Budyko framework. Hydrol. Earth Syst. Sci. Disc. 9 (6), 7571–7589.

Wang, Q., Robertson, D., Chiew, F., 2009. A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. Water Resour. Res. 45 (5), W05407.

Williams, C.A. et al., 2012. Climate and vegetation controls on the surface water balance: synthesis of evapotranspiration measured across a global network of flux towers. Water Resour. Res. 48 (6), W06523.

Xu, X., Scanlon, B.R., Schilling, K., Sun, A., 2013. Relative importance of climate and land surface changes on hydrologic changes in the US Midwest since the 1930s: implications for biofuel production. J. Hydrol. 497 (8), 110–120.

Zealand, C.M., Burn, D.H., Simonovic, S.P., 1999. Short term streamflow forecasting using artificial neural networks. J. Hydrol. 214 (1), 32–48.

Zhang, L., Dawes, W., Walker, G., 2001. Response of mean annual evapotranspiration to vegetation changes at catchment scale. Water Resour. Res. 37 (3), 701–708.

Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms. CRC Press.