Journal of Hydrology 561 (2018) 764–779

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Reference evapotranspiration forecasting based on local meteorological and global climate information screened by partial mutual information

Wei Fang[a], Shengzhi Huang[a,*], Qiang Huang[a], Guohe Huang[b], Erhao Meng[a], Jinkai Luan[a]

[a] State Key Laboratory of Eco-hydraulics in Northwest Arid Region of China, Xi'an University of Technology, Xi'an 710048, China
[b] Institute for Energy, Environment and Sustainable Communities, University of Regina, Regina, Saskatchewan S4S 0A2, Canada

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Purna Chandra Nayak, Associate Editor

Keywords:
Evapotranspiration
Partial mutual information
Climatic indices
Teleconnection

ABSTRACT

In this study, reference evapotranspiration (ET_0) forecasting models are developed for the least economically developed regions subject to meteorological data scarcity. Firstly, the partial mutual information (PMI) capable of capturing the linear and nonlinear dependence is investigated regarding its utility to identify relevant predictors and exclude those that are redundant through the comparison with partial linear correlation. An efficient input selection technique is crucial for decreasing model data requirements. Then, the interconnection between global climate indices and regional ET_0 is identified. Relevant climatic indices are introduced as additional predictors to comprise information regarding ET_0, which ought to be provided by meteorological data unavailable. The case study in the Jing River and Beiluo River basins, China, reveals that PMI outperforms the partial linear correlation in excluding the redundant information, favouring the yield of smaller predictor sets. The teleconnection analysis identifies the correlation between Nino 1 + 2 and regional ET_0, indicating influences of ENSO events on the evapotranspiration process in the study area. Furthermore, introducing Nino 1 + 2 as predictors helps to yield more accurate ET_0 forecasts. A model performance comparison also shows that nonlinear stochastic models (SVR or RF with input selection through PMI) do not always outperform linear models (MLR with inputs screen by linear correlation). However, the former can offer quite comparable performance depending on smaller predictor sets. Therefore, efforts such as screening model inputs through PMI and incorporating global climatic indices interconnected with ET_0 can benefit the development of ET_0 forecasting models suitable for data-scarce regions.

1. Introduction

Evapotranspiration is a crucial component in the hydrological cycle, simultaneously transferring water from land, oceans and plants to the atmosphere through evaporation and transpiration (Tabari et al., 2013). Estimating the reference evapotranspiration (ET_0) is essential for engineering applications like the irrigation scheduling as well as scientific research like the hydrological modelling. The FAO-56 Penman-Monteith (FAO-PM) equation (Allen et al., 1998) is recommended by the Food and Agriculture Organization (FAO) to be a standard model for estimating ET_0. Benefiting from a solid physical foundation, the FAO-PM equation with related adjustments can be used as a good estimator (Jato-Espino et al., 2016). Its main drawback, however, lies in its relatively high data requirement, which limits its application in many regions, especially in the least economically developed countries, where sufficient meteorological stations and reliable observations are often unavailable (Droogers and Allen, 2002). Therefore, it is of important significance to develop alternative models with lower data burden and computationally suitable for forecasting ET_0 in data-scarce regions.

The aforementioned limitation of the FAO-PM equation has led researchers to turn to numerous empirical models with reduced data requirements. Empirical models mainly include temperature-based (Hargreaves, Blaney-Criddle and Thornthwaite) equations and radiation-based (Priestley-Taylor, Makkink and Jensen-Haise) equations, some of which the FAO-PM equation evolved from. As no universal consensus has been achieved on their global applicability, additional parameter estimation is an indispensable step in applying empirical models to different climatic conditions (Droogers and Allen, 2002; Nandagiri and Kovoor, 2006). The other category of alternative models manages to capture the mapping relationship between selected inputs and ET_0 by means of statistical methods or artificial intelligence approaches covering from multiple linear regression, autoregressive moving average and support vector regression (Jato-Espino et al., 2016; Psilovikos and Elhag, 2013; Tabari et al., 2012; Cheng et al., 2016) to various neural networks and evolutionary algorithms (Falamarzi et al.,

* Corresponding author.
E-mail address: huangshengzhi@xaut.edu.cn (S. Huang).

https://doi.org/10.1016/j.jhydrol.2018.04.038
Received 6 January 2018; Received in revised form 12 April 2018; Accepted 14 April 2018
Available online 19 April 2018

2014; Shiri et al., 2014; Traore et al., 2016; Fang et al., 2017). For all these models, identifying the optimal input is a fundamental task and is a necessity to reduce the model data requirements. The conventional solution is to test several input combinations comprising only a portion of the meteorological variables available and then derive the optimal input set according to predetermined evaluation criteria (Parasuraman et al., 2007; Partal, 2016; Traore et al., 2016). Though a computationally efficient searching strategy, examining a fraction of all possible combinations instead of an exhaustive search still leaves doubt as to whether there are some combinations with lower data requirements outperforming the 'optimal' input set selected. The other strategy for screening model inputs is based on calculating the linear correlation coefficient, which statistically quantifies the linear dependence between each meteorological variable and ET_0 (Jain et al., 2008; Kişi, 2006). Meteorological variables with strong linear correlation with ET_0 are included in the model input set. This strategy, however, is argued to likely select redundant inputs that provide the same amount of information regarding ET_0. Afterward, the partial linear correlation is introduced to further eliminate the redundant information from the input set (Mallikarjuna et al., 2012). On the other hand, evapotranspiration is universally considered a nonlinear process dependent on interacting climatological variables. As a result, the nonlinear dynamics of the evapotranspiration process may not be well captured by only examining the linear correlation.

To this end, entropy and mutual information (MI), two important notions in information theory, are introduced to quantify more general (both linear and nonlinear) dependence. Entropy is known to be a measure of uncertainty for given variables and it is through the notion of entropy that MI is derived (Quilty et al., 2016). MI, also termed transinformation, is defined as the information content of one variable that is also contained by another variable and is formulated as the difference between total entropy of the two random variables and their joint entropy (Ahmadi et al., 2009; Yang et al., 2000). Ahmadi et al. (2009) and Nourani et al. (2015) have applied these two information-content-based criteria (namely, entropy and MI) to input selection for solar radiation estimation and rainfall-runoff modelling, respectively. Evaluating entropy and MI makes it possible for input selection to consider both linear and nonlinear dependence between input candidates and model output. However, as in the case of selecting input through the linear correlation coefficient, there is a disadvantage when using entropy and MI to screen meaningful inputs. This is, an input strongly correlated with the model output might provide redundant information that has been explained by previously selected inputs. To overcome this shortcoming, Sharma (2000) proposed partial mutual information (PMI) for evaluating the additional mutual information attained by adding a potential input to the model input set. In this study, the utility of the partial mutual information to identify relevant predictors for ET_0 is investigated and is compared with that of the partial linear correlation.

The past two decades have witnessed an increasing number of studies on the interconnections between hydrological variables and global climate patterns at multiple timescales. For precipitation, streamflow and groundwater levels, numerous research has identified their delayed response to variability in climatic indices, such as the North Atlantic Oscillation (NAO), Southern Oscillation Index (SOI) and Pacific-North American pattern (PNA) (Cai et al., 2010; Coleman and Budikova, 2013; Tremblay et al., 2011, Huang et al., 2018, Liu et al., 2018). Wang et al. (2006) revealed the strong influence of El Niño–Southern Oscillation (ENSO) events on regional precipitation in the Yellow River Basin, China, which resulted in a 51% decrease in runoff to the sea. Zhang et al. (2007) reported the spatially changing (in-phase or anti-phase) interconnection between ENSO and the annual maximum streamflow from the upper to the lower Yangtze River Basin, China. It was found by Xu et al. (2007) that approximately 20% of 481 gauging stations in China showed a significant correlation between precipitation and SOI, and a more negative correlation than positive was observed. Such interconnections have been exploited by forecast practices involving these hydrological variables successfully (Fan et al., 2015; Schepen et al., 2012; Yang et al., 2017). With respect to ET_0, Meza (2005) found that ET_0 variation in the Maipo River Basin, Chile, was influenced by phases of ENSO, concluding that during the winter and spring, there was up to a 30% difference in ET_0 between the El Niño and La Niña years. Sabziparvar et al. (2011) analysed the ET_0-SOI interconnection at 13 meteorological station sites in Iran. At most of the studied sites, winter and spring ENSO events influenced the ET_0 values of the following summer and autumn. Spatially, more significant impacts of ENSO forcing on ET_0 variability were observed at warm arid sites than at humid sites. Tabari et al. (2014) examined the ET_0-NAO interconnection during winter at 41 Iranian meteorological stations. The results disclosed the negative correlation between winter ET_0 and NOA index, and a negative phase of NAO led to a 3% increase in ET_0 values relative to those during a positive phase. In spite of studies reporting the apparent interconnection between regional ET_0 and global climate patterns, little attention has been paid to incorporating influential climatic indices into ET_0 forecasting practices. Therefore, this study employs global climatic indices as additional potential inputs of forecasting models to analyse their correlation with ET_0 in the study area and investigate their role in yielding a higher forecasting accuracy. The merit lies in that these climatic indices can be easily acquired from related research institutions and do not increase the data collection burden, and they can be universally applied to regions with meteorological data scarcity.

This study aims to (1) investigate the utility of partial mutual information to identify meaningful predictors for ET_0 through a comparison with the partial linear correlation, which merely measures the linear dependence; (2) recognize the interconnection between global climate indices and regional ET_0; and (3) recommend the optimal ET_0 forecasting models having both favourable performance and lower data requirements for regions subject to data scarcity. An appropriate input variable selection (IVS) technique benefits models through effectively decreasing the data requirements. In addition, introducing climatic indices may favour the explanation of variability in ET_0, which ought to be interpreted by the missing meteorological variables. Therefore, the study could have important implications for developing ET_0 forecasting models suitable for the least economically developed countries.

2. Model developments

2.1. An overview of ET_0 forecasting models

The procedure for developing ET_0 forecasting models is organized into four parts, which are depicted in Fig. 1.

2.1.1. Input candidate pools

Scenario 1 is utilized to compare the utility of the partial mutual information and partial linear correlation to screen predictors for ET_0. Under Scenario 1, the input candidate pool comprises all local meteorological variables characterizing variations in air temperature, air pressure, precipitation, humidity, solar radiation and wind speed. It is a prevailing means of composing the input candidate pool and has been used in many previous studies (Chatzithomas and Alexandris, 2015; Kumar et al., 2002; Tabari et al., 2012). Scenario 2 further comprises global climatic indices, in addition to the meteorological variables of Scenario 1, and can provide a comparison with Scenario 1 for investigating the effectiveness of climatic indices in enhancing model performance. Scenario 3 is used for developing ET_0 forecasting models suitable for the least economically developed regions. With consideration of the meteorological data scarcity in many such regions, the input candidate pool under the latter scenario only includes routinely measured meteorological variables (air temperature and sunshine duration), which are available at nearly all meteorological stations. Global climatic indices are further introduced as potential model inputs to

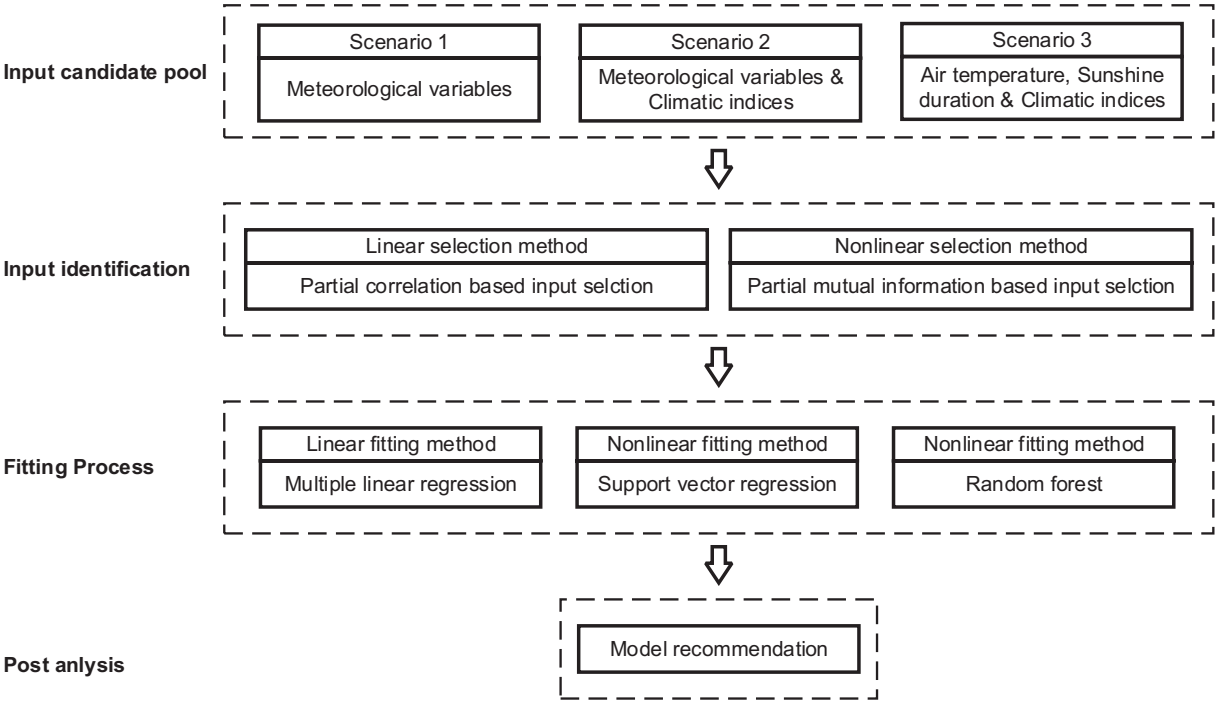

Fig. 1. Schematic description of developing ET_0 forecasting models.

make up the information regarding ET_0, which ought to be provided by meteorological variables unavailable, such as the humidity, wind speed and solar radiation.

2.1.2. Input identification

Variables relevant to ET_0 are identified from the candidate pool by means of partial-correlation-based or partial-mutual-information-based IVS methods, which can measure the linear or nonlinear dependence between random variables. Then, the relevant variables selected constitute the predictor set for the predictand, ET_0.

2.1.3. Fitting process

Predictors of ET_0 serve as inputs of the fitting methods adopted, including multiple linear regression, support vector regression and random forecast. Three methods compete to capture the mapping relationship between predictors and the predictand in linear or nonlinear manners.

2.1.4. Post analysis

Performance of ET_0 forecasting models combining different IVS methods and fitting methods is quantified according to evaluation criteria. Then, the optimal model is recommended that has both favourable forecasting skills and low data requirements to predict ET_0 in data-scarce regions.

2.2. Input variable selection techniques

IVS is the fundamental consideration during the development of accurate and cost-effective statistical models, whose task is to identify the fewest input variables required to interpret the behaviour of model output (May et al., 2011). 'Fewest' implies that both irrelevant and redundant variables need to be excluded from the resulting predictor set. In this study, the partial-correlation and partial-mutual-information-based IVS techniques are adopted.

2.2.1. Partial-correlation-based input selection (PCIS)

The Pearson correlation coefficient, measuring the linear dependence between two random variables, is defined as follows:

$$R_{XY} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (1)$$

where Y and X are a predictand and a potential input variable, respectively; (x_i, y_i) represents the i-th bivariate sample; and n denotes the sample size.

In an extensively used IVS algorithm, Pearson correlation coefficients between the predictand and each potential input variable are calculated and arranged in a descending order. Afterwards, a forward selection is performed, in which potential input variables whose correlation coefficients rank the top k or significantly differ from zero are selected as relevant variables or predictors of the predictand. Obviously, the Pearson-correlation-coefficient-based IVS algorithm follows the screening criterion of maximum relevance; however, without considering the redundancy between selected predictors. Redundancy means that two or more predictors can provide the same amount of information regarding the predictand. As a result, the predictor identification based on Pearson correlation coefficients ignoring the other crucial filtering criterion – the minimum redundancy – tends to choose some redundant predictors from the input candidate pool.

To further exclude redundant variables, this algorithm is modified by replacing the Pearson correlation coefficient with the partial correlation. The partial correlation coefficient quantifies the additional dependence between each input candidate and the predictand Y that cannot be accounted for by the predictors having been selected (Sharma, 2000). In detail, if the first predictor (Z_1) is identified, the partial correlation between Y and each potential input X in the candidate pool is expressed as follows:

$$R_{XY\cdot Z_1} = \frac{R_{XY}-R_{XZ_1}R_{YZ_1}}{\sqrt{(1-R_{XZ_1}^2)(1-R_{YZ_1}^2)}} \quad (2)$$

If the second predictor (Z_2) is subsequently identified, the partial correlation, conditional on two predictors selected (Z_1 and Z_2), is computed as (De La Fuente et al., 2004) follows:

$$R_{XY\cdot Z_1Z_2} = \frac{R_{XY\cdot Z_1} - R_{XZ_2\cdot Z_1}R_{YZ_2\cdot Z_1}}{\sqrt{(1-R^2_{XZ_2\cdot Z_1})(1-R^2_{YZ_2\cdot Z_1})}} \tag{3}$$

Similarly, a higher-order partial correlation coefficient can be calculated as more predictors are progressively screened out.

In the partial-correlation-based input variable selection (PCIS) algorithm, the modified correlation measurement and the forward selection guarantee the effective implementation of the minimum-redundancy – maximum-relevance (mRMR) criterion. The selection process will terminate if the maximum partial correlation regarding the remaining potential input variables no longer significantly differs from zero at the 95% confidence level.

2.2.2. Partial-mutual-information-based input selection (PMIS)

The PCIS algorithm, despite having been used extensively, is criticized for its fundamental assumption of a linearly structured dependence between predictors and predictands within the system to be modelled, as well as its sensitivity to the noise carried by samples (May et al., 2008). As a crucial component in the hydrological cycle, evapotranspiration is a highly complicated and nonlinear process driven by interacting climatological factors, such as precipitation, temperature, and wind speed (Kim and Kim, 2008; Kumar et al., 2002). Therefore, a PCIS algorithm capable of evaluating linear dependence in the system studied may not be suitable for addressing the nonlinear relationships of the evapotranspiration process.

An alternative to the PCIS algorithm is the partial-mutual-information-based input variable selection (PMIS) algorithm without an assumption of the dependence structure. It provides an emerging approach for detecting both linear and nonlinear dependence in a multivariate system, based on mutual information (MI) rooted in the information theory (Fraser and Swinney, 1986). MI between the predictand Y and a potential input variable X is given as follows:

$$I_{XY} = \int\int p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}dxdy \tag{4}$$

where $p_X(x)$ and $p_Y(y)$ signify marginal probability density functions (PDFs) of X and Y, respectively, and $p_{X,Y}(x, y)$ denotes the joint PDF. I_{XY} is often interpreted as the reduction in the uncertainty regarding Y owing to the observation of X. Evidently, no dependence between X and Y will lead to an I_{XY} equal to 0, and a higher I_{XY} value shows a stronger correlation between two random variables.

In a practical context, I_{XY} is estimated using a set of bivariate samples as follows:

$$I_{XY} \approx \frac{1}{n}\sum_{i=1}^{n}\log\frac{f_{X,Y}(x_i,y_i)}{f_X(x_i)f_Y(y_i)} \tag{5}$$

where (x_i, y_i) is the i-th sample point; n represents the sample size; $f_X(x)$, and $f_Y(y)$ and $f_{X,Y}(x, y)$ denote the estimation of the corresponding marginal and joint PDFs, respectively.

Estimating the marginal and joint PDFs can be performed using either a crude histogram or a kernel density estimator (KDE). Here, owing to its favourable accuracy and robustness, the latter coupled with the Gaussian kernel is employed to derive the following estimation of PDF:

$$f_{\mathbf{U}}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}\lambda^d\det(\mathbf{S})^{1/2}}\exp\left(-\frac{(\mathbf{u}-\mathbf{u}_i)^T\mathbf{S}^{-1}(\mathbf{u}-\mathbf{u}_i)}{2\lambda^2}\right) \tag{6}$$

where $f_{\mathbf{U}}(\mathbf{u})$ is the multivariate PDF estimation of d-dimensional variable set $\mathbf{U}$; $\mathbf{S}$ denotes the sample covariance matrix of $\mathbf{U}$; det () symbolizes the determinant operation; and λ represents the kernel bandwidth that largely determines the estimation accuracy of PDF and is empirically recommended by Sharma (2000) as Eq. (7):

$$\lambda = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)} \tag{7}$$

As in the case of the Pearson correlation coefficient, the forward selection scheme choosing the top-k values of MI between the predictand and each potential input variable serves as an effective means of excluding irrelevant variables, but it is likely to include those that are redundant. Hence, there is a need to introduce the partial mutual information (PMI) analogous to the partial relation. It quantifies how much remaining uncertainty regarding Y that has not yet been explained by the selected predictors can be further interpreted by the observation of X. PMI between X and Y, conditional on a predictor set $\mathbf{Z}$ with m members, is formulated as follows:

$$I'_{XY\cdot\mathbf{Z}} = \int p_{X',Y'}(x',y')\log\frac{p_{X',Y'}(x',y')}{p_{X'}(x')p_{Y'}(y')}dx'dy' \tag{8}$$

where X' and Y' represent the residual information contained in variables X and Y, which cannot be accounted for by $\mathbf{Z}$.

Sample-based estimation of PMI is expressed as follows:

$$I'_{XY\cdot\mathbf{Z}} \approx \frac{1}{n}\sum_{i=1}^{n}\log\frac{f_{X',Y'}(x'_i,y'_i)}{f_{X'}(x'_i)f_{Y'}(y'_i)} \tag{9}$$

where $f_{X'}(x')$ $f_{Y'}(y')$ and $f_{X',Y'}(x',y')$ are marginal and joint PDFs that can be estimated by following Eq. (6), respectively, and n is the number of samples.

$$f_{X'}(x') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}\lambda^d\det(\mathbf{S}_{X'X'})^{1/2}}\exp\left(-\frac{(x'-x'_i)\mathbf{S}^{-1}_{X'X'}(x'-x'_i)}{2\lambda^2}\right) \tag{10a}$$

$$f_{Y'}(y') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}\lambda^d\det(\mathbf{S}_{Y'Y'})^{1/2}}\exp\left(-\frac{(y'-y'_i)\mathbf{S}^{-1}_{Y'Y'}(y'-y'_i)}{2\lambda^2}\right) \tag{10b}$$

$$f_{X',Y'}(x',y') = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi)^{d/2}\lambda^d\det(\mathbf{S}_{X'Y'})^{1/2}}\exp\left(-\begin{bmatrix}x'-x'_i\\ y'-y'_i\end{bmatrix}^T\mathbf{S}^{-1}_{X'Y'}\begin{bmatrix}x'-x'_i\\ y'-y'_i\end{bmatrix}/2\lambda^2\right) \tag{10c}$$

$$x'_i = x_i - E[x\,|\mathbf{z}_i], \quad y'_i = y_i - E[y\,|\mathbf{z}_i] \tag{10d}$$

in which $E[]$ denotes the expectation operation. The conditional expectation $E[x\,|\mathbf{z}]$ is the mean of the conditional PDF ($p_{X|\mathbf{Z}}(x\,|\mathbf{z}) = p_{X,\mathbf{Z}}(x,\mathbf{z})/p_{\mathbf{Z}}(\mathbf{z})$); therefore, it can be estimated via the KDE according to Eq. (11) as follows:

$$E[x\,|\mathbf{z}] \approx \frac{1}{n}\sum_{i=1}^{n}\omega_i[x_i + (\mathbf{z}-\mathbf{z}_i)^T\mathbf{S}^{-1}_{\mathbf{ZZ}}\mathbf{S}_{\mathbf{XZ}}] \tag{11a}$$

$$\omega_i = \exp\left[-\frac{(\mathbf{z}-\mathbf{z}_i)^T\mathbf{S}^{-1}_{\mathbf{ZZ}}(\mathbf{z}-\mathbf{z}_i)}{2\lambda^2}\right]\Bigg/\sum_{j=1}^{n}\exp\left[-\frac{(\mathbf{z}-\mathbf{z}_j)^T\mathbf{S}^{-1}_{\mathbf{ZZ}}(\mathbf{z}-\mathbf{z}_j)}{2\lambda^2}\right] \tag{11b}$$

where $\mathbf{S}_{\mathbf{XZ}}$ signifies the cross-covariance of X and $\mathbf{Z}$, and $\mathbf{S}_{\mathbf{ZZ}}$ denotes the covariance of $\mathbf{Z}$. $E[y\,|\mathbf{z}]$ can be obtained in a similar manner.

In this study, the predictor-identification process based on the PIMS algorithm adopts Akaike information criterion (AIC) as the termination criterion and is detailed as follows:

(a) Construct an input candidate pool $\mathbf{C}$ based on prior knowledge regarding the system to be modelled. Initialize a null set $\mathbf{X}$ to place selected predictors. Assume $\mathbf{X}$ and the predictand to be the system input and output, respectively.
(b) To identify the first predictor, calculate the MI between the predictand and each potential input by following Eqs. (5)–(7). Add that with the maximum MI score to the predictor set $\mathbf{X}$ and remove it from the candidate pool $\mathbf{C}$.
(c) Use a general regression neural network (GRNN) to fit the relationship between the model input $\mathbf{X}$ and output. Calculate the AIC of the system and mark it by $AIC_{\mathbf{X}}$.

(d) As for recognizing more predictors, follow Eqs. (9)–(11) to evaluate the PMI between the predictand and each remaining input candidate in **C**.
(e) Identify the input candidate c_j having the maximum PMI score. Assume $\mathbf{X} \cap c_j$ to be the new predictor set. Use GRNN to fit the relationship between the predictand and $\mathbf{X} \cap c_j$. Calculate the corresponding AIC, $AIC_{\mathbf{X}\cap c_j}$. If $AIC_{\mathbf{X}\cap c_j}$ is smaller than $AIC_{\mathbf{X}}$, add c_j to the predictor set **X** and remove it from the candidate pool **C**.
(f) Repeat steps (c)–(e). Once $AIC_{\mathbf{X}\cap c_j}$ becomes larger than $AIC_{\mathbf{X}}$, the input selection is terminated.

Note that the termination criterion, AIC, can provide a tradeoff between the goodness-of-fit and the model complexity and is formulated as follows:

$$\mathrm{AIC} = n\log_e\left(\frac{1}{n}\sum_{i=1}^{n} u_i^2\right) + 2p \tag{12}$$

where n denotes the number of samples, u_i represents the model output residual resulting from the regression of ET_0 on the predictor set **X** through GRNN, and p is the Vapnik–Chernovenkis (VC) dimension characterizing the model complexity.

2.3. Fitting techniques

Three fitting techniques (multiple linear regression, support vector regression and random forest) compete to capture the mapping relationship between ET_0 and the screened predictor set. Such efforts aim at reducing the model uncertainties arising from selecting appropriate fitting techniques.

2.3.1. Multiple linear regression

Assume that the predictor set **X** of the predictand Y is composed of k variables, i.e., X_1, …, X_k. Multiple linear regression (MLR) approximates the input–output relationship of a multivariable system in the following manner (Grégoire, 2014):

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \tag{13}$$

where $\beta = [\beta_1,\ldots,\beta_k]$ are coefficients to be estimated; ε is the noise with the mean equal to zero and unknown variance. Usually, the core task, estimating coefficients $\boldsymbol{\beta}$, is accomplished using the least square method (Chatterjee and Hadi, 1986).

2.3.2. Support vector regression (SVR)

Support vector regression (SVR) first maps the predictor set into a higher dimensional feature space where the predictand can be linearly described and then subtly converts the linear expression in the feature space back to the original predictor space with the help of a Mercer kernel. The linear relationship in the higher feature space can be formulated as follows:

$$f(x) = \langle \mathbf{w}\cdot\varphi(\mathbf{x})\rangle + b \tag{14}$$

where **w** and b are the weight vector and offset coefficient, respectively; $\langle\cdot\rangle$ represents the dot product; and $\varphi()$ is the transformation function used for mapping **x** from the original predictor space into the a higher feature space.

w and b are estimated by minimizing the regression risk expressed as Eq. (15):

$$R_{SVR} = \frac{1}{2}\|w\|^2 + R_{emp} \tag{15}$$

where $\|w\|^2$ measures the flatness of the regression model, which indicates the model complexity; and R_{emp} is the empirical risk.

Usually, the regression risk is minimized by addressing it with a convex optimization. Important details regarding the algorithm can be found in Chang and Lin (2011), Cortes and Vapnik (1995), and Smola and Schölkopf (2004). Finally, the derived linear equation in the higher feature space is converted to a nonlinear form in the original predictor space by the kernel function $k()$.

Commonly used kernel functions include linear, polynomial, sigmoid and radial basis function (RBF) types. In this study, the RBF kernel presented in Eq. (16) is adopted:

$$k(\mathbf{x},\mathbf{x}_i) = \exp(-\gamma\|\mathbf{x}-\mathbf{x}_i\|^2) \tag{16}$$

where γ denotes the kernel width.

The performance of SVR is largely dependent on appropriately selecting the parameter combination of C, γ and ε. Therefore, there is a need to carefully tune these parameters.

2.3.3. Random forest (RF)

In 2001, Breiman (2001) developed the random forest (RF) for the purpose of enhancing the performance of classification and regression trees (CARTs) and reducing the risk of overfitting. Hereafter, the RF has been widely applied to streamflow and electricity price forecasting (Mei et al., 2014; Yang et al., 2017), as well as land cover and gene classification (Díaz-Uriarte and De Andres, 2006; Gislason et al., 2006). In the regression analysis, the RF is an ensemble of regression trees, and the algorithm is detailed as follows (Liaw and Wiener, 2002):

(a) Draw a bootstrap sample from the training set.
(b) Grow an unpruned regression tree to fit the bootstrap sample. At each node, select m variables from all predictors randomly. Pick the best split among the selected m variables to divide the node into two child nodes. The best split is defined as the one capable of minimizing the mean square error. Recursively repeat the split process until the predetermined termination criterion, such as the minimum members in child nodes or the maximum tree size, is met.
(c) Repeat steps (a) and (b) and aggregate B trees as an RF.
(d) Produce the final prediction by averaging the outputs of all the trees.

The performance of RF is sensitive to the selection of the maximum number of variables used to grow a regression tree m and the number of regression trees B in the forest. Therefore, these two parameters need to be carefully tuned.

2.4. Model calibration

As previously stated, there are several parameters largely governing the performance of the fitting methods. The model calibration aims at searching for the optimal combination of these parameters for SVR and RF. The calibration mechanism adopted by this study combines a two-stage calibration strategy and the well-known shuffled complex evolution (SCE-UA) algorithm (Duan et al., 1993, 1992).

Firstly, to prevent the model from overfitting, the calibration period is divided into a calibration subperiod and a test subperiod. SVR or RF with each parameter combination candidate is trained by samples in the calibration subperiod. Note that the calibrated models are evaluated in the test subperiod prior to being directly applied to the validation period. The fitness of each parameter combination is characterized by the smaller value of the Nash-Sutcliffe efficiency (NSE) coefficients during the calibration and test subperiods, which is given by Eq. (17):

$$\mathrm{fitness} = \min(\mathrm{NSE}^{\mathrm{cal}},\mathrm{NSE}^{\mathrm{tes}}) \tag{17}$$

Then, the SCE-UA algorithm is responsible for evolutionally and iteratively generating a number of parameter combination candidates within a predetermined parameter space. Eventually, the parameter combination with the maximum fitness is identified to be the optimal for SVR or RF.

Previous model calibration is accustomed to determining fitness of parameter combinations only based on model performance during the calibration period. However, the two-stage calibration strategy

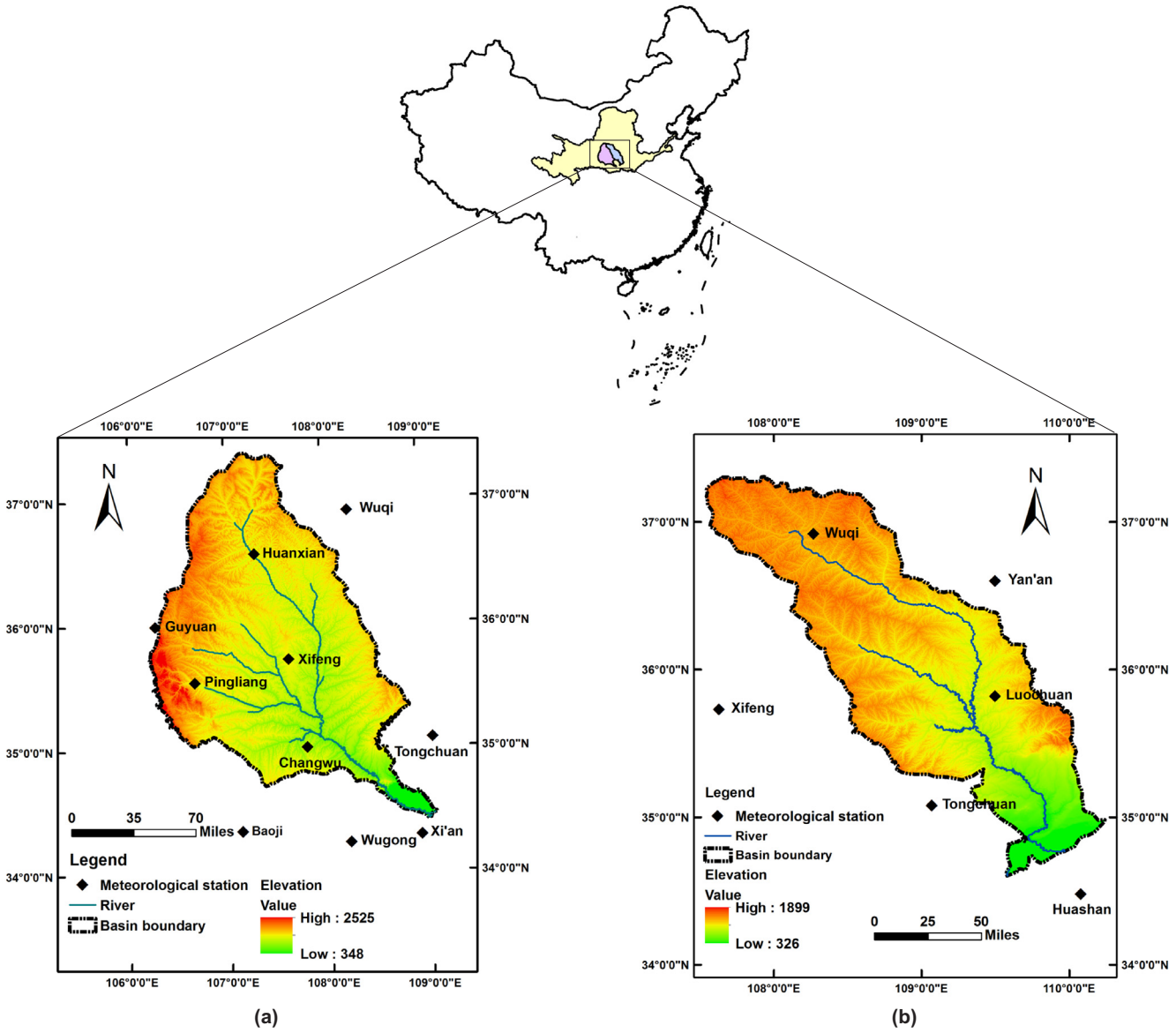

Fig. 2. Location of (a) the Jing River Basin and (b) Beiluo River Basin.

provides a more conservative but safer alternative, which could effectively restrain the overfitting risk and therefore benefit models in maintaining consistent in-sample and out-of-sample performance.

2.5. Evaluation criteria

The goodness-of-fit of ET_0 forecasting models is evaluated by three statistical measures, namely NSE, percent bias (PBIAS) and RMSE-observation standard deviation ratio (RSR).

NSE is computed using Eq. (18) as follows:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n} (E_i^{obs} - E_i^{fore})^2}{\sum_{i=1}^{n} (E_i^{obs} - \overline{E}^{obs})^2} \tag{18}$$

where E_i^{obs} and E_i^{fore} denote the i-th observed and forecasted values of ET_0, respectively; $\overline{E}^{obs}$ represents the mean of the observed ET_0; and n is the total number of observations. NSE varies from negative infinity to 1, with 1 suggesting a perfect match between forecasts and observations, and a value less than 0 indicating that $\overline{E}^{obs}$ outperforms E_i^{fore} in providing the better prediction.

The PBIAS shown in Eq. (19) measures the average tendency of the forecasted ET_0 relative to observations. A positive or negative value of PBIAS indicates the overestimation or underestimation bias. A PBIAS nearer to zero corresponds to a more accurate prediction.

$$\mathrm{PBIAS} = \frac{\sum_{i=1}^{n} (E_i^{obs} - E_i^{fore}) \times 100}{\sum_{i=1}^{n} E_i^{obs}}\% \tag{19}$$

RMSE is a statistic used extensively in evaluating model performance. In RSR, RMSE is standardized by the standard deviation of observations. The merit of applying RSR is that Singh et al. (2005) have proposed a guideline to quantify which RMSE is eligible to be identified at a low level based on a comparison with the standard deviation of observations as follows:

Scenario 1

Observations of 12 meteorological variables at current time step (t)

$\overline{T}_t$ $\overline{T}_{\max,t}$ $\overline{T}_{\min,t}$ $T_{ext+,t}$ $T_{ext-,t}$ SH_t SP_t RH_t U_t P_t VP_t AP_t

Scenario 2	Scenario 3
① Observations of 12 meteorological variables at current time step (t): $\overline{T}_t$ $\overline{T}_{\max,t}$ $\overline{T}_{\min,t}$ $T_{ext+,t}$ $T_{ext-,t}$ SH_t SP_t RH_t U_t P_t VP_t AP_t	① Observations of routinely measured meteorological variables at current time step (t): $\overline{T}_t$ $\overline{T}_{\max,t}$ $\overline{T}_{\min,t}$ $T_{ext+,t}$ $T_{ext-,t}$ SH_t

② Values of 24 climatic indices both at current time step (t) and lagging up to 12 months

AO_t	AO_{t-1}	…	AO_{t-12}	Nino 4_t	Nino 4_{t-1}	…	Nino 4_{t-12}
AMM_t	AMM_{t-1}	…	AMM_{t-12}	NOA_t	NOA_{t-1}	…	NOA_{t-12}
AMO_t	AMO_{t-1}	…	AMO_{t-12}	NTA_t	NTA_{t-1}	…	NTA_{t-12}
CAR_t	CAR_{t-1}	…	CAR_{t-12}	NP_t	NP_{t-1}	…	NP_{t-12}
EA/WR_t	EA/WR_{t-1}	…	EA/WR_{t-12}	ONI_t	ONI_{t-1}	…	ONI_{t-12}
EP/NP_t	EP/NP_{t-1}	…	EP/NP_{t-12}	PDO_t	PDO_{t-1}	…	PDO_{t-12}
GMT_t	GMT_{t-1}	…	GMT_{t-12}	PNA_t	PNA_{t-1}	…	PNA_{t-12}
$GIAM_t$	$GIAM_{t-1}$	…	$GIAM_{t-12}$	QBO_t	QBO_{t-1}	…	QBO_{t-12}
MEI_t	MEI_{t-1}	…	MEI_{t-12}	SOI_t	SOI_{t-1}	…	SOI_{t-12}
Nino $1+2_t$	Nino $1+2_{t-1}$	…	Nino $1+2_{t-12}$	SF_t	SF_{t-1}	…	SF_{t-12}
Nino 3.4_t	Nino 3.4_{t-1}	…	Nino 3.4_{t-12}	TNI_t	TNI_{t-1}	…	TNI_{t-12}
Nino 3_t	Nino 3_{t-1}	…	Nino 3_{t-12}	TPI_t	TPI_{t-1}	…	TPI_{t-12}

Fig. 3. Input candidate pools under three scenarios.

$$\mathrm{RSR} = \frac{\mathrm{RMSE}}{\mathrm{STDEV}^{obs}} = \frac{\sqrt{\sum_{i=1}^{n}(E_i^{obs}-E_i^{fore})^2}}{\sqrt{\sum_{i=1}^{n}(E_i^{obs}-\overline{E}^{obs})^2}} \tag{20}$$

In addition, the hit rate and the fraction of the predictions within a factor of two of the observations (FAC2) are also calculated for the validation period. They are formulated as follows (Tominaga et al., 2015):

$$\text{hit rate} = \frac{1}{n}\sum_{i=1}^{n} I_i \tag{21}$$

$$\text{in which } I_i = \begin{cases} 1, & \text{if } \left|\frac{E_i^{fore}-E_i^{obs}}{E_i^{fore}}\right| \leqslant 0.25 \\ 0, & \text{else} \end{cases}$$

$$\mathrm{FAC2} = \frac{1}{n}\sum_{i=1}^{n} I_i' \tag{22}$$

$$\text{in which } I_i' = \begin{cases} 1, & \text{if } \frac{1}{2} \leqslant \frac{E_i^{fore}}{E_i^{obs}} \leqslant 2 \\ 0, & \text{else} \end{cases}$$

3. Study area and data

Two basins, namely, the Jing River basin (JRB) and the Beiluo River basin (BRB) in China, were taken as a case study.

3.1. Overview of the Jing River and Beiluo River

The Jing River and Beiluo River, which are depicted in Fig. 2, are second-order tributaries of the Yellow River, China (Huang et al., 2014). The Jing River flows 455 km from its headwaters in the Ningxia Hui Autonomous Region, draining an area of 45,400 km^2. The mean annual discharge of the Jing River is 1.832 billion m^3. The Beiluo River, with a main channel length of 680 km, is the longest river in Shaanxi Province. Annually, approximately 0.997 billion m^3 of runoff flows within its drainage area, estimated to be 26,900 km^2. The two rivers mainly travel across the Loess Plateau, which is known as one of China's most ecologically fragile regions having suffered from severe soil erosion and desertification for centuries (Li et al., 2017). Over 350 million tons of highly erodible loess are annually delivered from surface runoff to the Yellow River.

The JRB and BRB have a continental monsoonal climate featuring intensive precipitation and high temperatures during the summer and rare precipitation and low temperatures during the winter (Liu et al., 2017). Annual precipitation and reference evapotranspiration are approximately 550 mm and 1100 mm in the JRB, respectively. They are 520 mm and 1100 mm in the BRB, respectively. The reference evapotranspiration is much higher than precipitation in both basins and may be largely attributed to the sparse vegetative cover as a result of extensive human interventions, such as deforestation, over-cultivation and over-grazing.

3.2. Data collection

Monthly meteorological data gauged at 13 meteorological stations situated within or adjacent to the JRB and BRB are retrieved from the China Meteorological Data Service Center (CMDC) (http://www.cma.gov.cn/2011qxfw/2011qsjgx/). These data covering a period from 1966 to 2010 have observations of 12 variables, including the mean temperature ($\overline{T}$), mean of daily maximum temperature ($\overline{T}_{\max}$), mean of daily minimum temperature ($\overline{T}_{\min}$), extreme high temperature (T_{ext+}), extreme low temperature (T_{ext-}), sunshine duration (*SH*), percentage of sunshine (*SP*), mean relative humidity (*RH*), mean wind speed (*U*), precipitation (*P*), mean vapor pressure (*VP*) and mean air pressure (*AP*). Observations obtained at one station site are not competent to describe meteorological conditions of the whole basin. Therefore, the weight regarding each meteorological station is computed using the Thiessen polygon method, and then area-weighted observations of corresponding meteorological variables were yielded for the two studied basins (the JRB and BRB). As is indicated in Fig. 3, observations of the 12 meteorological variables during the current month constitute the input candidate pool in Scenario 1.

In addition to local meteorological information, the 24 global climatic indices listed in Table 1 are also employed to analyse whether they can contribute to a more accurate prediction of ET_0. Selected climatic indices are mainly associated with oceanic and atmospheric phenomena, such as the El Niño–Southern Oscillation (ENSO) and many teleconnection patterns. Their monthly values from 1965 to 2010 were retrieved from the Earth System Research Laboratory of the National Oceanic and Atmospheric Administration (https://www.esrl.noaa.gov/psd/data/climateindices/list/). As has been noted in Subsection 2.1, both the 12 meteorological variables and the 24 climatic indices are employed under Scenario 2 as potential model inputs. Under Scenario 3, the 24 climatic indices and routinely measured meteorological variables (air temperature and sunshine duration) available at nearly all meteorological stations are both introduced into the input candidate pool. Particularly, potential input variables related to the 24 climatic indices cover their monthly observations at the current time step and those lagging from 1 month to 12 months, due to the consideration of the propagation speed and transport paths of water vapor on a global scale. Input candidate pools under Scenarios 2 and 3 are depicted in Fig. 3.

As for the modelling target, the reference evapotranspiration is calculated using the FAO-PM equation (Allen et al., 1998) as follows:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma\frac{900}{T+273}U_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)} \quad (23)$$

in which Δ is the slope of the vapor pressure curve; R_n represents the net radiation reaching the crop surface; G denotes the soil heat flux density; γ signifies the psychrometric constant; T represents the average air temperature at 2 m height; U_2 represents the wind speed at 2 m; and e_s and e_a represent the saturation and actual vapor pressure, respectively. The preparation of all the data needed for the ET_0 calculation follows the procedure presented in Chapter 3 of the FAO Irrigation and Drainage Paper 56 (Allen et al., 1998).

4. Results analysis and discussion

Under Scenario 1, the utility of PMIS to identify the relevant predictors for ET_0 and eliminate the redundant predictors is investigated through a comparison with PCIS. Under Scenario 2, whether global climatic indices contribute to a more accurate ET_0 forecast is investigated. ET_0 forecasting models suitable for the least economically developed regions subject to data scarcity are eventually recommended under Scenario 3.

4.1. Scenario 1: ET_0 forecasting based on local meteorological information

Monthly data employed for the development of models cover a time period from 1966 to 2010. According to the model calibration mechanism formulated in Subsection 2.4, the dataset is divided into two parts. The first part (1966–2001), accounting for 80% of the whole studied period, is used for model calibration, and the second part (2002–2010), comprising the remaining 20%, is used for validation purposes. Subsequently, the calibration period is further split into a calibration subperiod (1966–1993) and a test subperiod (1994–2001) to effectively prevent models from overfitting. The other key issue in the model development involves initializing the parameter space for the fitting techniques. As for SVR, three parameters to be tuned (namely C, γ and ε) are set to vary from 2^{-10}, 2^{-10} and 2^{-8} to 2^{10}, 2^{10} and 2^{-1}, respectively. With respect to RF, the maximum number of input variables used to grow regression trees is set to equal the total number of employed predictors. On the basis of findings regarding RF presented by Yang et al. (2016), the number of regression trees is set at 100. This dataset division and parameter configuration are adopted by all three scenarios.

As shown in Fig. 3, the input candidate pool in Scenario 1 comprises observations of 12 meteorological variables at the current time step. Table 2 presents the predictor set screened by PCIS from the input candidate pool for the JRB. The predictors are listed in the sequence in which they have been chosen. The first selected predictor is $^{-}T_{max.t}$, which has the largest Pearson correlation coefficient (0.932) among the 12 potential inputs. The corresponding p-value is far less than 0.05 indicating that the positive linear correlation between $^{-}T_{max.t}$ and ET_0 is significant at the 95% confidence level. Then, the linear dependence of ET_0 on the remaining potential inputs that cannot be accounted for by $^{-}T_{max.t}$ is quantified by calculating the partial correlation. RH_t is found to own the maximum absolute value of partial correlation (0.682) among the remaining 11 potential inputs and therefore is chosen to be the second predictor. Unlike $^{-}T_{max.t}$, RH_t is negatively related to ET_0. Subsequently, T_t with a partial correlation value of 0.408 is considered

Table 1
Description of 24 climate indices.

Index name	Description	Index name	Description
AO	First leading mode from the EOF analysis of monthly mean height anomalies	Nino 4	Central Tropical Pacific SST (5N-5S, 160E-150 W)
AMM	Atlantic Meridional Mode	NOA	North Atlantic Oscillation
AMO	Atlantic multidecadal Oscillation	NTA	North Tropical Atlantic SST Index
CAR	Caribbean Sea surface temperature (SST) Index	NP	North Pacific pattern
EA/WR	East Atlantic/ West Russia pattern	ONI	Oceanic Nino Index
EP/NP	East Pacific/North Pacific Oscillation	PDO	Pacific Decadal Oscillation
GMT	Global Mean Lan/Ocean Temperature Index	PNA	Pacific North American Index
GIAM	Globally Integrated Angular Momentum	QBO	Quasi-Biennial Oscillation
MEI	Multivariate ENSO Index	SOI	Southern Oscillation Index
Nino 1 + 2	Extreme Eastern Tropical Pacific SST (0-10S, 90 W-80 W)	SF	Solar Flux
Nino 3.4	East Central Tropical Pacific SST (5N-5S, 170–120 W)	TNI	Indices of El Niño evolution
Nino 3	Eastern Tropical Pacific SST (5N-5S,150 W-90 W)	TPI	Tripole Index for the Interdecadal Pacific Oscillation

Table 2
Input variables selected from input candidate pool under Scenario 1 for the JRB.

IVS	Variable name	Partial correlation	P-value	IVS	Variable name	PMI	AIC
PCIS	$^{-}T_{max.t}$	0.932	3.334×10^{-149}	PMIS	$^{-}T_{max.t}$	1.144	-705
	RH_t	-0.682	4.252×10^{-47}		RH_t	0.435	-972
	$^{-}T_t$	0.408	7.821×10^{-15}		$^{-}T_t$	0.292	-1045
	$^{-}T_{min.t}$	-0.622	4.603×10^{-37}		$^{-}T_{min.t}$	0.147	-1059
	SH_t	0.443	2.001×10^{-17}		SH_t	0.147	-1082
	SP_t	-0.825	1.398×10^{-83}		SP_t	0.316	-1181
	VP_t	0.393	1.243×10^{-13}		$T_{ext+.t}$	0.136	-1205
	$T_{ext+.t}$	0.286	1.342×10^{-7}		/	/	/

to reveal the most amount of additional dependence regarding ET_0 compared with the other nine potential inputs. As a result, $^{-}T_t$ joins the predictor set as the third member, followed by $^{-}T_{min.t}$, SH_t, SP_t, VP_t,$^{-}T_{ext+.t}$ being progressively selected in a similar manner. Potential inputs not included into the predictor set are those with partial correlation values not significantly different form zero at the 95% confidence level. By measuring the linear dependence between meteorological variables and ET_0, PCIS identifies a total of eight predictors for ET_0 in the JBR. Table 2 presents the predictor set screened by PMIS for JRB. Predictors are also ranked in the sequence in which they were selected. In PMIS, the PMI between each of the potential inputs and ET_0 is calculated, conditional on predictors having been selected. $^{-}T_{max.t}$ has the highest PMI score of 1.144 among the 12 potential inputs and is identified as the first predictor. The inclusion of RH_t, whose PMI score of 0.435 exceeds its 10 counterparts, in the predictor set yields a decreasing AIC value from -705 to -972; therefore, it is selected to be the second predictor. A downward trend in AIC is maintained when $^{-}T_t$, having the highest PMI score (0.292) relative to the other nine potential inputs, is included. Similarly, $^{-}T_{min.t}$, SH_t, SP_t, $^{-}T_{ext+.t}$ are progressively identified as predictors until the AIC value starts to increase. By measuring both linear and nonlinear dependence between meteorological variables and ET_0, PMIS chooses seven predictors for ET_0 in the BRB. It is noticeable that PCIS and PMIS select the same first six predictors, indicating that their strong correlation with ET_0 has been captured using the two IVS methods. The only difference is that PMIS excluding VP_t yields a smaller predictor set relative to PCIS.

The utility of the predictor sets obtained by PCIS and PMIS for the JRB is examined by fitting their relationship with ET_0 over the calibration period (1966–2001) and then testing it over the calibration period (2002–2010). Three fitting methods (namely, MLR, SVR and RF) are employed to reduce the model uncertainties arising from the appropriate selection of fitting techniques. Table 3 presents the performance evaluation of ET_0 forecasting models combining diverse IVS techniques and fitting techniques. A predictor set consisting of all 12 potential inputs is denoted by 'M12', and the corresponding models are used as benchmarks for comparison purposes. For all models, the consistent performance during the calibration, test and validation periods suggests that the overfitting risk has been addressed well. In the JRB, it was observed that for a given predictor set, SVR invariably outperforms the other fitting methods in terms of two out of the three evaluation statistics, namely, NSE and RSR, while, MLR yields a better PBIAS than that of SVR and RF. As a result, SVR performs better against its two counterparts in capturing the relationship between predictors and ET_0 in the JRB. With respect to different predictor sets, the M12-SVR model presents the best performance with NSE, PBIS and RSR equal to 0.997, −1.407 and 0.058, respectively. However, quite comparable performance was found in the PCIS-SVR and PMIS-SVR models, signifying that both PCIS and PMIS are effective means to select relevant variables for ET_0 in JRB. With lower data requirements, the PCIS-SVR and PMIS-SVR models provide more economical alternatives compared with the M12-SVR model. In addition, the PMIS-SVR model with a smaller predictor set achieved nearly the same model performance as the PCIS-SVR model, also suggesting that PCIS identifies a redundant predictor, VP_t.

In the BRB, PCIS as shown in Table 4 selects as many as 11 predictors to interpret the variability of ET_0. The first predictor chosen is $^{-}T_{max.t}$ with a Pearson correlation coefficient of 0.926. RH_t is found to have the maximum absolute value of partial correlation (0.697) among the remaining 11 potential inputs and is selected to be the second predictor. Subsequently, $^{-}T_t$, SH_t, SP_t, $^{-}T_{min.t}$, VP_t, U_t,$^{-}T_{ext+.t}$, P_t,$^{-}T_{ext-.t}$, whose partial correlation values are significantly different from zero at the 95% confidence level, are identified. Compared with that of PCIS, PMIS yields a smaller predictor set composed of nine components by measuring both linear and nonlinear dependence. Table 4 shows that$^{-}T_{max.t}$ has the highest PMI score of 1.099 among all 12 potential inputs and is the first to enter the predictor set. Then, RH_t is selected to be the second predictor owing to its highest PMI score of 0.458 compared with the other 10 potential inputs and a decreased AIC value from -667 to -956. A descending trend in AIC value continues to be observed, as $^{-}T_t$, $^{-}T_{min.t}$, U_t, SH_t, SP_t,$^{-}T_{ext+.t}$,$^{-}T_{ext-.t}$ are included in the predictor set. It is noted that all predictors obtained through PMIS are also identified by PCIS. Meanwhile, excluding VP_t and P_t results in PMIS yielding a smaller predictor set than that of PCIS.

The utility of predictor sets obtained using the two IVS techniques for the BRB is examined. As listed in Table 5, the consistent performance during the calibration, test and validation periods indicates that all models have been prevented from overfitting. For a given predictor

Table 3
Performance of ET_0 forecasting models developed for the JRB under Scenario 1.

Model	Input number	Calibration			Test			Validation				
		NSE	PBIAS	RSR	NSE	PBIAS	RSR	NSE	PBIAS	RSR	Hit rate	FAC2
M12-MLR	12	0.993	−0.165	0.084	/	/	/	0.994	**0.636**	0.081	0.972	1.000
M12-SVR		0.999	−0.108	0.029	0.998	−1.194	0.045	**0.997**	−1.407	**0.058**	**1.000**	1.000
M12-RF		0.994	0.036	0.079	0.982	0.460	0.133	0.980	−0.763	0.141	0.981	1.000
PCIS-MLR	8	0.993	−0.165	0.085	/	/	/	0.993	**0.636**	0.082	0.972	1.000
PCIS-SVR		0.997	0.148	0.053	0.997	−1.278	0.056	**0.995**	−2.956	**0.069**	**1.000**	1.000
PCIS-RF		0.989	0.030	0.103	0.964	−2.019	0.189	0.964	−5.070	0.188	0.935	1.000
PMIS-MLR	7	0.991	0.082	0.095	/	/	/	0.991	**−0.315**	0.095	0.972	1.000
PMIS -SVR		0.997	0.141	0.054	0.997	−1.238	0.056	**0.995**	−2.764	**0.070**	**1.000**	1.000
PMIS-RF		0.990	0.183	0.101	0.966	−1.552	0.183	0.965	−5.054	0.188	0.926	1.000

Table 4
Input variables selected from input candidate pool under Scenario 1 for BRB.

IVS	Variable name	Partial correlation	P-value	IVS	Variable name	PMI	AIC
PCIS	$^{-}T_{max.t}$	0.926	1.298×10^{-143}	PMIS	$^{-}T_{max.t}$	1.099	−667
	RH_t	−0.697	5.103×10^{-50}		RH_t	0.458	−956
	$^{-}T_t$	0.467	1.640×10^{-19}		$^{-}T_t$	0.335	−1035
	SH_t	0.584	8.134×10^{-32}		$^{-}T_{min.t}$	0.181	−1051
	SP_t	−0.826	2.716×10^{-84}		U_t	0.163	−1054
	$^{-}T_{min.t}$	−0.337	3.118×10^{-10}		SH_t	0.124	−1063
	VP_t	0.329	8.674×10^{-10}		SP_t	0.307	−1187
	U_t	0.369	4.514×10^{-12}		$T_{ext+.t}$	0.125	−1209
	$^{-}T_{ext+.t}$	0.167	2.434×10^{-3}		$T_{ext-.t}$	0.112	−1225
	P_t	−0.140	1.126×10^{-2}		/	/	/
	$T_{ext-.t}$	0.114	4.013×10^{-2}		/	/	/

set, SVR presents the best NSE and RSR among the three fitting methods. MLR is observed to have better PBIAS than that of both SVR and RF. Therefore, SVR is more capable of capturing the relationship between ET_0 and its predictors in the BRB. With regard to different predictor sets, the M12-SVR model employing all 12 potential inputs yields the best performance. However, fairly comparable performance is presented by the PCIS-SVR and PMIS-SVR models with their equivalent NSE values of 0.995, PBIS values of −2.956 and −2.764, and RSR values of 0.069 and 0.070, suggesting that PCIS and PMIS can serve as effective approaches for identifying variables relevant to ET_0 in the BRB. Decreased data requirements together with a similar performance make the PCIS-SVR and PMIS-SVR models more economical alternatives to the M12-SVR model. In addition, the PMIS-SVR model using a smaller predictor set achieves nearly the same performance as the PCIS-SVR model. It is suggested that PCIS includes two redundant predictors, U_t and P_t.

The results under Scenario 1 reveal that PCIS and PMIS are capable of identifying meaningful predictors for ET_0 in the JRB and BRB from numerous meteorological variables available. Thus, they are in favour of deriving foresting models with lower data requirements. Additionally, PCIS tends to include some redundant predictors. However, PMIS effectively excludes the redundant information by simultaneously measuring linear and nonlinear dependence. It should be noted that though ET_0 can be predicted quite well at the basin scale with numerous meteorological variables alone, models developed under Scenario 1 are not suitable for regions where usually there are not adequate observations of meteorological variables available.

4.2. Scenario 2: ET_0 forecasting based on meteorological information and climatic indices

Compared with Scenario 1, Scenario 2 further includes 24 climatic indices as potential inputs so as to investigate their correlation to ET_0 and whether they can contribute to more accurate forecasts.

In the JRB, PCIS identifies as many as 32 meteorological variables and climatic indices as predictors for ET_0, among which the first nine are listed in Table 6. PMIS yields a much smaller predictor set containing eight members. In the BRB, a total of 28 input variables are selected by PCIS to explain the variability of ET_0, while 11 predictors are selected through PMIS as shown in Table 7. Note that Nino 1 + 2, one of the eastern Pacific SST indices used for characterizing ENSO events, is identified by both PMIS and PCIS to be predictors and ranks higher in the predictor set, implying that ET_0 in the JRB and BRB may be strongly correlated with ENSO events.

Afterwards, the utility of these predictor sets obtained are examined by the model performance as presented in Tables 8 and 9. 'M12C24' denotes a predictor set composed of all potential inputs under Scenario 1 and 24 climatic indices. For the JRB, a comparison between Tables 3 and 8 shows a general improvement in model performance due to the further inclusion of relevant climatic indices in predictor sets. Similar performance enhancement is also observed in Tables 5 and 9 for the BRB.

Results under Scenario 2 are a reminder that climatic indices are likely to carry additional information regarding ET_0, and introducing those that are relevant through the appropriate IVS techniques can favour the yield of more accurate ET_0 forecasts.

4.3. Scenario 3: ET_0 forecasting models recommended for data-scarce regions

Under Scenario 3, ET_0 forecasting models are developed for data-scarce regions. Only routinely measured meteorological variables, such as air temperatures (T_t, $^{-}T_{max.t}$, $^{-}T_{min.t}$, $T_{ext+.t}$ and $T_{ext-.t}$) and sunshine duration (SH_t), are employed. In comparison with other meteorological variables such as solar radiation, vapor pressure, relative humidity and wind speed, measuring air temperature and solar duration requires quite simple instruments, which supports the global availability of these observations. Meanwhile, 24 climatic indices are introduced into the input candidate pool. Potential inputs related to climatic indices cover their monthly values at the current time step and those lagging from

Table 5
Performance of ET_0 forecasting models developed for BRB under Scenario 1.

Model	Input number	Calibration			Test			Validation				
		NSE	PBIAS	RSR	NSE	PBIAS	RSR	NSE	PBIAS	RSR	Hit rate	FAC2
M12-MLR	12	0.992	−0.186	0.088	/	/	/	0.993	**0.719**	0.086	0.954	1.000
M12-SVR		0.997	0.248	0.052	0.996	−2.160	0.063	**0.996**	−1.226	**0.065**	**1.000**	1.000
M12-RF		0.993	0.057	0.082	0.967	1.151	0.180	0.951	3.971	0.220	0.972	1.000
PCIS-MLR	11	0.992	−0.118	0.089	/	/	/	0.992	**0.456**	0.089	0.963	1.000
PCIS-SVR		0.998	0.145	0.050	0.995	−2.177	0.070	**0.994**	−2.572	**0.079**	**1.000**	1.000
PCIS-RF		0.990	0.116	0.099	0.966	−0.956	0.184	0.953	−5.316	0.215	0.954	1.000
PMIS-MLR	9	0.989	0.173	0.103	/	/	/	0.988	**−0.668**	0.107	0.954	1.000
PMIS -SVR		0.997	−0.060	0.056	0.995	−2.392	0.069	**0.994**	−2.734	**0.074**	**1.000**	1.000
PMIS-RF		0.990	0.049	0.101	0.965	−0.742	0.101	0.961	−4.042	0.197	0.954	1.000

Table 6
Input variables selected from input candidate pool under Scenario 2 for the JRB.

IVS	Variable name	Partial correlation	P-value	IVS	Variable name	PMI	AIC
PCIS	$\bar{}T_{max.t}$	0.932	3.334×10^{-149}	PMIS	$\bar{}T_{max.t}$	1.144	−704.6
	Nino 1 + 2 $_{t-7}$	−0.726	4.042×10^{-56}		Nino 1 + 2 $_{t-6}$	0.451	−974.7
	SH_t	0.687	6.073×10^{-48}		SH_t	0.316	−1096
	SP_t	−0.737	2.799×10^{-58}		$\bar{}T_t$	0.3243	−1259
	RH_t	−0.478	2.475×10^{-20}		Nino 1 + 2 $_{t-8}$	0.169	−1252
	EP/NP_{t-3}	0.492	1.268×10^{-21}		$\bar{}T_{min.t}$	0.3438	−1348
	EP/NP_{t-10}	0.522	1.791×10^{-24}		$T_{ext+.t}$	0.1389	−1351
	EP/NP_{t-2}	−0.384	5.591×10^{-13}		RH_t	0.1183	−1393
	$\bar{}T_t$	0.319	3.572×10^{-9}		/	/	/
	⋮	⋮	⋮		/	/	/

Table 7
Input variables selected from input candidate pool under Scenario 2 for BRB.

IVS	Variable name	Partial correlation	P-value	IVS	Variable name	PMI	AIC
PCIS	$\bar{}T_{max.t}$	0.926	1.298×10^{-143}	PMIS	$\bar{}T_{max.t}$	1.099	−667.3
	Nino 1 + 2 $_{t-7}$	−0.752	2.591×10^{-62}		Nino 1 + 2 $_{t-6}$	0.4729	−988.2
	SH_t	0.683	3.872×10^{-47}		SH_t	0.3139	−1106
	SP_t	−0.686	1.262×10^{-47}		$\bar{}T_t$	0.3241	−1225
	RH_t	−0.481	1.291×10^{-20}		Nino 1 + 2 $_{t-3}$	0.1874	−1235
	EP/NP_{t-3}	0.529	2.822×10^{-25}		RH_t	0.1904	−1236
	EP/NP_{t-10}	0.497	5.091×10^{-22}		$\bar{}T_{min.t}$	0.375	−1249
	$\bar{}T_t$	0.366	7.488×10^{-12}		U_t	0.1559	−1252
	EP/NP_{t-6}	−0.355	3.640×10^{-11}		$T_{ext+.t}$	0.124	−1264
	U_t	0.302	2.655×10^{-08}		$T_{ext-.t}$	0.1174	−1284
	⋮	⋮	⋮		SP_t	0.1202	−1292

1 month to 12 months. Therefore, there are in total 318 potential inputs under Scenario 3.

In the JRB, PMIS identifies seven predictors as listed in Table 10 to interpret the variability of ET_0. In detail, the predictor set screened by PMIS comprises all meteorological variables except $T_{ext-.t}$, and Nino 1 + 2, with lag lengths of two and six months. Nino 1 + 2 is utilized to characterize the evolution of ENSO events. As these predictors progressively join in the predictor set, a decreasing trend in AIC values is continuously observed. Compared with PMIS, PCIS selects as many as 32 predictors, among which the first nine predictors are listed in Table 10. The other 23 are climatic indices of different lag months. Similar to PMIS, PCIS is found to choose all meteorological variables except $T_{ext-.t}$. Predictors related to Nino 1 + 2 are ranked at higher positions in the predictor set, indicating their strong correlation with ET_0 in the JRB. However, it is noted that PCIS and PMIS identify Nino 1 + 2 with slightly different lag lengths.

Subsequently, the utility of predictor sets obtained by PCIS and PMIS for the JRB is examined. For comparison purposes, two predictor sets are adopted as benchmarks. One denoted by 'T5' contains five temperature variables, and for the other, symbolized by 'T5S', sunshine duration is further added. The performances of the ET_0 forecasting models with the four predictor sets are presented in Table 11. For T5, T5S and the predictor set acquired by PMIS, SVR is generally found to outperform MLR and RF. MLR offers the best performance compared with that of SVR and RF when fitting the relationship between ET_0 and the predictor set obtained through PCIS. With respect to different predictor sets, a comparison between the T5-SVR and T5S-SVR models exhibits that adding the solar duration contributes to enhanced model performance. More importantly, introducing climatic indices into predictor sets favours yielding a distinct improvement in model performance relative to T5 and T5S, which is intuitively shown by the more compact concentration of forecasts around their targets in Fig. 4. A comparison of two predictor sets containing both meteorological variables and climatic indices shows that the PMIS-SVR model with seven predictors had superior performance compared with that of the PCIS-MLR model with as many as 32 predictors, due to a 0.02 increase in NSE value, a 0.009 decrease in RSR value and a 0.009 increase in hit rate. It is suggested that some redundant predictors are selected by PCIS. Therefore, with lower data requirements and superior performance, the PMIS-SVR model is recommended to forecast ET_0 in the JRB.

In the BRB, PMIS identifies seven predictors as listed in Table 12, including all temperature variables (except $T_{ext-.t}$) and Nino 1 + 2 with

Table 8
Performance of ET_0 forecasting models developed for the JRB under Scenario 2.

Model	Input number	Calibration			Test			Validation				
		NSE	PBIAS	RSR	NSE	PBIAS	RSR	NSE	PBIAS	RSR	Hit rate	FAC2
M12C24-MLR	324	0.995	0.102	0.067	/	/	/	0.994	**0.314**	0.076	0.976	1.000
M12C24-SVR		0.999	0.046	0.016	0.999	−0.718	0.030	**0.998**	1.569	**0.048**	**1.000**	1.000
M12C24-RF		0.995	−0.152	0.072	0.976	2.066	0.155	0.977	5.150	0.152	0.968	1.000
PCIS-MLR	32	0.995	1.141	0.070	/	/	/	0.994	**0.516**	0.071	0.975	1.000
PCIS-SVR		0.997	0.109	0.038	0.997	0.283	0.052	**0.997**	−1.240	**0.060**	**1.000**	1.000
PCIS-RF		0.994	0.196	0.075	0.979	0.590	0.144	0.972	−2.299	0.165	0.955	1.000
PMIS-MLR	8	0.994	0.155	0.093	/	/	/	0.993	**−0.209**	0.094	0.975	1.000
PMIS -SVR		0.998	0.124	0.047	0.998	0.111	0.048	**0.997**	−1.009	**0.057**	**1.000**	1.000
PMIS-RF		0.993	0.033	0.082	0.978	−0.843	0.148	0.976	−5.560	0.153	0.942	1.000

Table 9
Performance of ET_0 forecasting models developed for the BRB under Scenario 2.

Model	Input number	Calibration			Test			Validation				
		NSE	PBIAS	RSR	NSE	PBIAS	RSR	NSE	PBIAS	RSR	Hit rate	FAC2
M12C24-MLR	324	0.995	−0.101	0.073	/	/	/	0.996	**0.648**	0.073	0.965	1.000
M12C24-SVR		0.999	0.107	0.036	0.998	1.081	0.054	**0.998**	−1.721	**0.061**	**1.000**	1.000
M12C24-RF		0.995	−0.084	0.072	0.961	2.519	0.196	0.956	3.731	0.210	0.979	1.000
PCIS-MLR	28	0.996	0.044	0.057	/	/	/	0.995	**0.581**	0.056	0.973	1.000
PCIS-SVR		0.998	0.209	0.049	0.996	−1.605	0.066	**0.996**	−1.168	**0.070**	**1.000**	1.000
PCIS-RF		0.994	0.057	0.075	0.975	−0.024	0.158	0.971	−3.842	0.171	0.965	1.000
PMIS-MLR	11	0.990	0.293	0.098	/	/	/	0.990	**−0.934**	0.102	0.969	1.000
PMIS -SVR		0.998	0.046	0.046	0.994	−2.857	0.074	**0.996**	−3.791	**0.068**	**1.000**	1.000
PMIS-RF		0.994	−0.240	0.080	0.976	−1.288	0.155	0.971	−4.567	0.169	0.967	1.000

Table 10
Input variables selected from input candidate pool under Scenario 3 for the JRB.

IVS	Variable name	Partial correlation	P-value	IVS	Variable name	PMI	AIC
PCIS	$^{-}T_{max.t}$	0.932	3.334×10^{-149}	PMIS	$^{-}T_{max.t}$	1.099	−667.3
	Nino $1+2_{t\text{-}7}$	−0.726	4.042×10^{-56}		Nino $1+2_{t\text{-}6}$	0.473	−988.2
	SH_t	0.687	6.073×10^{-48}		SH_t	0.314	−1106
	Nino $1+2_{t\text{-}4}$	0.528	2.406×10^{-25}		$^{-}T_t$	0.324	−1235
	$^{-}T_t$	0.361	1.157×10^{-11}		Nino $1+2_{t\text{-}3}$	0.187	−1252
	$^{-}T_{min.t}$	−0.467	2.603×10^{-19}		$^{-}T_{min.t}$	0.158	−1273
	Nino $1+2_{t\text{-}11}$	0.370	3.587×10^{-12}		$T_{ext+.t}$	0.101	−1284
	AMO_t	−0.256	1.342×10^{-7}		/	/	/
	$T_{ext+.t}$	0.183	0.001		/	/	/
	⋮	⋮	⋮		/	/	/

Table 11
Performance of ET_0 forecasting models developed for the JRB under Scenario 3.

Model	Input number	Calibration			Test			Validation				
		NSE	PBIAS	RSR	NSE	PBIAS	RSR	NSE	PBIAS	RSR	Hit rate	FAC2
T5-MLR	5	0.957	−0.952	0.207	/	/	/	0.960	3.677	0.200	0.889	0.991
T5-SVR		0.977	0.210	0.153	0.977	−1.872	0.151	**0.978**	**0.331**	**0.146**	**0.981**	1.000
T5-RF		0.976	−0.206	0.156	0.917	−3.405	0.287	0.916	−6.211	0.289	0.907	1.000
T5S-MLR	6	0.967	−0.519	0.182	/	/	/	0.969	2.003	**0.175**	0.898	0.991
T5S-SVR		0.979	0.345	0.144	0.979	−0.962	0.145	**0.983**	**1.092**	0.131	**0.972**	1.000
T5S-RF		0.983	−0.061	0.129	0.956	−1.653	0.208	0.950	−5.811	0.223	0.917	1.000
PCIS-MLR	32	0.991	−0.302	0.095	/	/	/	**0.990**	**1.165**	**0.098**	0.991	1.000
PCIS-SVR		0.995	−0.290	0.068	0.982	4.777	0.133	0.986	4.042	0.119	0.991	1.000
PCIS-RF		0.994	−0.138	0.080	0.975	0.850	0.159	0.976	−2.368	0.155	0.991	1.000
PMIS-MLR	7	0.984	0.101	0.126	/	/	/	0.980	**−0.392**	0.140	0.954	0.991
PMIS -SVR		0.994	−0.176	0.079	0.993	−1.278	0.084	**0.992**	−2.626	**0.089**	**1.000**	1.000
PMIS-RF		0.992	−0.011	0.087	0.977	−0.823	0.150	0.980	−3.157	0.142	0.991	1.000

lag lengths of three and six months. In comparison, PCIS yields a much larger predictor set containing 32 elements. The first nine of these are listed in Table 12 and show that Nino 1 + 2 with the lag lengths of four and seven months are ranked at higher positions in the predictor set, indicating their strong linear correlation with ET_0 in the BRB. It is noticeable that the Nino 1 + 2 selected by PMIS and PCIS are of different time lags.

The utility of predictor sets obtained using the two IVS techniques are further evaluated by model performance as shown in Table 13. For T5, T5S and the predictor set acquired through PMIS, SVR invariably offers superior model performance compared with that of MLR and RF in terms of NSE and RSR. However, the mapping relationship between the predictor set obtained by PCIS and ET_0 in the BRB is more accurately fitted by MLR. With respect to different predictor sets, the comparison between the T5-SVR and T5S-SVR models exhibits an improved model performance due to the employment of solar duration as an additional predictor. More importantly, introducing climatic indices into predictor sets favours achieving a distinct improvement in model performance relative to T5 and T5S, which is intuitively presented in Fig. 5. In detail, compared with the T5S-SVR model, the PMIS-SVR model is observed to have a 0.012 increase in NSE, a 0.037 decrease in RSR and a 0.019 increase in hit rate. The PCIS-MLR model offers a 0.019 increase in NSE; a decrease of 3.167% and 0.066 in the PBIAS and RSR, respectively; and a 0.009 increase in the hit rate.

Under Scenario 3, the similar performance of the PCIS and PMIS models suggests that PCIS tends to include some redundant predictors. Meanwhile, it is found that introducing climatic indices favours yielding a more accurate ET_0 forecast in terms of NSE, RSR and hit rate. Nino 1 + 2, one of the ENSO indices, is selected by both PCIS and PMIS as a predictor and ranks higher in the predictor set, revealing that the evapotranspiration process in the JRB and BRB is strongly influenced by ENSO events.

Compared with the predictors under Scenario 1, those under Scenario 3 are all routinely measured meteorological variables and

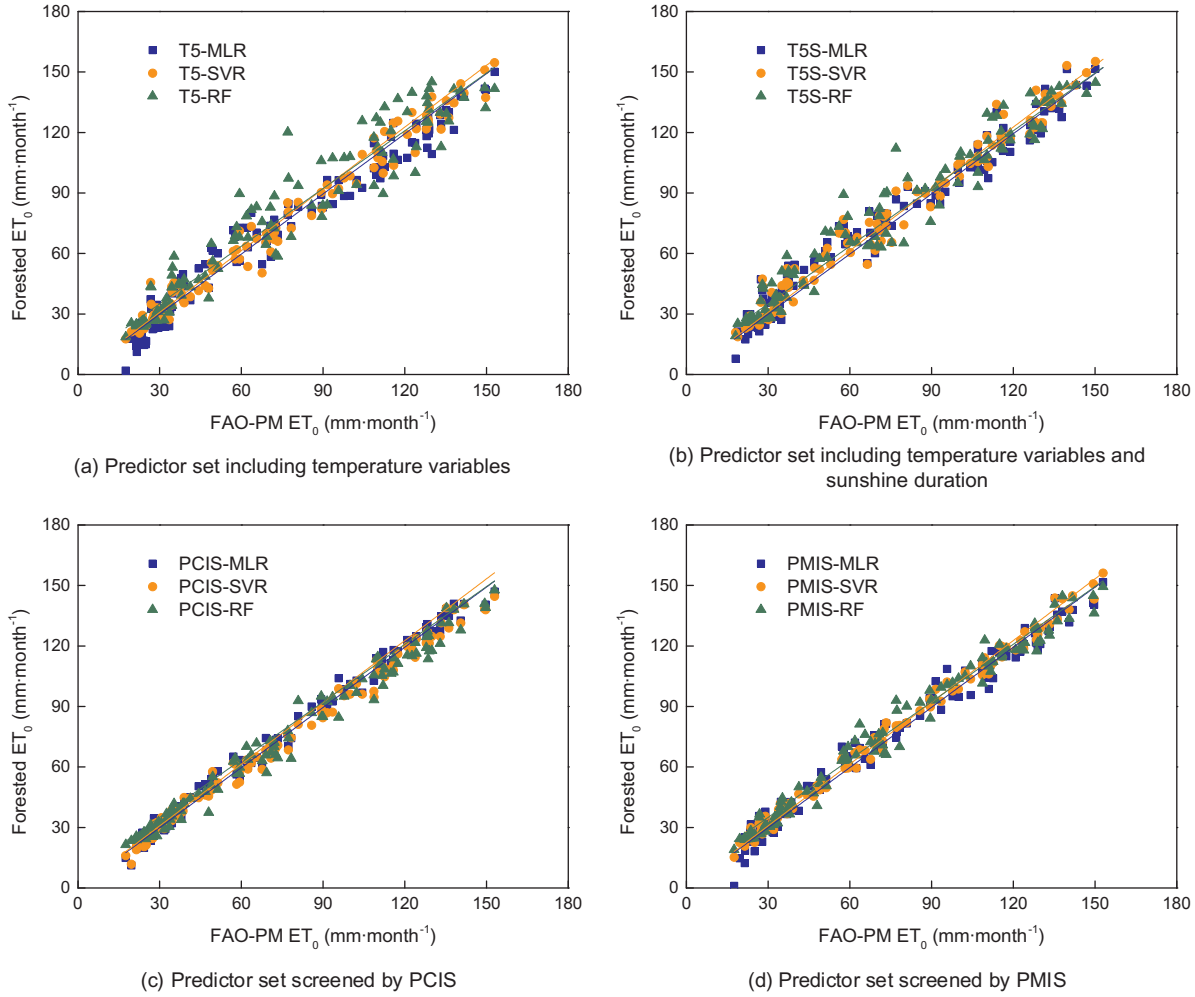

Fig. 4. Scatter plots of ET_0 forecasted under Scenario 2 versus ET_0 calculated by FAO-PM equation during the validation period in the JRB.

Table 12
Input variables selected from input candidate pool under Scenario 3 for the BRB.

IVS	Variable name	Partial correlation	P-value	IVS	Variable name	PMI	AIC
PCIS	$^{-}T_{max.t}$	0.932	3.334×10^{-149}	PMIS	$^{-}T_{max.t}$	1.099	−667.3
	Nino 1 + 2 $_{t-7}$	−0.726	4.042×10^{-56}		Nino 1 + 2 $_{t-6}$	0.473	−988.2
	SH_t	0.687	6.073×10^{-48}		SH_t	0.314	−1106
	Nino 1 + 2 $_{t-4}$	0.528	2.406×10^{-25}		$^{-}T_t$	0.324	−1235
	$^{-}T_t$	0.361	1.157×10^{-11}		Nino 1 + 2 $_{t-3}$	0.187	−1252
	$^{-}T_{min.t}$	−0.467	2.603×10^{-19}		$^{-}T_{min.t}$	0.158	−1273
	Nino 1 + 2 $_{t-11}$	0.370	3.587×10^{-12}		$T_{ext+.t}$	0.101	−1284
	AMO_t	−0.256	1.342×10^{-7}		/	/	/
	$T_{ext+.t}$	0.183	0.001		/	/	/
	⋮	⋮	⋮		/	/	/

climatic indices, making the models developed under Scenario 3 more suitable for forecasting ET_0 in the least economically developed regions. As is shown in Tables 11 and 13, the PMIS-SVR models with seven predictors offer favourable forecasting skills, and are therefore recommended for forecasting ET_0 in the JRB and BRB.

5. Conclusions

This study aimed to develop ET_0 forecasting models for the least economically developed regions subject to meteorological data scarcity, mainly through (1) exploring appropriate input variable selection techniques to effectively reduce model data requirements and (2) introducing global climatic indices as additional model inputs for creating information regarding ET_0, which ought to be provided by meteorological variables unavailable. First, it was investigated whether PMIS was capable of identifying relevant predictors and excluding those that were redundant. A comparison was also made with PCIS. Then, the interconnection between global climatic indices and regional ET_0 was recognized by PMIS and PCIS, and relevant climatic indices were incorporated into the models as additional predictors. Finally, models with both lower data requirements and favourable performance were recommended.

ET_0 forecasting models were developed for two study areas, namely, the JRB and BRB in China. The results indicated that PMIS and PCIS were both effective approaches for identifying the relevant predictors for regional ET_0. However, PCIS only measuring the linear dependence

Table 13
Performance of ET_0 forecasting models developed for the BRB under Scenario 3.

Model	Input number	Calibration			Test			Validation				
		NSE	PBIAS	RSR	NSE	PBIAS	RSR	NSE	PBIAS	RSR	Hit rate	FAC2
T5-MLR	5	0.954	0.189	0.215	/	/	/	0.950	**−0.730**	0.222	0.852	0.991
T5-SVR		0.974	0.509	0.162	0.974	−2.187	0.162	**0.968**	−3.383	**0.177**	**0.944**	1.000
T5-RF		0.969	−0.094	0.176	0.905	0.450	0.306	0.905	−4.161	0.307	0.843	1.000
T5S-MLR	6	0.966	0.966	0.184	/	/	/	0.963	**−3.737**	0.191	0.898	0.991
T5S-SVR		0.979	0.964	0.145	0.977	−4.394	0.152	**0.970**	−4.727	**0.171**	**0.972**	1.000
T5S-RF		0.981	−0.106	0.138	0.952	−2.594	0.219	0.936	−7.306	0.251	0.889	1.000
PCIS-MLR	32	0.992	−0.239	0.091	/	/	/	**0.989**	**0.926**	**0.105**	0.981	1.000
PCIS-SVR		0.995	0.052	0.068	0.991	2.499	0.095	0.988	1.559	0.111	0.944	1.000
PCIS-RF		0.994	−0.031	0.078	0.976	0.017	0.153	0.973	−3.903	0.163	**0.991**	1.000
PMIS-MLR	7	0.984	1.130	0.128	/	/	/	0.978	**−4.372**	0.148	0.954	0.991
PMIS -SVR		0.993	0.042	0.083	0.991	−2.197	0.991	**0.982**	−5.694	**0.134**	**0.991**	1.000
PMIS-RF		0.992	−0.026	0.092	0.979	−1.215	0.146	0.973	−5.146	0.164	**0.991**	1.000

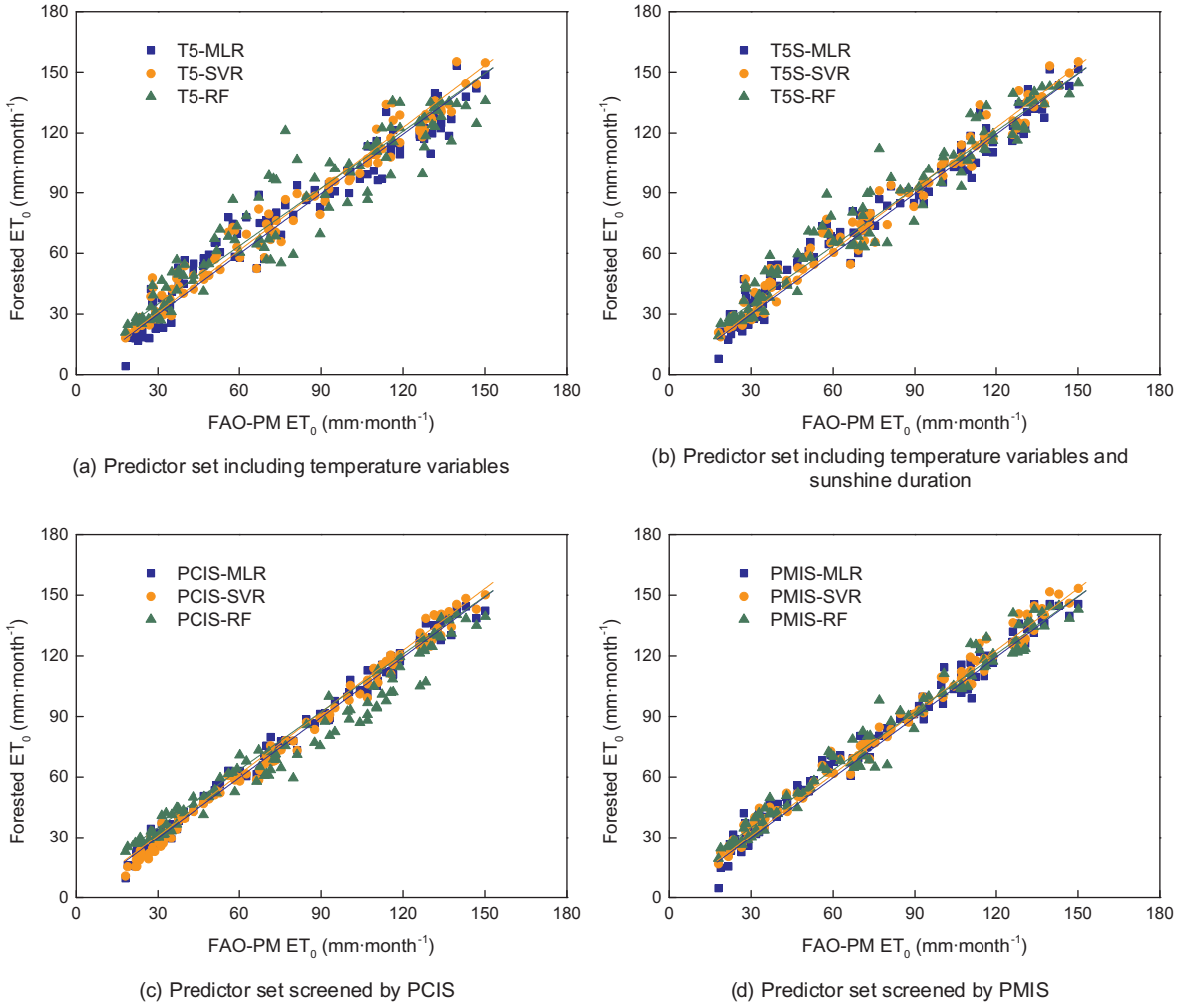

Fig. 5. Scatter plots of ET_0 forecasted under Scenario 2 versus ET_0 calculated by FAO-PM equation during the validation period in the BRB.

tended to select redundant predictors. PMIS presented better performance in excluding redundant information by simultaneously evaluating the linear and nonlinear correlations. Therefore, smaller predictor sets were yielded by PMIS relative to PCIS, which is crucial for model development in data-scarce regions. Furthermore, Nino 1 + 2 characterizing ENSO evolutions was identified by both PMIS and PCIS to be correlated with ET_0, revealing ENSO influences on the evapotranspiration process in the study areas. Introducing Nino 1 + 2 into the models yielded more accurate ET_0 forecasts. Among the various models investigated, the non-linear stochastic models (SVR or RF with inputs selected through PMIS) did not always improve the accuracy of the linear models (MLR with inputs screened by PCIS). However, the PMIS-SVR model was able to offer quite comparable performance depending on smaller predictor sets, and was, therefore, recommended to predict ET_0 in the JRB and BRB. These findings suggest that selecting model inputs through PMIS as well as introducing global climatic

indices into input candidate pools favours developing ET_0 forecasting models suitable for the least economically developed regions.

Although PMIS is proven to be a competitive alternative capable of identifying meaningful predictors for ET_0, previous studies on solar radiation estimation (Ahmadi et al., 2009; Remesan et al., 2008; Reyhani et al., 2005) have shown the superiority of the Gamma test (GT) over MI in selecting the best predictor set. Therefore, future work could be carried out to investigate the applicability of GT in ET_0 estimation and a further comparison between GT and PMIS could be made.

Acknowledgments

This study was joint funded by the National Natural Science Foundation of China (grant number 51709221), the Planning project of science and technology of water resources of Shaanxi (grant number 2017slkj-19), the National Department Public Benefit Research Foundation of Ministry of Water Resources (grant number 201501058), China Scholarship Council (grant number 201608610170), the doctorate innovation funding of Xi'an University of Technology (grant number 310-252071712) and the project of School of Water Resources and Hydropower of Xi'an University of Technology (grant number 2016ZZKT-15). Interested readers can access the data used in this study by contacting the first author.

References

Ahmadi, A., Han, D., Karamouz, M., Remesan, R., 2009. Input data selection for solar radiation estimation. Hydrol. Process. 23 (19), 2754–2764.
Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. FAO, Rome, 300 (9) D05109.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Cai, W., Van Rensch, P., Cowan, T., Sullivan, A., 2010. Asymmetry in ENSO teleconnection with regional rainfall, its multidecadal variability, and impact. J. Clim. 23 (18), 4944–4955.
Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) 2 (3), 27.
Chatterjee, S., Hadi, A.S., 1986. Influential observations, high leverage points, and outliers in linear regression. Statist. Sci. 379–393.
Chatzithomas, C., Alexandris, S., 2015. Solar radiation and relative humidity based, empirical method, to estimate hourly reference evapotranspiration. Agric. Water Manage. 152, 188–197.
Cheng, G., Dong, C., Huang, G., Baetz, B.W., Han, J., 2016. Discrete principal-monotonicity inference for hydro-system analysis under irregular nonlinearities, data uncertainties, and multivariate dependencies Part I: methodology development. Hydrol. Process. 30 (23), 4255–4272.
Coleman, J.S., Budikova, D., 2013. Eastern US summer streamflow during extreme phases of the North Atlantic Oscillation. J. Geophys. Res. Atmos. 118 (10), 4181–4193.
Cortes, C., Vapnik, V., 1995. Support vector machine. Mach. Learn. 20 (3), 273–297.
Díaz-Uriarte, R., De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7 (1), 3.
De La Fuente, A., Bing, N., Hoeschele, I., Mendes, P., 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20 (18), 3565–3574.
Droogers, P., Allen, R.G., 2002. Estimating reference evapotranspiration under inaccurate data conditions. Irrigat. Drain. Syst. 16 (1), 33–45.
Duan, Q., Gupta, V.K., Sorooshian, S., 1993. Shuffled complex evolution approach for effective and efficient global minimization. J. Optimiz. Theor. Appl. 76 (3), 501–521.
Duan, Q., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28 (4), 1015–1031.
Falamarzi, Y., Palizdan, N., Huang, Y.F., Lee, T.S., 2014. Estimating evapotranspiration from temperature and wind speed data using artificial and wavelet neural networks (WNNs). Agr. Water Manage. 140, 26–36.
Fan, Y.R., Huang, W., Huang, G.H., Li, Z., Li, Y.P., Wang, X.Q., Cheng, G.H., Jin, L., 2015. A stepwise-cluster forecasting approach for monthly streamflows based on climate teleconnections. Stoch. Environ. Res. Risk Assess. 29 (6), 1557–1569.
Fang, W., Huang, Q., Huang, S., Yang, J., Meng, E., Li, Y., 2017. Optimal sizing of utility-scale photovoltaic power generation complementarily operating with hydropower: a case study of the world's largest hydro-photovoltaic plant. Energy Convers. Manage. 136, 161–172.
Fraser, A.M., Swinney, H.L., 1986. Independent coordinates for strange attractors from mutual information. Phys. Rev. A 33 (2), 1134.
Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Patt. Recogn. Lett. 27 (4), 294–300.
Grégoire, G., 2014. Multiple linear regression. Eur. Astron. Soc. Public. Ser. 66, 45–72.
Huang, S., Chang, J., Huang, Q., Chen, Y., 2014. Spatio-temporal changes and frequency analysis of drought in the Wei River Basin, China. Water Resour. Manage. 28 (10), 3095–3110.
Huang, S., Ma, L., Chang, J., et al., 2018. Drought structure change characteristic across China based on an integrated drought index. J. Hydrol in press.
Jain, S., Nayak, P., Sudheer, K., 2008. Models for estimating evapotranspiration using artificial neural networks, and their physical interpretation. Hydrol. Process. 22 (13), 2225–2234.
Jato-Espino, D., Charlesworth, S.M., Perales-Momparler, S., Andrés-Doménech, I., 2016. Prediction of evapotranspiration in a mediterranean region using basic meteorological variables. J. Hydrol. Eng. 22 (4), 04016064.
Kim, S., Kim, H.S., 2008. Neural networks and genetic algorithm approach for nonlinear evaporation and evapotranspiration modeling. J. Hydrol. 351 (3), 299–317.
Kişi, Ö., 2006. Evapotranspiration estimation using feed-forward neural networks. Hydrol. Res. 37 (3), 247–260.
Kumar, M., Raghuwanshi, N., Singh, R., Wallender, W., Pruitt, W., 2002. Estimating evapotranspiration using artificial neural network. J. Irrig. Drain. Eng. 128 (4), 224–233.
Li, P., Mu, X., Holden, J., Wu, Y., Irvine, B., Wang, F., Gao, P., Zhao, G., Sun, W., 2017. Comparison of soil erosion models used to study the Chinese Loess Plateau. Earth Sci. Rev. 170, 17–30.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2 (3), 18–22.
Liu, S., Huang, S., Huang, Q., Xie, Y., Leng, G., Luan, J., Song, X., Wei, X., Li, X., 2017. Identification of the non-stationarity of extreme precipitation events and correlations with large-scale ocean-atmospheric circulation patterns: a case study in the Wei River Basin, China. J. Hydrol. 548, 184–195.
Liu, S., Huang, S., Xie, Y., Huang, Q., Leng, G., Hou, B., Zhang, Y., Wei, X., 2018. Spatial-temporal changes of maximum and minimum temperatures in the Wei River Basin, China: Changing patterns, causes and implications. Atmos. Res. 204, 1–11.
P. Mallikarjuna S. Jyothy K.S. Reddy 2012. Daily reference evapotranspiration estimation using linear regression and ANN models J. Institut. Eng. (India): Ser. A 93 4 215 221
May, R., Dandy, G., Maier, H., 2011. Review of input variable selection methods for artificial neural networks, Artificial neural networks-methodological advances and biomedical applications. InTech.
May, R.J., Maier, H.R., Dandy, G.C., Fernando, T.G., 2008. Non-linear variable selection for artificial neural networks using partial mutual information. Environ. Modell. Software 23 (10), 1312–1326.
Mei, J., He, D., Harley, R., Habetler, T., Qu, G., 2014. A random forest method for real-time price forecasting in new york electricity market, PES General Meeting| Conference & Exposition, 2014 IEEE. IEEE 1–5.
Meza, F.J., 2005. Variability of reference evapotranspiration and water demands. Association to ENSO in the Maipo river basin, Chile. Global Planet. Change 47 (2), 212–220.
Nandagiri, L., Kovoor, G.M., 2006. Performance evaluation of reference evapotranspiration equations across a range of Indian climates. J. Irrigat. Drain. Eng. 132 (3), 238–249.
Nourani, V., Khanghah, T., Baghanam, A., 2015. Application of entropy concept for input selection of wavelet-ANN based rainfall-runoff modeling. J. Environ. Inform. 26 (1).
Parasuraman, K., Elshorbagy, A., Carey, S.K., 2007. Modelling the dynamics of the evapotranspiration process using genetic programming. Hydrol. Sci. J. 52 (3), 563–578.
Partal, T., 2016. Comparison of wavelet based hybrid models for daily evapotranspiration estimation using meteorological data. KSCE J. Civ. Eng. 20 (5), 2050–2058.
Psilovikos, A., Elhag, M., 2013. Forecasting of remotely sensed daily evapotranspiration data over Nile Delta region. Egypt Water Resour. Manage. 27 (12), 4115–4130.
Quilty, J., Adamowski, J., Khalil, B., Rathinasamy, M., 2016. Bootstrap rank-ordered conditional mutual information (broCMI): a nonlinear input variable selection method for water resources modeling. Water Resour. Res. 52 (3), 2299–2326.
Remesan, R., Shamim, M., Han, D., 2008. Model data selection using gamma test for daily solar radiation estimation. Hydrol. Process. 22 (21), 4301–4309.
Reyhani, N., Hao, J., Ji, Y., Lendasse, A., 2005. Mutual information and gamma test for input selection.
Sabziparvar, A., Mirmasoudi, S., Tabari, H., Nazemosadat, M., Maryanaji, Z., 2011. ENSO teleconnection impacts on reference evapotranspiration variability in some warm climates of Iran. Int. J. Climatol. 31 (11), 1710–1723.
Schepen, A., Wang, Q., Robertson, D., 2012. Evidence for using lagged climate indices to forecast Australian seasonal rainfall. J. Clim. 25 (4), 1230–1246.
Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: part 1—A strategy for system predictor identification. J. Hydrol. 239 (1), 232–239.
Shiri, J., Sadraddini, A.A., Nazemi, A.H., Kisi, O., Landeras, G., Fard, A.F., Marti, P., 2014. Generalizability of Gene Expression Programming-based approaches for estimating daily reference evapotranspiration in coastal stations of Iran. J. Hydrol. 508, 1–11.
Singh, J., Knapp, H.V., Arnold, J., Demissie, M., 2005. Hydrological modeling of the Iroquois River watershed using HSPF and SWAT. JAWRA 41 (2), 343–360.
Smola, A.J., Schölkopf, B., 2004. A tutorial on support vector regression. Statist. Comput. 14 (3), 199–222.
Tabari, H., Grismer, M.E., Trajkovic, S., 2013. Comparative analysis of 31 reference evapotranspiration methods under humid conditions. Irrig. Sci. 31 (2), 107–117.
Tabari, H., Kisi, O., Ezani, A., Talaee, P.H., 2012. SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment. J. Hydrol. 444, 78–89.
Tabari, H., Talaee, P.H., Some'e, B.S., Willems, P., 2014. Possible influences of North Atlantic Oscillation on winter reference evapotranspiration in Iran. Global Planet. Change 117, 28–39.
Tominaga, Y., Akabayashi, S.-I., Kitahara, T., Arinami, Y., 2015. Air flow around isolated gable-roof buildings with different roof pitches: Wind tunnel experiments and CFD simulations. Build. Environ. 84, 204–213.
Traore, S., Luo, Y., Fipps, G., 2016. Deployment of artificial neural network for short-term

forecasting of evapotranspiration using public weather forecast restricted messages. Agric. Water Manage. 163, 363–379.
Tremblay, L., Larocque, M., Anctil, F., Rivard, C., 2011. Teleconnections and interannual variability in Canadian groundwater levels. J. Hydrol. 410 (3), 178–188.
Wang, H., Yang, Z., Saito, Y., Liu, J.P., Sun, X., 2006. Interannual and seasonal variation of the Huanghe (Yellow River) water discharge over the past 50 years: connections to impacts from ENSO events and dams. Global Planet. Change 50 (3), 212–225.
Xu, Z., Li, J., Takeuchi, K., Ishidaira, H., 2007. Long-term trend of precipitation in China and its association with the El Niño–southern oscillation. Hydrol. Process. 21 (1), 61–71.
Yang, H.H., Van Vuuren, S., Sharma, S., Hermansky, H., 2000. Relevance of time–frequency features for phonetic and speaker-channel classification. Speech Commun. 31 (1), 35–50.
Yang, T., Asanjan, A.A., Welles, E., Gao, X., Sorooshian, S., Liu, X., 2017. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resour. Res. 53 (4), 2786–2812.
Yang, T., Gao, X., Sorooshian, S., Li, X., 2016. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. Water Resour. Res. 52 (3), 1626–1651.
Zhang, Q., Xu, C., Jiang, T., Wu, Y., 2007. Possible influence of ENSO on annual maximum streamflow of the Yangtze River, China. J. Hydrol. 333 (2), 265–274.