Application of ensemble empirical mode decomposition based on machine learning methodologies in forecasting monthly pan evaporation

Mohammad Rezaie-Balf, Ozgur Kisi and Lloyd H. C. Chua

ABSTRACT

Accurate prediction of pan evaporation (PE) is one of the crucial factors in water resources management and planning in agriculture. In this research, two hybrid models, self-adaptive time-frequency methodology, ensemble empirical mode decomposition (EEMD) coupled with support vector machine (EEMD-SVM) and EEMD model tree (EEMD-MT), were employed to forecast monthly PE. The EEMD-SVM and EEMD-MT were compared with single SVM and MT models in forecasting monthly PE, measured between 1975 and 2008, at Siirt and Diyarbakir stations in Turkey. The results were evaluated using four assessment criteria, Nash–Sutcliffe Efficiency (NSE), root mean square error (RMSE), performance index (PI), Willmott’s index (WI), and Legates–McCabe’s index (LMI). The EEMD-MT model respectively improved the accuracy of MT by 36 and 44.7% with respect to NSE and WI in the testing stage for the Siirt station. For the Diyarbakir station, the improvements in results were less spectacular, with improvements in NSE (1.7%) and WI (2.2%), respectively, in the testing stage. The overall results indicate that the proposed pre-processing technique is very promising for complex time series forecasting and further studies incorporating this technique are recommended.

Key words | ensemble EMD, forecasting model tree, pan evaporation, support vector machines

INTRODUCTION

Evaporation is probably the most complicated and difficult parameter to forecast among all the elements of the hydrological cycle due to the complex interactions among the hydrologic components of water surface, land, atmosphere process, and vegetation. Hence, proper forecasting of evaporation is a significant issue in water resources management and agriculture, particularly in semi-arid and arid areas. Because of temperature differences, this hydrological phenomenon is a nonlinear process which occurs in nature (Irmak et al. 2002; Trajkovic & Kolakovic 2010; Tezel & Buyukyildiz 2016; Malik et al. 2017). There are many parameters influencing pan evaporation (PE) rates, including air temperature ($T_A$), solar radiation ($S_R$), sunshine hours ($H_S$), relative humidity ($R_H$) and wind speed ($W_S$). Although accurate estimation of $P_E$ is of high importance, especially in studies related to integrated water resources management, the effects of various weather parameters on pan evaporation in different regions are still not well understood.

In general, there are two main approaches, namely direct and indirect procedures, for calculating and forecasting evaporation. Direct measurements of $P_E$ are widely applied for predicting evaporation. Indirect methods such as energy-budget, water budgets, and Penman–Monteith techniques are based on various climate variables (e.g. $T_A$, $W_S$ and $S_R$), which are not easily accessible. It is not possible nowadays to extract a mathematical relationship that includes all of the physical processes due to the nonlinearity and complexity of the evaporation process. Several researchers therefore recommend the use of models (Tezel & Buyukyildiz 2016; Wang et al. 2017; Tan et al. 2018) that do
not require a priori knowledge of the underlying physical processes or excessive data. Such models are considered to be more adequate (adaptable) in real-decision support systems compared to the physical-based methods (Kisi 2007; Ch et al. 2014; Deo & Şahin 2015). Considering that evaporation is highly nonlinear, these techniques are able to map the nonlinearities associated with evaporation and its input parameters to a high degree of accuracy.

In the last decade, machine learning (ML) methods such as Adaptive Neuro Fuzzy Inference System (ANFIS), artificial neural network (ANN), model tree (MT), gene expression programming (GEP), extreme learning machine (ELM) and support vector machine (SVM) have been successfully employed for solving a wide range of environmental and water engineering problems (Adamowski & Karapataki 2010; Chua et al. 2011; Sattar 2013; Ch et al. 2014; He et al. 2014; Ebtehaj et al. 2015; Deo et al. 2016; Najafzadeh et al. 2016; Li et al. 2017; Mouatadid & Adamowski 2017; Rezaie-Balf & Kisi 2017; Rezaie-Balf et al. 2017a, 2017b; Yu et al. 2017). Many studies have also been conducted using ML methods for predicting pan evaporation (Tabari et al. 2012; Lin et al. 2013; Kim et al. 2015; Tezel & Buyukildiz 2016; Kisi et al. 2016; Malik et al. 2017; Wang et al. 2017), although the development of powerful approaches and methods with high levels of reliability to achieve accurate forecasts remains a major challenge. This is because time series PE data are generally highly nonlinear and seasonal. Under such situations, applying the raw data to the model directly may not produce accurate results if the data are highly nonlinear. In this context, using a preprocessing technique may enhance the model’s performance (Wu et al. 2010).

In recent years, researchers using data pre-processing tools such as principal component analysis (PCA), continuous wavelet transform (CWT), moving average (MA), wavelet multi-resolution analysis (WMRA), maximum entropy spectral analysis (MESA), and singular spectrum analysis (SSA) have been successful in improving forecasting accuracy (Hu et al. 2007; Kisi 2009; Sang et al. 2009; Wu et al. 2009; Wu & Chau 2011; Nourani et al. 2015; Sang et al. 2015; Rezaie-Balf et al. 2017a, 2017b; Tiwari & Adamowski 2017; Wen et al. 2017).

More recently, new noise assisted data analysis techniques, empirical mode decomposition (EMD) and ensemble empirical mode decomposition (EEMD), were pioneered by Huang et al. (1998) and Wu & Huang (2009), respectively. The EEMD approach is an advanced form of traditional decomposition techniques (e.g. wavelet decomposition and Fourier decomposition), and is an intuitive, empirical and self-adaptive data processing tool which was developed for non-linear and non-stationary time series. Recently, several successful applications have been presented which utilized EEMD to forecast time series phenomena such as rainfall and runoff (Karthikeyan & Kumar 2013; Latifoglu et al. 2015; Wang et al. 2015; Zhao & Chen 2015; Barge & Sharif 2016), water quality parameters (Fijani et al. 2019), soil moisture (Prasad et al. 2018), and electricity demand (Al-Musaylh et al. 2018). For instance, Napolitano et al. (2011) carried out a study to explore several aspects of ANN in forecasting daily stream flow time series using EMD. They showed that using EMD analysis, the ANN can have a higher prediction accuracy and reliability. Likewise, Sang et al. (2012) presented a hybrid EMD and maximum entropy spectral analysis method. They proved that this method, in addition to improving the period identification, also can improve overall period identification by being able to discriminate period, trend, and noise. Wang et al. (2013) applied EEMD as an adaptive data analysis tool for decomposing annual rainfall time series in modeling rainfall-runoff process using a SVM model. They employed particle swarm optimization (PSO) to calculate optimal SVM control parameters. Similar to previous investigations, they showed their proposed model’s efficiency and capability in rainfall-runoff forecasting. Moreover, a four-stage model, EEMD-based radial basis function neural network (RBFN) and linear neural network (LNN), was developed by Di et al. (2014) for hydrological time series forecasting. They considered six datasets and showed that their model outperformed the RBFNN, EMD-RBFNN, and EMD-WA-RBFNN-LNN methods.

The overall goal of the current study is to propose a new hybrid model to improve the monthly $P_E$ forecasting. The SVM and MT techniques were investigated and improvements of the models’ accuracy were tested by coupling them with the EEMD procedure. The data used are the monthly $P_E$ data recorded at Siirt and Diyarbakir stations, located in Turkey. The rest of the present research is organized as follows. The next section provides a case study and data analysis, followed by a section describing the
methodology of EEMD and each forecasting method (i.e. MT and SVM). Next, the statistical indicators applied in this study are provided. The penultimate section represents the forecasting results of the hybrid models and the comparison results, and the final section consists of a summary and conclusion of outline results.

STUDY AREA AND DATASET

To validate the performance of the proposed approach, the monthly climatic data and \( P_E \) from Siirt station (latitude 37° 55’ N, longitude 41° 56’ E, and altitude 896 m) and Diyarbakir station (latitude 37° 54’ N, longitude 40° 13’ E, and altitude 649 m), operated by the Turkish Meteorological Organization (TMO) were utilized. All the climatic data were obtained from TMO. It is the only legal organization that provides meteorological information in Turkey. TMO measures pan evaporation, which are not publicly available, using Standard A pans (World Meteorological Organization standards). The 187 datasets of climatic variables comprising wind speed (\( U_2 \)), temperature (\( T_\alpha \)), relative humidity (\( R_h \)) and solar radiation (\( S_R \)) of the aforementioned stations consisting of 35 years (1975–2008) are described in Table 1. Furthermore, data were not available for January, February, March and December and about 5% of the whole data were missing for May and April. Data were directly used in this study without any pre-processing (e.g. filling in missing ones). Seventy-five per cent of the whole data was used for training of the applied models and the obtained models were tested by using the remaining 25% data.

The location of the study area is illustrated in Figure 1. In this area, average annual precipitation is between 400 and 800 mm, falling mainly in the winter months of December to January. The mean values of the pan evaporation during spring and summer (main crop growing seasons) at Siirt and Diyarbakir are 8.81 and 7.47 mm, respectively. Summers are hot and dry and the annual temperature ranges from 16.0 to 46.8 °C (Keshtegar & Kisi 2007). According to the de Martonne (1926) aridity index, both Siirt and Diyarbakir stations have a dry, semiarid (semi-desert) climate.

| Descriptive statistics of the meteorological data at Siirt and Diyarbakir stations |
|----------------------------------|---|---|---|---|---|
| Statistical parameters          | T (°C) | \( R_S \) (MJ/m²) | \( R_h \) (%) | \( U_2 \) (m/s) | \( P_E \) (mm/day) |
| Siirt station (training data set)| Minimum | 10.1 | 216.16 | 17.9 | 0.7 | 1.4 |
|                                | Maximum | 33.5 | 859.12 | 72.3 | 2.8 | 15.7 |
|                                | Mean    | 23.34 | 585.25 | 39.99 | 1.67 | 7.89 |
|                                | \( S_x \) | 6.058 | 152.61 | 12.36 | 0.45 | 3.34 |
|                                | \( C_{sx} \) | -0.37 | -0.58 | 0.48 | 0.13 | 0.008 |
|                                | \( K_x \) | -0.94 | -0.405 | -0.52 | -0.79 | -0.83 |
| Siirt station (testing data set)| Minimum | 10.4 | 321.14 | 25.6 | 0.6 | 2 |
|                                | Maximum | 32.3 | 688.38 | 69.2 | 1.4 | 16.6 |
|                                | Mean    | 24.72 | 547.22 | 40.04 | 1.11 | 9.55 |
|                                | \( S_x \) | 5.69 | 116.48 | 11.62 | 0.19 | 3.68 |
|                                | \( C_{sx} \) | -0.48 | -0.57 | 0.77 | -0.76 | -0.06 |
|                                | \( K_x \) | -0.65 | -0.91 | -0.21 | 0.19 | -0.81 |
| Diyarbakir station (training data set)| Minimum | 8 | 163.41 | 13.4 | 0.2 | 1.2 |
|                                | Maximum | 33.4 | 917.65 | 81.2 | 4.9 | 20.2 |
|                                | Mean    | 23.21 | 643.05 | 42.27 | 2.51 | 8.38 |
|                                | \( S_x \) | 6.472 | 176.21 | 15.01 | 1.01 | 4.04 |
|                                | \( C_{sx} \) | -0.37 | -0.66 | 0.48 | -0.51 | 0.12 |
|                                | \( K_x \) | -0.96 | -0.37 | -0.88 | -0.24 | -0.81 |
| Diyarbakir station (testing data set)| Minimum | 10.3 | 415.35 | 10.7 | 1.1 | 3.6 |
|                                | Maximum | 32.6 | 890.11 | 79.3 | 4 | 16.8 |
|                                | Mean    | 24.52 | 694.96 | 33.61 | 2.43 | 10.03 |
|                                | \( S_x \) | 6.33 | 144.52 | 17.93 | 0.69 | 3.76 |
|                                | \( C_{sx} \) | -0.32 | -0.39 | 0.93 | 0.21 | -0.14 |
|                                | \( K_x \) | -1.07 | -0.97 | 0.33 | -0.35 | -1.21 |
|                                | 10.3 | 415.35 | 10.7 | 1.1 | 3.6 |

\( X_{min}, X_{max}, X_{mean}, S_x, C_{sx}, \) and \( K_x \) denote the minimum, maximum, mean, standard deviation, skewness, and kurtosis, respectively.

MODELS USED

Two ML algorithms, SVM and MT, were developed and coupled with EEMD as a data pre-processing procedure for forecasting monthly \( P_E \) in Siirt and Diyarbakir. In this section, the description of EEMD, SVM and MT are briefly presented.
Ensemble empirical mode decomposition (EEMD)

The EEMD method, proposed by Wu & Huang (2009), is an empirical procedure used to represent a nonlinear and non-stationary signal from original data. This data pre-processing method is an improvement of the EMD, self-adaptive time-frequency transformation procedure, and does not rely on process information. EMD is used mainly for decomposing the original time series data into a finite and low number of oscillatory modes depending on the local characteristic time scale (Huang et al. 1998). The oscillatory modes are revealed by intrinsic mode function (IMFs) components, embedded in the data. These component signals are represented as a sum of zero-mean well behaved fast and slow oscillation modes (IMFs) which have two requirements (Huang et al. 1998; Wu & Huang 2009): (i) Throughout the whole length of data, the number of zero-crossings and the number of extrema must either be equal or at least differ by one; and (ii) At any point, the mean value of the local maxima and minima envelope is zero. According to these properties, some meaningful IMFs can be well defined. Generally, an IMF indicates a simple oscillatory mode, compared to simple harmonic function. Based on this definition, a shifting process of original time series can be briefly expressed as follows (Huang et al. 1998).

Step 1: Identify all extrema (local maxima and minima) points of the given time series $y(t)$;

Step 2: Connect by spline interpolation, to create an upper envelope of local maxima points $e_{\text{max}}(t)$, and a lower envelope $e_{\text{min}}(t)$ of all minima points;

Step 3: Compute the mean $m(t)$ of two envelopes by:

$$m(t) = \frac{(e_{\text{max}}(t) + e_{\text{min}}(t))}{2}$$

(1)
Step 4: Subtract the mean from the data to establish an IMF candidate; \( h(t) = y(t) - m(t) \).

Step 5: If \( h(t) \) satisfies the two properties of IMF as indicated by a predefined stopping criterion, \( h(t) \) is indicated as the first IMF (written as \( c_1(t) \) where 1 is its index); If \( h(t) \) is not an IMF, \( y(t) \) is replaced with \( h(t) \) and iterate steps 1–4 until \( h(t) \) satisfies the two conditions of IMFs.

Step 6: The residue \( r_1(t) = y(t) - c_1(t) \) is then treated as new data subjected to the same shifting procedure as defined above for the next IMF from \( r_1(t) \). In the final step, the shifting process can be stopped, when the residue \( r(t) \) has a monotonous trend or has one local maxima and minima point from which no more IMFs can be extracted (Huang et al. 2003). At the end of this shifting process, the original signal \( y(t) \) can be reconstructed as the sum of IMFs and residual as:

\[
y(t) = \sum_{i=1}^{n} c_i(t) + r_n(t)
\]  

where \( r_n(t) \) represents the final residue, \( n \) is the number of IMFs, and \( c_i(t) \) are almost orthogonal to each other, and all have zero means. Further information for the EMD technique and the stopping criteria can be obtained from Huang et al. (1998, 2003). It has been shown that EMD can be unstable due to mode mixing (Wu & Huang 2009). Mode mixing can occur when a single IMF comprises widely disparate scale components, or a similar scale component residing in different IMFs (Lei et al. 2009). To overcome the scale separation issue without presenting a subjective intermittence test, a new noise-assisted data analysis technique is proposed, the ensemble EMD (EEMD), which characterizes the true IMF components as the average of an ensemble of trials, each comprising the signal in addition to a white noise of finite amplitude. In this regard, the impacts of the decomposition using the EEMD are that the added white noise series cancel each other in the final mean of the corresponding IMFs; the mean IMFs stays within the natural dyadic filter windows and thus significantly reduces the possibility of mode mixing and preserves the dyadic property.

Support vector machine (SVM)

Support vector machines, firstly presented by Cortes & Vapnik (1995), is an effective tool for solving nonlinear problems by relating input variables to a defined output. The least square SVM is an advanced type of SVM, developed by Suykens & Vandewalle (1999), and is used for chaotic time series prediction. The SVM on the basis of the structural risk minimization (SRM) principle, minimizes the expected error in order to reduce the overfitting. In SVM, in order to separate the data patterns, the input data are mapped into a higher dimensional feature space.

Consider a training data set \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), where \( x_i \) is the input value including \( m \) features, \( y_i \) is output value relevant to \( x_i \) and \( n \) is the length of the data set. The SVM regression function is given as:

\[
y = \langle w, x \rangle + b
\]  

where \( w \) and \( b \) are the weights vector in the feature space with the dimension of \( x \) and bias term, respectively, and \( \langle *, * \rangle \) indicates the inner product. The regression issue can be defined as a process for minimization of the regularized risk function as:

\[
\frac{1}{2} W^T W + C \sum_{i=1}^{n} \xi_i + C \sum_{i=1}^{n} \xi_i^2
\]  

subject to the condition:

\[
\begin{align*}
(w, x) r + h - y_i & \leq \varepsilon + \xi_i^r \\
y_i - (w, x) - h & \leq \varepsilon + \xi_i^l \\
\xi_i^r, \xi_i^l & \geq 0 \quad i = 1, 2, \ldots, n
\end{align*}
\]  

where \( C \) is a positive constant that determines the degree of penalized loss for prediction error, \( \xi_i^r \) and \( \xi_i^l \) are the slack variables to specify the distance from observed values to the corresponding boundary values of \( \varepsilon \). It is expected that the most of the data set fall within the range of \( \varepsilon \) and errors \( \xi_i^r \) and \( \xi_i^l \) occur when the data fall outside this range. This optimization is solved by maximizing algorithm of quadratic programming that is illustrated in Equation (6):

\[
\sum_{i=1}^{n} y_i (a_i - \alpha_i) - \varepsilon \sum_{i=1}^{n} (a_i + \alpha_i) - 0.5 \sum_{i,j=1}^{n} (a_i - \alpha_i) (a_j - \alpha_j) K(x_i, x_j)
\]
subject to the condition:

\[
\begin{cases}
\sum_{i=1}^{n} (\alpha_i - \alpha_{\bar{i}}) = 0 \\
0 \leq \alpha_i, \alpha_{\bar{i}} \leq C \\
i = 1, 2, \ldots, N
\end{cases}
\] (7)

where \(\alpha_i\) and \(\alpha_{\bar{i}}\) are the Lagrange multipliers, and \(K(.)\) is the kernel function defined as an inner product of points \(\varnothing(x_i)\) and \(\varnothing(x_{\bar{i}})\), defined as:

\[K(x_i, x_{\bar{i}}) = \varnothing(x_i) \cdot \varnothing(x_{\bar{i}})\] (8)

The solution to the objective function in Equation (6) is optimal and unique and the solution can be expressed by Lagrange multipliers as:

\[y = \sum_{i=1}^{n} (\alpha_i - \alpha_{\bar{i}})K(x, x_i) + h\] (9)

Various kernel functions have been used in SVM to determine the most efficient model for a given type of dataset (Gu et al. 2010; Kisi & Parmar 2016).

---

**Figure 2** | Splitting the input space and prediction by the model tree for a new data record: (a) splitting of the input space \((X1 \times X2)\) by the M5p model tree algorithm; (b) predicting a new data record by the model tree.
Model tree (MT)

The MT is a state-of-the-art hierarchical algorithm, presented by Quinlan (1992), for providing the relation between input-output parameters. This algorithm is used to solve the problem by dividing the data space into several sub-problems (sub-spaces) and builds piecewise linear regression models for each sub-domain. The data records, in classification trees, are classified by sorting the tree from root node to some leaf nodes of the tree. The MT is based on the well-known CART algorithm (Breiman et al. 1984), which deals with continuous-class learning attributes. This algorithm is known to be one of the most efficient methods to present physically meaningful insights for a given phenomenon (Kisi et al. 2016). Figure 2 illustrates the tree-building process, within four linear regression models and the knowledge extracted from the structure for corresponding sub-domains. Furthermore, a general tree structure of MT approach is illustrated in Figure 2(b). In this figure, if $X_1$ and $X_2$ are less than 3 and 1.5, respectively, the fifth model should be considered as a form of $Y = a_0 + a_1X_1 + a_2X_2$.

Where $X_1$ and $X_2$ are inputs, $Y$ is output, $a_0$, $a_1$ and $a_2$ are regression coefficients. Within the MT algorithm, the tree stores a linear model which predicts the class values.

![Figure 3](https://iwaponline.com/hr/article-pdf/50/2/498/549273/nnh0500498.pdf) | Workflow of the models construction procedure.
of the portion of data set reaching to the leaf. According to certain attributes of the data, the records are split into different portions. To determine the best attribute for splitting the data set at each node, standard deviation is utilized as a splitting criterion. Trees are constructed using the standard deviation reduction (SDR) which maximizes the expected

![Figure 4](https://iwaponline.com/hr/article-pdf/50/2/498/549273/eho5000498.pdf)

**Figure 4** | Decomposition of monthly $P_T$ time series in Siirt station.
error reduction for each node as follows (Quinlan 1992):

$$SDR = sd(E) - \sum_{i} \frac{|E_i|}{|E|} sd(E_i)$$ (10)

where $E$ is the set of instances that reaches the leaf(node); $E_i$ represents a subset of input data to the parent node; and $sd$ is the standard deviation. To overcome the overfitting problem and gain better generalization, pruning processes are performed to prune back the overgrown trees in the next stage. In the pruning process, the sub-trees (inner nodes) are transformed into leaf nodes by replacing them with linear regression functions. After pruning, a smoothing procedure is applied to the disjointed linear models of the neighboring leaves. In the smoothing procedure, all the leaf models are amalgamated along the path back to the root to compose the final model in order to provide final predictions. Detailed information on MT can be found in Wang & Witten (1996) and Talebi et al. (2017).

### STATISTICAL PERFORMANCE INDICATORS

In the present research, to assess the EEMD-SVM, EEMD-MT, SVM and MT models, several error indicators were used:

1. Nash–Sutcliffe Efficiency (NSE): This criterion is taken into account to evaluate the ability of hydrological models. The highest value of unity demonstrates a perfect fit between observed and predicted $PE$ where a negative value means that the model performs worse than the arithmetic mean of the time series (Nash & Sutcliffe 1970):

$$NSE = 1 - \frac{\sum_{i=1}^{N} (PE_{for} - PE_{obs})^2}{\sum_{i=1}^{N} (PE_{obs} - \bar{PE}_{obs})^2}$$ (11)

2. Root mean square error (RMSE): RMSE shows the standard deviation of the forecasting errors or residuals. RMSE varies from zero for perfect estimates to large positive values for poor estimates (Hyndman & Koehler 2006):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (PE_{for} - PE_{obs})^2}$$ (12)

3. Performance index (PI): PI values vary from 0 to $+\infty$, with lower values showing better accuracy. A PI value close to zero (e.g. PI $\leq 0.2$) indicates that the model predicts the observed values well (Gandomi & Roke 2015):

$$PI = \frac{RMSE/PE_{obs}}{1 + \frac{\sum_{i=1}^{N} (PE_{obs} - PE_{for}) (PE_{for} - \bar{PE}_{for})}{\sqrt{\sum_{i=1}^{N} (PE_{obs} - PE_{for})^2 \sum_{i=1}^{N} (PE_{for} - \bar{PE}_{for})^2}}}$$ (13)

4. Willmott’s index of agreement (WI): WI is a standardized measure for model prediction error and ranges from 0 to 1. The values close to 0 show poor accuracy while the values close to 1 reveal good performance (Willmott et al. 2012):

$$WI = 1 - \frac{\sum_{i=1}^{N} (PE_{obs} - PE_{for})^2}{\sum_{i=1}^{N} (PE_{for} - \bar{PE}_{for})^2 + |PE_{obs} - \bar{PE}_{obs}|^2}$$ (14)

Legates–McCabe’s index (LMI): This criterion considers absolute values for computation (Legates & McCabe 1999). Therefore, LMI is not inflated by the squared values and is insensitive to outliers making it simple and easy to interpret:

$$LMI = 1 - \frac{\sum_{i=1}^{N} |PE_{for} - PE_{for}|}{\sum_{i=1}^{N} |PE_{obs} - PE_{obs}|}$$

where $PE_{obs}$ and $PE_{for}$ are the observed and forecasting values, respectively; $\bar{PE}_{obs}$ and $\bar{PE}_{for}$ represent the mean of

<table>
<thead>
<tr>
<th>Models</th>
<th>NSE</th>
<th>RMSE (mm/day)</th>
<th>PI (mm/day)</th>
<th>WI</th>
<th>LMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.842</td>
<td>1.304</td>
<td>0.126</td>
<td>0.942</td>
<td>0.672</td>
</tr>
<tr>
<td>MT</td>
<td>0.901</td>
<td>1.047</td>
<td>0.110</td>
<td>0.962</td>
<td>0.724</td>
</tr>
<tr>
<td>EEMD-SVM</td>
<td>0.929</td>
<td>0.887</td>
<td>0.093</td>
<td>0.978</td>
<td>0.774</td>
</tr>
<tr>
<td>EEMD-MT</td>
<td>0.947</td>
<td>0.765</td>
<td>0.079</td>
<td>0.986</td>
<td>0.792</td>
</tr>
<tr>
<td>SVM</td>
<td>0.463</td>
<td>2.663</td>
<td>0.137</td>
<td>0.814</td>
<td>0.341</td>
</tr>
<tr>
<td>MT</td>
<td>0.653</td>
<td>2.140</td>
<td>0.101</td>
<td>0.893</td>
<td>0.492</td>
</tr>
<tr>
<td>EEMD-SVM</td>
<td>0.796</td>
<td>1.641</td>
<td>0.069</td>
<td>0.934</td>
<td>0.591</td>
</tr>
<tr>
<td>EEMD-MT</td>
<td>0.893</td>
<td>1.183</td>
<td>0.051</td>
<td>0.969</td>
<td>0.699</td>
</tr>
</tbody>
</table>

The bold numbers represent the values of performance criteria for the best fitted models.
Figure 5 | Scatter plots of observed vs. estimated for proposed models in training and testing stages at Siirt station.
the observed and forecasted values, respectively; and \( N \) is the number of data.

**MODELS APPLICATION AND RESULTS ANALYSIS**

**Models implementation based on EEMD for \( P_E \) forecasting**

The main goal of the EEMD-MT/SVM models is to forecast monthly \( P_E \) at different stations in Turkey. According to Figure 3, the workflow of the proposed approach contains three main steps to enhance the performance of proposed forecasting techniques that can be expressed as follows:

1. In Step 1, the EEMD procedure is firstly utilized to decompose the all original input/output time series \( y(t) \) into several IMF components \( C_i(t) \) \( (i = 1, 2, 3, \ldots, n) \) and one residual component \( r_n(t) \).
2. Next, for each extracted IMF component and the residual component (for example IMF1), the SVM and MT are

![Figure 6](https://iwaponline.com/hr/article-pdf/50/2/498/549273/nh0500498.pdf)

**Figure 6** Time series plot of proposed models in testing stage at Siirt station.
established as forecasting models to simulate the decomposed IMF and residual components, and to predict each component by the same sub-series (IMF1) of four input variables, respectively.

3. In Step 3, the forecasted values of all extracted IMF and residue components by the SVM and MT models are aggregated to generate the modelled $P_E$ time series.

To summarize, the hybrid EEMD-SVM/MT forecasting tools establish the idea of ‘decomposition and ensemble’. The decomposition is used to simplify the forecasting process, and the ensemble is utilized to formulate a consensus forecasting on the original time series.

**Forecasting of monthly $P_E$ using proposed models at Siirt station**

The SVM, MT, EEMD-SVM, and EEMD-MT models were employed for forecasting monthly pan evaporation using $T_A$, $R_H$, $W_S$, and $S_R$ variables as inputs in Siirt station. These are the most effective variables on $P_E$ and previous studies (Tabari et al. 2012; Kisi et al. 2016; Malik et al. 2017; Wang et al. 2017) also used these variables for $P_E$ forecasting. Although standalone SVM and MT use original four inputs, $T_A$, $R_H$, $W_S$, and $S_R$, decomposed IMFs of input variables in each frequency are used to develop EEMD-SVM/MT models. The number of ensemble and the amplitude of white noise should be set before using EEMD. The relationship of the standard deviation of error ($e_n$) which is based on the ensemble number ($n$) and the amplitude of the added white noise ($\epsilon$) is described by Equation (15) (Wu & Huang 2009):

$$e_n = \frac{\epsilon}{\sqrt{n}}$$

(15)

According to previous studies (Wu & Huang 2009), in general, the amplitude of added noise should be about 0.2 times the standard deviation of the dataset and the number of ensembles used was 500. For more detailed information on the analysis of noise and ensemble averaging effects on decomposition, the readers are referred to Wu & Huang (2009). As an example, decomposed time series of pan evaporation data are represented in Figure 4. As seen from the figure, original monthly $P_E$ time series have been decomposed into eight independent IMF components in the order from the highest frequency (IMF1) to the lowest frequency (IMF8), and one residue component. Similarly, $T_A$, $R_H$, $W_S$, and $S_R$ inputs were also decomposed in IMF components. IMF1 of each input variable ($T_A$, $R_H$, $W_S$, and $S_R$) was used for forecasting IMF1 of output ($P_E$). Simultaneously, IMF2, IMF3, … of each input variable were used for forecasting IMF2, IMF3, … of $P_E$. Thereafter, all forecasted values of IMSs were summed to obtain original $P_E$.

Next, the accuracy of the proposed models in forecasting of monthly $P_E$ was assessed at both training and testing stages (Table 2). Based on the comparison between SVM and MT approaches, the MT is found to perform better than the SVM in terms of statistical metrics during training and testing stages at Siirt station. The quantitative evaluation results in Table 2 also show that EEMD-MT has a lower value of RMSE (0.765 mm) and PI (0.079 mm) compared to the EEMD-SVM model in the training stage for the Siirt station. Moreover, the EEMD-MT provided more accurate results in terms of NSE (0.89), WI (0.969), and LMI (0.699) than the SVM and MT.

In the testing stage, with respect to the measured value of $P_E$, it can be seen from Table 2, in terms of NSE and WI, that the EEMD-MT and EEMD-SVM models are able to model $P_E$ at Siirt station well. The SVM model has the highest error in

<table>
<thead>
<tr>
<th>Models</th>
<th>NSE</th>
<th>RMSE (mm/day)</th>
<th>PI (mm/day)</th>
<th>WI</th>
<th>LMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total available data in training stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.866</td>
<td>1.521</td>
<td>0.196</td>
<td>0.925</td>
<td>0.711</td>
</tr>
<tr>
<td>MT</td>
<td>0.882</td>
<td>1.451</td>
<td>0.176</td>
<td>0.953</td>
<td>0.755</td>
</tr>
<tr>
<td>EEMD-SVM</td>
<td>0.901</td>
<td>1.270</td>
<td>0.138</td>
<td>0.972</td>
<td>0.773</td>
</tr>
<tr>
<td>EEMD-MT</td>
<td>0.928</td>
<td>1.075</td>
<td>0.125</td>
<td>0.981</td>
<td>0.789</td>
</tr>
<tr>
<td>Total available data in testing stage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.869</td>
<td>1.462</td>
<td>0.149</td>
<td>0.958</td>
<td>0.703</td>
</tr>
<tr>
<td>MT</td>
<td>0.907</td>
<td>1.362</td>
<td>0.139</td>
<td>0.963</td>
<td>0.712</td>
</tr>
<tr>
<td>EEMD-SVM</td>
<td>0.902</td>
<td>1.268</td>
<td>0.138</td>
<td>0.973</td>
<td>0.733</td>
</tr>
<tr>
<td>EEMD-MT</td>
<td>0.923</td>
<td>1.098</td>
<td>0.054</td>
<td>0.985</td>
<td>0.795</td>
</tr>
</tbody>
</table>

The bold numbers represent the values of performance criteria for the best fitted models.
Figure 7 | Scatter plots of observed vs. estimated for proposed models in training and testing stages at Diyarbakir station.
estimating $P_E$ (RMSE = 2.663 mm; PI = 0.137 mm) compared to the other models. In addition, the EEMD-MT model improves the MT model with 44.71 and 49.50% reductions in term of RMSE and PI, respectively, and a 36, 8 and 42% increase in NSE, WI, and LMI, respectively.

Comparison of the modelled $P_E$ is made against measured values in the scatter plots (Figure 5) for both training and testing stages. The figure illustrates the superiority of the EEMD-MT model during the test stage for Siirt station. The superiority of the EEMD-MT model over the SVM, MT, and EEMD-SVM models is also clear from Figure 6. These results support our hypothesis that the EEMD method is suitable for decomposing monthly $P_E$ time series, and the idea of ‘decomposition and ensemble’ is feasible and the proposed EEMD-SVM/MT models can overcome drawbacks of the individual SVM and MT models by generating a synergetic effect in forecasting. Thus the monthly $P_E$ data decomposed by using the EEMD procedure can improve the forecast results at Siirt.

Figure 8 | Time series plot of proposed models in testing stage at Diyarbakir station.
Forecasting of monthly $P_E$ using proposed models at Diyarbakir station

Similar to the methodology used for Siirt station, four proposed models (i.e. SVM, MT, EEMD-SVM, and EEMD-MT) were developed for the purpose of forecasting monthly $P_E$ at Diyarbakir station. Specifically, SVM and MT were adopted as benchmark models and the results compared with the conjunction models which used EEMD to pre-process the original monthly $P_E$ data, adopting the phase-space reconstruction method to design the input vectors. The training and testing data sets were identical across all models. Table 3 indicates the performance results for the four models in terms of the four aforementioned statistical error metrics. The results show that the hybrid techniques have the lowest RMSE and PI values and the highest NSE and WI, outperforming the SVM and MT models, at both training and testing stages. In the training stage, the comparison of the SVM and MT models shows that the MT model has better accuracy ($NSE = 0.882$, $WI = 0.953$, and $LMI = 0.755$) in forecasting of $P_E$; however, these models’ results are significantly improved by coupling with the EEMD technique. The EEMD-MT approach outperformed all other models (Table 3). By integrating the EEMD with SVM, RMSE and PI were reduced to 1.270 and 0.138 mm in the training stage, respectively. It can be seen from Table 3 that by coupling with EEMD, the RMSE and PI of single MT were reduced by 25.9–28.9 and 19.4–75.5% in the training and testing stages, respectively. Improvements in the forecast results were approximately 5.2, 6.1 and 4.5% respectively, for NSE, WI and LMI values at the training stage, and 1.7, 2.2 and 11.6%, respectively, in the testing stage.

The measured and forecasted values of $P_E$ during training and testing stages are compared in Figures 7 and 8 for the four models. It can be seen from Figure 8 that SVM and MT models are able to follow the general trend of measured $P_E$ quite well, although these models are not able to match the extreme high and low values accurately, especially at the latter part of the testing period. In fact, the SVM and MT models under-predict $P_E$ for the testing data from 2006 onwards, whereas the EEMD-SVM/MT models are able to provide accurate estimates during this period of time (with the exception of 2008 where it is noticed that $P_E$ is under-predicted by all models). In general, the results of this analysis illustrate that the proposed EEMD-SVM/MT models are able to obtain better results than the SVM/MT models, shown by the improvements in different evaluation measures.

Models assessment at spring and summer seasons

To further evaluate the performance of the proposed models, the results were assessed in the spring and summer seasons only, since these are the main crop growing periods. Hence, further analysis was conducted for the period from May to October for 1975–2008 in both stations. Table 4 illustrates the results of the EEMD-SVM and EEMD-MT models in forecasting $P_E$ from May to October for Siirt station. As can be seen in Table 4, the EEMD-MT has a lower error in terms of RMSE and PI for all months compared to the EEMD-SVM model. It is notable that the models’ accuracy (NSE, WI and LMI) from May to July reduces but then increases for the months of August, September, and October.

In Diyarbakir station, the forecasted $P_E$ values in both the spring and summer seasons with the EEMD-SVM/MT models have a similar trend to the Siirt station (Table 5). Also, it can be said that the EEMD-SVM has inadequate accuracy in most of the months and it could just be a
Moving to a more detailed analysis on seasonal forecasting of the proposed models, Figure 9 illustrates RMSE and PI indices along May–October for the aforementioned period. As can be seen in both stations, the EEMD-MT model has lower error compared to EEMD-SVM. This indicates that the EEMD-SVM model may not provide a complete solution to every case. This also shows that accurate prediction of monthly $P_E$ is a complex issue.

### Table 5

<table>
<thead>
<tr>
<th>Models</th>
<th>Month</th>
<th>NSE</th>
<th>RMSE (mm/day)</th>
<th>PI (mm/day)</th>
<th>WI</th>
<th>LMI</th>
</tr>
</thead>
<tbody>
<tr>
<td>EEMD-SVM</td>
<td>May</td>
<td>0.677</td>
<td>0.802</td>
<td>0.068</td>
<td>0.924</td>
<td>0.723</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td>0.664</td>
<td>1.159</td>
<td>0.092</td>
<td>0.927</td>
<td>0.715</td>
</tr>
<tr>
<td></td>
<td>July</td>
<td>0.183</td>
<td>1.498</td>
<td>0.085</td>
<td>0.812</td>
<td>0.562</td>
</tr>
<tr>
<td></td>
<td>August</td>
<td>0.380</td>
<td>0.973</td>
<td>0.055</td>
<td>0.825</td>
<td>0.593</td>
</tr>
<tr>
<td></td>
<td>September</td>
<td>0.026</td>
<td>1.039</td>
<td>0.082</td>
<td>0.760</td>
<td>0.463</td>
</tr>
<tr>
<td></td>
<td>October</td>
<td>0.535</td>
<td>0.820</td>
<td>0.106</td>
<td>0.849</td>
<td>0.617</td>
</tr>
<tr>
<td>EEMD-MT</td>
<td>May</td>
<td>0.805</td>
<td>0.623</td>
<td>0.053</td>
<td>0.953</td>
<td>0.753</td>
</tr>
<tr>
<td></td>
<td>June</td>
<td>0.523</td>
<td>1.381</td>
<td>0.073</td>
<td>0.832</td>
<td>0.622</td>
</tr>
<tr>
<td></td>
<td>July</td>
<td>0.620</td>
<td>0.848</td>
<td>0.035</td>
<td>0.892</td>
<td>0.662</td>
</tr>
<tr>
<td></td>
<td>August</td>
<td>0.652</td>
<td>0.729</td>
<td>0.029</td>
<td>0.896</td>
<td>0.678</td>
</tr>
<tr>
<td></td>
<td>September</td>
<td>0.601</td>
<td>0.666</td>
<td>0.038</td>
<td>0.903</td>
<td>0.692</td>
</tr>
<tr>
<td></td>
<td>October</td>
<td>0.866</td>
<td>0.440</td>
<td>0.037</td>
<td>0.969</td>
<td>0.783</td>
</tr>
</tbody>
</table>

Predicting a hydro-meteorological time series is often complicated and frustrating. The series may appear completely random; a volatile sequence showing no signs of predictability. Time series are full of patterns and relationships. Decomposition aims to identify and separate them into distinct components, each with specific properties and behavior. In this study, the EEMD procedure was applied to tackle the trends and random behavior of time series data which may impact on accuracy of $P_E$ forecasting. This research attempted to improve SVM and MT models in the prediction of pan evaporation by coupling with the EEMD procedure. The $P_E$ time series from two climatic stations in Turkey, Siirt and Diyarbakir, were utilized for validation of the proposed coupled models. Firstly, the SVM and MT models alone were applied to forecast monthly $P_E$. In order to enhance the forecasting accuracy
of the typical models, coupled EEMD-SVM and EEMD-MT models were then used. For the coupled models, the original time series data were decomposed into eight IMFs and one residual for $P_E$ modeling process. Input variables included $T_A$, $R_{ht}$, $W_S$, and $S_R$. The performance of the proposed models was assessed in terms of NSE, RMSE, PI, WI, and LMI in the training and testing phases. At Siir station, the EEMD-MT model provided more accurate results in terms of NSE (0.89), WI (0.97), and LMI (0.70) than the SVM model. The proposed results showed that the SVM and MT models coupling accuracy was achieved with the EEMD-MT model (NSE (0.89), WI (0.97), and LMI (0.70) than the SVM model. EEMD-MT model provided more accurate results in terms of NSE, RMSE, PI, WI, and LMI compared with other pre-processing methods such as CEEMDAN and improved CEEMDAN in future studies.

REFERENCES


First received 12 April 2018; accepted in revised form 18 October 2018. Available online 8 November 2018