# *Estimation of monthly pan evaporation using support vector machine in Three Gorges Reservoir Area, China*

Article

Accepted Version

# Estimation of monthly pan evaporation using support vector machine in Three Gorges Reservoir Area, China

Ji-Long Chen [a], Hong Yang [b], Ming-Quan Lv [a], Zuo-lin Xiao [c*], Sheng-Jun Wu [a]

[a] Key Laboratory on Water Environment of Reservoir Watershed, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 401122, China
[b] Department of Geography and Environmental Science, University of Reading, Reading, Whiteknights, RG6 6AB, UK
[c] The Key Laboratory of GIS Application Research, Chongqing Normal University, Chongqing 401331, China

*Corresponding author:
Email address: xiaoll@cqnu.edu.cn

**Abstract:** Pan evaporation plays a critical role in estimating water budget and modelling crop water requirements. However, it has been measured at a very limited number of meteorological stations. Estimation of pan evaporation from measured meteorological variables offers an important alternative and drawn increasing attention in the recent years. This paper investigated the performance of support vector machine (SVM) in estimation of monthly pan evaporation using commonly measured meteorological variables in Three Gorges Reservoir Area in China. Evaluation suggested that SVM models showed remarkable performances and significantly outperformed the empirical model. The SVM model with polynomial as kernel function outperformed that with radial basis function. In the case of unavailable measurements of pan evaporation and meteorological variables to construct the SVM model, pan evaporation can be well estimated by SVM model developed using data at other sites. The results indicated that the SVM method would be a promising alternative over the traditional approaches for estimating pan evaporation from measured meteorological variables.
**Key words:** pan evaporation, support vector machine, meteorological variables, Three Gorges Reservoir Area

# 1 Introduction

The critical importance of pan evaporation in estimating water budget and modelling crop water requirements has been well documented (Li et al. 2018; Flammini et al. 2018). It has been widely used as an indicator for estimating free water evaporation and reference evapotranspiration (Wang et al. 2017a). The realistic method for measuring pan evaporation is to use Class A pan evaporimeter. However, due to the large cost of the measurement devices and its difficult maintenance (Kisi et al. 2016), pan evaporation has been measured only at a very limited number of meteorological stations (Shirsath and Singh 2010). Therefore, many attempts have been made to estimate pan evaporation with modelling approach (Malik and Kumar 2015).

Two major methods have been generally used to calculate pan evaporation. One is mechanism method which simulates physical processes of evaporation based on the principles of energy budget and mass exchanges. Martínez et al. (2006) developed a multilayer model based on the discretization of the pan water volume into several layers. Choudhary and Klauda (2016) presented a physical model based on theoretical models for mass and energy transfer. Gentine et al. (2016) introduced a new methodology which is theoretically based on the budget of heat and moisture in the boundary layer. However, it is still difficult to build an accurate model representing all the processes due to the nonlinear and complex process of evaporation (Lin et al. 2013; Sanikhani et al. 2012). Moreover, the physical models are generally complex and their calibrations require a large amount of data and computational costs (Tan et al. 2007), limiting the practical applications of physical models in researches in environment, agriculture and ecology. As a consequence, the other alternative has been widely developed.

In particular, a large number of data-driven models have been created. For example, the empirical model and machine learning algorithm have been extensively investigated. Stephens and Stewart (1963) developed an empirical model using radiation and air temperature. This model was found to perform best among 23 models in extremely arid areas (Al-Shalan and Salih 1987). Hanson (1989) presented an empirical equation using radiation and air temperature in the USA. Linacre (1977) proposed a simple model using temperature in Australia. Rotstayn et al. (2006) coupled the radiative component and the aerodynamic component to develop the PenPan model, which was later validated by Roderick et al. (2007) and Johnson and Sharma (2010) across Australia. Lim et al. (2016) modified the PenPan model to present the PenPan-V2 model, which was found to outperform the original PenPan model in Australia. Patel and Majmundar (2016) obtained empirical relations as functions of air temperature, relative humidity, wind velocity, and sunshine duration in India. Andreasen et al. (2017) developed multilinear regression models using various combinations of meteorological variables in the USA. The main benefit of empirical models is that the meteorological variables are routinely measured and easily available. However, they can only be applied to the places with similar climatic conditions (Goyal et al. 2014). Moreover, the empirical models cannot provide accurate estimations due to the complex process of evaporation (Shalamu 2011).

Many researchers have explored the potential of machine learning algorithms in estimating pan evaporation. Sudheer et al. (2002) investigated the abilities of artificial neural networks (ANN) using meteorological variables, and the results showed that ANN model performed better than empirical models in Georgia, USA. Later, a large number of works evaluated the performances of ANN in Iran (Tabari et al. 2012), Turkey (Tezel and Buyukyildiz 2016), India (Malik et al. 2017), and China (Wang et al. 2017a), and the results confirmed the superiority of ANN over the empirical

models (Kim et al. 2013, 2014, 2015; Kisi et al. 2016; Malik and Kumar 2015). Guven and Kisi (2011) presented genetic programming (GP) to estimate pan evaporation in California. Evaluations suggested that GP model gave better estimation than ANN and empirical models (Guven and Kisi 2013; Kim et al. 2015; Kişi and Tombul 2013). Malik et al. (2017) compared the performances of multi-layer perceptron neural network (MLPNN), self-organizing map neural network (SOMNN), ANFIS, ANN and empirical models in Indian. Wang et al. (2017c) investigated the performances of fuzzy genetic (FG), adaptive neuro-fuzzy inference systems (ANFIS) and M5 model tree and empirical models in China, and the results indicated that FG model generally produced better results than other models.

In recent years, a novel machine learning algorithm, support vector machine (SVM), has been widely applied and proven to be a robust algorithm for estimation (Chen et al. 2015). Some researches have explored the potential of SVM in the estimation of pan evaporation from meteorological variables. Kim et al. (2012) evaluated the performance of SVM in Korea and Iran, and results suggested that SVM was superior to ANN and empirical models. Lin et al. (2013) compared the accuracy of SVM and ANN models and reported that SVM generated better estimation than ANN. Goyal et al. (2014), Kisi (2015), and Tezel and Buyukyildiz (2016) evaluated the performances of SVM, ANN and empirical models and also reported that SVM outperformed other models. Pammar and Deka (2017) explored the potential of SVM for pan evaporation estimation in India. Wang et al. (2017b) compared the performances of SVM, FG, ANFIS, multivariate adaptive regression spline(MARS) and empirical models in China, and the overall results indicated that SVM performed better than other methods.

China Three Gorge Project (TGP), officially launched in 1994 and fully operated in 2008, ranked as the largest hydropower project in the world (Xu et al., 2013). The climatic impacts of TGP in Three Gorges Reservoir Area (TGRA) were seriously debated since the late 1950s and remain so afterwards (Fu et al., 2010). Researches indicated that TGP has affected regional air temperature (Yao et al. 2013) and precipitation (Lv et al. 2017). These changes of meteorological variables may affect evaporation in TGRA, which would change the energy budget and mass exchange (Yang et al. 2012). Therefore, it is of great significance to investigate the evaporation changes induced by the construction of TGP for understanding its climatic impacts. However, due to the lack of sufficient measurements of evaporation, it is still largely unknown about the effect of the construction and operation of TGP on regional evaporation. Therefore, it is important and urgent to estimate evaporation with novel method in TGRA.

With respect to the rising popularity of SVM and the increasing need of evaporation for studying the climatic impacts of TGP. The main objectives of this study are (1) to explore the performance of SVM for estimating month pan evaporation using commonly measured meteorological variables in TGRA; (2) to compare the SVM model with empirical model; and (3) to explore the feasibility of pan evaporation estimation at one site using SVM model developed using data from other sites.

## 2 Materials and method

### 2.1 Study area

TGRA (Figure 1) is located in the upstream of the Yangtze River, to the east of Sichuan Basin, to the north of Daba Mountain, and bordering the western Yangtze Plain. It stretches along the Yangtze River from Jiangjin county in Chongqing municipality to Yichang county in Hubei province, with the area of $5.79\times10^4$ km$^2$. The geography is complex and the elevation generally decreases from northeast to southwest. The region is dominated by mountainous and hilly areas.

TGRA is located in the transfer zone between the northern temperate zone and the subtropical zone. The climate of TGRA is subtropical monsoon climate which is characterized by four distinct seasons with a hot and humid summer, a warm to cool winter, and plenty of precipitation with a highly pronounced summer peak. Annual mean temperature is between 16.5℃and 19℃, and annual precipitation is about 1100 mm (Lv et al. 2016).

### 2.2 Sites and data

Six meteorological stations measuring pan evaporation and meteorological variables including solar radiation, sunshine duration, atmospheric pressure, vapor pressure, wind velocity, air temperature and relative humidity were used in the current study (Figure 1). The mapping of stations roughly range from 29° 35 ' to 31° 3' N (latitude), from 106°28' to 111° 30' E (longitude), and from 133.1 m to 607.3 m altitude. Table 1 shows the detailed information about the meteorological stations.

Monthly data records covering the period between 1970 and 2013 were obtained from the Chinese National Meteorological Information Center. Pan evaporation (mm) was measured by using a metal pan, 20 cm in diameter and 10 cm high, installed 70 cm above the ground. Solar radiation (MJ $m^2$) was measured by Pyranometer (Chen et al., 2010). The type of Pyranometer used by CMA was changed in 1993 (Chen et al., 2019). However, the homogeneity of the radiation data was unlikely to be affected because the instrument has been calibrated to the same standard following the guidelines of the World Meteorological Organization (WMO) (Chen and Li 2014; Yang et al., 2009). Sunshine duration (h) was measured by using Jordan sunshine recorder. Air temperatures (°C) was measured by using mercury and alcohol thermometers. Vapor pressure (Kpa) was measured by using adjustable cistern barometer. Relative humidity (%) was measured by using aspirated psychrometer at 2m height (CMA 1979). Wind velocity (m/s) was measured by using EL wind electric anemometer at 10 m height, which was transformed to speed at 2m height by a logarithmic model proposed by FAO56 (Allen et al. 1998). All the instruments were calibrated periodically and all the measurements were made following the guide of World Meteorological Organization (Chen and Li 2013).

Although preliminary quality control tests were conducted by the supplier, meteorological measurements may still contain errors due to inappropriate measuring manipulations and occasional voltage instability (Qin et al. 2011). We further checked the data according to the quality control scheme presented by Feng et al. (2004) and Tang et al (2010). First, records with missing data which were replaced by 32766 were removed. Second, the data with evident systematic and operational errors were removed. For example, actual global radiation exceeds extra-terrestrial radiation, sunshine duration exceeds potential sunshine duration, and the relative humidity exceeds 100%. Lastly, we used a relative simple method to build the relationship between pan evaporation and the meteorological variables and then to exclude noisy data and suspected data. More details can be found in Feng et al., (2004) and Tang et al. (2010). Two data sets were subsequently created for each station, and the first 70% of the record were used for modelling and the remaining 30% were used for validation.

### 2.3 Data description

Figure 2 shows the interannual variation of pan evaporation in TGRA. Pan evaporation showed a decreasing trend from 1970 to 2013, with the rate of -81.14mm/10y. After the Yangtze River interception and water impoundment of Three Gorges Dam in 1997, pan evaporation decreased with the rate of -84.66 mm/10y, which is insignificantly different from that in the period 1970 - 2013. Figure 3 shows the distributions of the monthly daily meteorological variables of the studied sites.

Monthly daily solar radiation varied between 4.42 and 16.22 MJ $m^{-2}$, monthly daily sunshine duration varied between 1.38 and 6.38h, and monthly daily air temperature varied between 6.4 and 28.15℃. These three meteorological variables showed similar change patterns with maximum in July and minimum in January (Figure 3.a-b). Monthly daily vapor pressure ranged from the minimum of 0.75 kPa in January and the maximum of 2.84 kPa in June (Figs.3c), which was generally opposite to that of atmospheric pressure with the minimum of 96.9 kPa in June and the maximum of 99.06 kPa in January (Figure 3c). Monthly daily relative humidity ranged between 72.56% and 79.88%, and monthly daily wind velocity ranged between 1.15 and 1.5m/s (Figure 3d), without clear seasonal pattern.

**2.4 Theory of support vector machine**

SVM is a machine learning algorithm developed by Vapnik (1995) and has been widely applied and proven to be a robust algorithm for estimation (Chen et al. 2013; 2015). SVM has several features compared with traditional machine learning models which perform the empirical risk minimization. SVM is based on the machine learning theory and principle of structural risk minimization which minimizes the upper bound of generalization error. SVM constructs the regression function using kernel functions defined in a high dimensional space. It delivers a unique solution since the optimal problem is convex. More detailed principle of SVM can be found in Vapnik (1995; 1998).

Given a set of data points $(x_i, d_i)$, SVM constructs regression function as the following form:

$$f(x) = w\phi(x) + b \tag{1}$$

where $\phi(x)$ is high dimensional feature space. $W$ and b are coefficients estimated by minimizing the following function:

$$R_{SVMs}(C) = C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i) + \frac{1}{2}\|w\|^2 \tag{2}$$

where $C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i)$ is empirical error measured by loss function $L\varepsilon$ . $\frac{1}{2}\|w\|^2$ is regularization term. $C$ is a constant used to specify the trade-off between $C\frac{1}{n}\sum_{i=1}^{n} L(d_i, y_i)$ and $\frac{1}{2}\|w\|^2$ .

$$L\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon & |d - y| \geq \varepsilon \\ 0 & otherwise \end{cases} \tag{3}$$

where $\varepsilon$ is prescribed accuracy representing for the tube size.

After introducing slack variables $\zeta_i$ and $\zeta_i^*$, Eq. (2) is transformed to the following constrained form:

$$\text{Minimize} \quad R_{SVMs}(w, \zeta^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\left(\zeta_i + \zeta_i^*\right) \tag{4}$$

$$\text{Subjected to} \quad \begin{matrix} d_i - w\phi(x_i) - b_i \leq \varepsilon + \zeta_i \\ w\phi(x_i) + b_i - d_i \leq \varepsilon + \zeta_i^*, \zeta_i^* \geq 0 \end{matrix} \tag{5}$$

The optimization problem (5) can be solved using Lagrange multipliers.

$$W(a_i, a_i^*) = \sum_{i=1}^{n} y_i(a_i - a_i^*) - \varepsilon\sum_{i=1}^{n}(a_i + a_i^*) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(a_i - a_i^*)(a_j + a_j^*)(\phi(x_i)\cdot\phi(x_j)) \quad (6)$$

$$\text{Subjected to } \sum_{i=1}^{n}(a_i - a_i^*) = 0 \quad (7)$$

$$0 \le a_i, a_i^* \le C$$

The regression function becomes:

$$f(x) = \sum_{i=1}^{n}(a_i - a_i^*)\phi(x_i)\cdot\phi(x_i) + b \quad (8)$$

By introducing kernel function $K(x_i, x_j)$, Eq. (8) can be rewritten as follow:

$$f(x, a_i, a_i^*) = \sum_{i=1}^{n}(a_i - a_i^*)K(x, x_i) + b \quad (9)$$

$K(x_i, x_j)$ is equal to the inner product of two vectors $x_i$ and $x_j$ in the feature space $\phi(x_i)$ and $\phi(x_j)$. Any function that meets Mercer's condition can be used as kernel function. Four basic kernel functions including linear, polynomial, radial basis function (RBF) and sigmoid are provide by SVM. The linear kernel function is a special case of radial basis function (Vapnik 1996). The sigmoid kernel function behaves like RBF and is not valid under some conditions (Vapnik 1998). Therefore, the polynomial kernel function $K(x_i, x_j) = (x_i.x_j + 1)^d$ and radial basis kernel function $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$ are widely used, where $d$ and $\gamma$ are the kernel function parameters.

The SVM algorithm was implemented by MATLAB software with the LIBSVM tool developed by Chang and Lin (2001). The main procedures included selection of the input attributes, data scaling, selection of kernel function, training and testing (Hsu et al. 2003). Seven meteorological variables including solar radiation, sunshine duration, atmospheric pressure, vapor pressure, wind velocity, air temperature and relative humidity were used as input variables, while pan evaporation was used as dependent variables. Data scaling can avoid calculation difficulties and improve data fitting. In this work, all the data were scaled to the range [0, 1] using the same linearly scaling method. The polynomial and radial basis kernel functions were employed and compared. The optimum ranges of SVM parameters were determined by grid search (Gestel et al. 2004), and the optimum values were then obtained from range using cross validation (Cherkassky and Mulier 1998). The training data were divided into subsets with equal size, each subset was tested once using the trained SVM model on the remaining subsets. After the optimum values of parameters were found, the final model was trained on the whole training data set using the selected parameters (Hsu et al. 2003), and the accuracy was evaluated using the testing data.

### 2.5 Empirical model

An empirical model for the estimation of pan evaporation using meteorological variables was developed as the following form:

$$Pan = aR_a + bS + cAp + dVp + eW + fT + gR + h \quad (10)$$

Where $R_a$, $S$, $A_p$, $V_p$, $W$, $T$ and $R$ were solar radiation, sunshine duration, atmospheric pressure, vapor pressure, wind velocity, air temperature and relative humidity, respectively, $a$-$h$ were empirical coefficients which were determined by least square regression to minimize the fitting squared residuals using the modelling dataset. The calibrated coefficients are presented in Table 2.

### 2.6. Error indicators

Root mean square error (RMSE) and relative root mean square error (RRMSE) (%) are used to evaluate the accuracy of developed models, higher values of RMSE, and RRMSE indicate better performances. Coefficient of determination ($R^2$) is used to measure the fit of model, and higher the

value, better the fit.

## 3 Results and discussions

### 3.1 Pan evaporation in TGRA

Annual mean pan evaporation of in TGRA varied between 1028.57 and 1655.02 mm with the average of 1264.72mm, which was much lower than the average pan evaporation over China (Yang and Yang 2012). It mainly occurred in summer, which accounted for 42.5% of the annual amounts, whereas it was very low in winter, accounting for 10.28% of the annual amounts. Pan evaporation generally increased northeastward along the main stream of Yangtze River, and it was correlated significant with longitude ($r=0.763$, $p<0.01$) and latitude ($r=0.672$, $p<0.01$) (Figure 4.a-b). Higher pan evaporation in northeastern region might be contributed by the higher global solar radiation, sunshine duration and wind velocity in this region. As shown in Figure 4c-e, pan evaporation was positively correlated with global solar radiation ($r=0.904$, $p<0.001$), sunshine duration ($r=0.853$, $p<0.001$) and wind velocity($r=0.732$, $p<0.01$), while lower pan evaporation in southwestern region might be due to the higher relative humidity and water vapor in this region, as it negatively correlated with them with the correlation coefficients of 0.923 ($p<0.001$) and 0.787 ($p<0.01$)(Fig.4f-g), respectively.

### 3.2 Performance of developed models

Error indicators of the developed models were presented in Table 3. Overall, empirical model gave reasonable estimations with the RMSE varying between 10.073 and 21.535 mm (average 13.79 mm) and RRMSE varying between 9.73% and 22.22% (averaged 13.44%). The new empirical model performed much better than the PenPan model in the USA (Hobbins et al. 2012), Spain (Azorin-Molina et al. 2015), and China (Liu and Sun 2016) and the Romanenko model and Meyer model in Turkey (Tezel and Buyukyildiz 2016). Yu et al. (2017) modified the PenPan model to present a new empirical model with better accuracy in northwest China. However, both models were inferior to our new empirical model. The possible reason may be due to that we used more meteorological variables, which could increase the fit of the new empirical model. It is worth noting that the main difficulty in limiting the universal applicability of our developed empirical model to other regions was the empirical coefficients which varied greatly from sites to sites. This was as a result of its complexity, intrinsic quality of equipments, the topographical and the local climate characteristics (Liu and Sun, 2016). Nevertheless, it was reasonable that our results were applicable to the regions with similar climatic conditions and topography. For other places of interest, the coefficients of the empirical model can be calibrated following the scheme of this study with relative ease, making the model appropriate for widespread applications.

All SVM models gave good performances with the RMSE varying between 4.808 and 14.333mm (average 8.136mm) and RRMSE varying between 5.62% and 14.79% (averaged 7.92%), suggesting that SVM showed remarkable performance in the estimation of monthly pan evaporation from meteorological variables. In recent years, many works had explored the potential of machine learning algorithms in estimating pan evaporation. The SVM models in our work performed much better than ANN algorithm in Turkey (Tezel and Buyukyildiz 2016), China (Wang et al. 2017a; b) and India (Pammar and Deka 2017), the FG and ANFIS algorithms in China (Wang et al.2017a), the GP algorithm in Turkey (Guven and Kisi 2013). While the accuracy of SVM models were influenced by kernel functions. $SVM_{pol}$ had the average RMSE of 7.967 mm and RRMSE of 7.75%, which were slightly lower than the RMSE of 8.305 mm and RRMSE of 8.09% for $SVM_{rbf}$,

suggesting that SVM model using polynomial as kernel function outperformed that using radial basis function. However, in many researches, RBF was commonly used and recommended other than polynomial kernel function (Wang et al. 2017c). It was probable that polynomial kernel function has more hyperparameters, which influenced the complexity of the computations (Dong et al. 2005). Our results indicated the importance of selecting an appropriate kernel function for the accuracy of SVM model.

### 3.3 Comparison between empirical model and SVM models

Comparisons showed that $SVM_{pol}$ had 35.72%~52.26% (average 43.24%) lower RMSE, and 4.03%~7.94% (average 5.68%) lower RRMSE than the emperical model, and the $SVM_{rbf}$ had 33.44%~52.15% (average 40.88%) lower RMSE, and 4%~7.43% (average 5.56%) lower RRMSE than the emperical model, suggesting that SVM models significantly outperformed the empirical model. Out results were consistent with those from Goyal et al. (2014), Kisi (2015), Tezel and Buyukyildiz (2016), Pammar and Deka (2017), and Wang et al.(2017a; b; c) who compared the performances of SVM model against empirical models and reported that SVM performed better than empirical models. Favorable agreements had been observed between estimations and measurements, with the $R^2$ varying between 0.953 and 0.993 (average 0.981) for $SVM_{pol}$ (Fig.5), and between 0.916 and 0.972 (average 0.957) for $SVM_{rbf}$ (Fig.6). The estimations of SVM models were much closer to the observations than those of the empirical model (Fig.7), further confirming the superiority of SVM over the empirical model.

### 3.4 Error variation of SVM models

Figure 8 showed the error variation of the best SVM model ($SVM_{pol}$). RMSE at each site showed similar variation pattern with maximum in summer and minimum in winter, and this may be due to the influence of the East Asian monsoon climate with large fluctuation of the meteorological variables in summer, resulting in the higher errors in this season. However, RRMSE generally showed an opposite changing pattern with maximum in winter and minimum in summer, contributing to the higher pan evaporation flux in summer than in winter. Excepted for Yichang, RMSE was positively correlated with global solar radiation, sunshine duration, air temperature, wind velocity and water vapor (Table 4), generally indicating that higher values of these meteorological variables would result in higher errors of $SVM_{pol}$, while RMSE was negatively correlated with atmospheric pressure. The correlation results further confirmed that errors of $SVM_{pol}$ was significantly affected by the climatic dynamics. RRMSE showed an opposite correlation trend to that of RMSE (Table 5).

### 3.5 Assessing the transferability of SVM model developed at one site to other sites

Long-term measured pan evaporation and meteorological variables were required to construct the SVM model. It was therefore open to question how to apply the model to the sites without measurements of pan evaporation. Therefore, estimations of pan evaporation using the SVM model developed at other sites were explored, and the performances were presented in Tables 6 and 7, in which the error indicators represented for the accuracy of SVM developed using the data from the sites in the first row for the estimation of pan evaporation at the sites in first column.

In the case of unavailable data to construct the model, monthly pan evaporation can be estimated by SVM model developed at other sites with reasonable accuracy. Wang et al. (2017b) investigated the applicability of ANN using data from nearby station, and the results indicated that RMSE increased by more than 50% in the absence of local data, suggesting that SVM was superior to ANN for estimations of pan evaporation using data from other sites. The acceptable alternative

sites had been highlighted for each site. As it can be seen, all the sites can been used to estimate pan evaporation at Wanzhou with the average RMSE of 7.637 mm and RRMSE of 8.74%, indicating that pan evaporation at this place was relatively easier to be estimated with the model developed at other sites. While at Badong and Yichang, RMSE were relatively higher than those of models developed using their own data. The average RMSE and RRMSE in each column indicated the representativeness of modelling sites for estimation of pan evaporation at other sites. Overall, Chongqing and Fengdu had relatively lower RMSE and RRMSE, implying that SVM models developed at these sites could be used to estimate pan evaporation at any location without measurements of pan evaporation in TGRA.

The findings had implications relative to the deployment of resource in hydrological and agricultural studies. Since many researchers were making efforts to obtain the data of pan evaporation at the places of interest where was not equipped with the evaporimeter. The results suggested that pan evaporation could be estimated by using SVM model developed at other suitable sites. For the studied sites in TGRA, selection of a suitable site can be guided by the results in this work, while for other sites, it depended on the climatic conditions, topography and the distance between the two sites. Therefore, future researches are still required to guide the selection of such suitable site.

## 4 Conclusions

Estimation of monthly pan evaporation using SVM algorithm with the commonly measured meteorological variables was investigated in TRGA. Our results suggested that SVM models showed remarkable performances and significantly outperformed the empirical model. While the accuracy of SVM models were influenced by kernel functions. The model using polynomial kernel function outperformed that using radial basis function. In the case of unavailable measurements of pan evaporation and meteorological variables to construct the SVM model, pan evaporation can be estimated by using SVM model developed at other sites with reasonable accuracy. The finding had potential attraction in generating monthly pan evaporation data for studies in agriculture, hydrology and ecology, especially for the large, remote areas where ground measurements were limited. However, further studies are needed to guide the selection of a suitable site where the data can be used to construct a transferable SVM model to the sites without measurements of pan evaporation. Moreover, development of powerful models with high level of reliability and better accuracy was still a challenging task. Recently, Pammar and Deka (2017) found that coupling different approaches to build a hybrid model was an effective and promising way. Thus, it is important for the future work to explore the hybrid model combining the SVM and other data statistic and analysis algorithms.

## Acknowledgment

The work was supported by National Natural Science Foundation of China (41401051), Chongqing Science and Technology project (cstc2015jcyjA00007). We thank the National Meteorological Information Center for providing the long-term records of pan evaporation and meteorological variables.

## References


Allen RG, Pereira LS, Raes D, Smith M. Crop evapotranspiration-guidelines for computing crop water requirements. In: FAO Irrigation and drainage paper 56. Rome: Food and Agriculture Organization of the United Nations; 1998.

Al-Shalan A, Salih AMA (1987) Evapotranspiration estimation in extremely arid areas. J Irrig Drain Eng ASCE 113(4):565-574.

Andreasen M, Rosenberry DO, Stannard DI (2017) Estimating daily lake evaporation from biweekly energy-budget data. Hydrol Process 31/DOI: 10.1002/hyp.11375

Azorin-Molina C, Vicente-Serrano SM, Sanchez-Lorenzo A, McVicar TR, Morán-Tejeda E, Revuelto J, Kenawy AE, Martín-Hernández N,Tomas-Burguera M (2015) Atmospheric evaporative demand observations, estimates and driving factors in Spain (1961–2011). J Hydrol 523: 262-277.

Chang C, Lin C (2001) LIBSVM: A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen JL, He L, Chen Q, Lv MQ, Zhu HL, Wen ZF, Wu SJ (2019) Study of monthly mean daily diffuse and direct beam radiation estimation with MODIS atmospheric product. Renewable Energy 132: 221-232.

Chen JL, Li GS (2013) Estimation of monthly average daily solar radiation from measured meteorological data in Yangtze River Basin in China. Int J Climatol 33: 487-498.

Chen, JL, Li GS (2014) Evaluation of support vector machine for estimation of solar radiation from measured meteorological variables. Theoretical and Applied Climatology 115: 627-638.

Chen JL, Li GS, Wu SJ (2013) Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. Energy Convers Manage 75: 311-318.

Chen JL, Li GS, Xiao BB, Wen ZF, Lv MQ, Chen CD, Jiang Y, Wang XX, Wu SJ (2015) Assessing the transferability of support vector machine model for estimation of global solar radiation from air temperature. Energ Convers Manage 89: 318-329.

Chen JL, Liu HB, Wu W, Xie DT (2011) Estimation of monthly solar radiation from measured temperatures using support vector machines-A case study. Renewable Energy 36(1): 413-20.

Cherkassky V, Mulier F (1998) Learning from data: Concepts, theory, and methods. Wiley, New York.

China Meteorological Administration (CMA) (1979) Specifications for surface meteorological observation. QX/T45-2007, p. 187.

Choudhary R, Klauda JB (2016) The Simultaneous Mass and Energy Evaporation (SM2E) Model. J Occup Environ Hyg 13(4):243-253.

Dong B, Cao C, Lee SE (2005) Applying support vector machines to predict building energy consumption in tropical region. Energy Buildings 37: 545-553

Feng S, Hu Q, Qian W (2004) Quality control of daily meteorological data in China, 1951–2000: a new dataset. Int J Climatol 24(7): 853-870.

Flammini A, Corradini C, Morbidelli R, Saltalippi C, Picciafuoco T, Giráldez JV (2018) Experimental Analyses of the Evaporation Dynamics in Bare Soils under Natural Conditions. Water Resour Manag 32(3):1153-1166.

Fu BJ, Wu BF, Lu YH, Xu ZH, Cao JH, Niu D, Yang GS, Zhou YM (2001) Three Gorges Project: Efforts and challenges for the environment. Prog Phys Geog 34(6): 741-754

Gentine P, Chhang A, Rigden A, Salvucci G (2016) Evaporation estimates using weather statio data and boundary layer theory. Geophys Res Lett 10.1002/2016GL070819.

Gestel TV, Suykens JAK, Baesens B, Viaene S, Vanthienen J, Dedene G, Moor BD, Vandewalle J (2004) Benchmarking least squares support vector machine classifiers. Mach Learn 54(1): 5-32.

Goyal MK, Bharti B, Quilty J, Adamowski J, Pandey A (2014) Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. Expert Syst Appl 41(11):5267-5276.

Guven A, Kişi O (2011) Daily pan evaporation modeling using linear genetic programming technique. J Hydrol 29(2):135-145.

Guven A, Kisi O (2013) Monthly pan evaporation modeling using linear genetic programming J Hydrol 503: 178-185.

Hanson CL (1989) Prediction of Class A pan evaporation in southwest Idaho. J Irrig Drain Eng, ASCE 115(2): 166-171.

Hobbins M, Wood A, Streubel D, Werner K (2012) What drives the variability of evaporative demand across the conterminous United States?. J Hydrometeorol 13:1195-1214.

Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. Available at: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Johnson F, Sharma A (2010) A comparison of Australian open water body evaporation trends for current and future climates estimated from Class A evaporation pan and general circulation models, J Hydrometeorol 11: 105-121.

Kim S, Shiri J, Kisi O (2012) Pan Evaporation Modeling Using Neural Computing Approach for Different Climatic Zones. Water Resour Manag 26(11):3231-3249.

Kim S, Shiri J, Kisi O, Singh VP (2013) Estimating Daily Pan Evaporation Using Different Data-Driven Methods and Lag-Time Patterns. Water Resour Manag 27(7):2267-2286.

Kim S, Shiri J, Singh VP, Kisi O, Landeras G (2015) Predicting daily pan evaporation by soft computing models with limited climatic data. Hydrol Sci J 60 (6): 1120-1136.

Kim S, Singh VP, Seo Y (2014) Evaluation of pan evaporation modeling with two different neural networks and weather station data. Theor Appl Climatol 17 (1): 1-13.

Kisi O (2015) Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 528:312-320.

Kisi O, Genc O, Dinc S, Zounemat-Kermani M (2016) Daily pan evaporation modeling using chi-squared automatic interaction detector, neural networks, classification and regression tree. Comput Electron in Agr 122:112-117.

Kişi Ö, Tombul M (2013) Modeling monthly pan evaporations using fuzzy genetic approach. J Hydrol 477: 203-212.

Li M, Chu R, Shen S, Armt I (2018) Dynamic analysis of pan evaporation variations in the Huai River Basin, a climate transition zone in eastern China. Sci Total Environ 625: 496-509.

Lim WH, Roderick ML, Farquhar GD (2016) A mathematical model of pan evaporation under steady state conditions. J Hydrol 540:641-658.

Lin GF, Lin HY, Wu MC (2013) Development of a support-vector-machine-based model for daily pan evaporation estimation. Hydrol Process 27(22):3115-3127.

Linacre ET (1977) A simple formula for estimating evaporation rates in various climates, using temperature data alone. Agric Met 18(6): 409-424.

Liu W, Sun F (2016) Assessing estimates of evaporative demand in climate models using observed pan evaporation over China. J Geophys Res Atmos 121/10.1002/2016JD025166.

Lv MQ, Chen JL, Mirza Zakaria A, Chen CD, Wen ZF, Jiang Y, Ma MH, Wu SJ (2016) Spatial distribution and temporal variation of reference evapotranspiration in the Three Gorges Reservoir area during 1960-2013. Int J Climatol 36: 4497-4511.

Lv MQ, Wu SJ, Chen JL, Chen CD, Wen ZF, Huang YY (2017) Changes in extremeprecipitation In the Yangtze River basin And its association with global mean temperature and ENSO. Int J Climatol doi:10.1002/joc.5311

Malik A, Kumar A (2015) Pan Evaporation Simulation Based on Daily Meteorological Data Using Soft Computing Techniques and Multiple Linear Regression. Water Resour Manag 29(6):1859-1872.

Malik A, Kumar A, Kisi O (2017) Monthly pan-evaporation estimation in Indian central Himalayas using different heuristic approaches and climate based models. Comput Electron in Agr 143: 302-313.

Martínez JMM, Alvarez VM, González-Real MM Baille A (2006) A simulation model for predicting hourly pan evaporation from meteorological data. J Hydrol 318(1-4):250-261.

Pammar L, Deka PC (2007) Daily pan evaporation modeling in climatically contrasting zones with hybridization of wavelet transform and support vector machines. Paddy Water Environ 15(4):711-722.

Patel JN, Majmundar BP (2016) Development of Evaporation Estimation Methods for a Reservoir in Gujarat India. J AmWater Works Ass 108(9):E489-E500.

Qin J, Chen Z, Yang K, Liang S, Tang W (2011) Estimation of monthly-mean daily global solar radiation based on MODIS and TRMM products. Appl Energ 88: 2480-2489.

Roderick ML, Rotstayn LD, Farquhar GD, Hobbins MT (2007) On the attribution of changing pan evaporation. Geophys Res Lett 34: L17403, doi:10.1029

Rotstayn LD, Roderick ML, Farquhar GD (2006) A simple pan-evaporation model for analysis of climate simulation: Evaluation over Australia. Geophys Res Lett 33: L17715, doi:10.1029.

Sanikhani H, Kisi O, Nikpour M R, Dinpashoh Y (2012) Estimation of Daily Pan Evaporation Using Two Different Adaptive Neuro-Fuzzy Computing Techniques. Water Resour Manag 26(15):4347-4365.

Shalamu A, King JP, Moreno J, Bawazir AS (2011) Modeling of daily pan evaporation using partial least squares regression. Sci China Technol Sci 54(1):163-174.

Shirsath PB, Singh AK (2010) A comparative study of daily pan evaporation estimation using ANN, regression and climate based models. Water Resour Manag 24(8):1571-1581.

Stephens JC, Stewart EH (1963) A comparison of procedures for computing evaporation and evapotranspiration. Publication 62, International Association of Scientific Hydrology. International Union of Geodynamics and Geophysics, Berkeley 123-133.

Sudheer KP, Gosain AK, Rangan DM, Saheb SM (2002) Modeling evaporation using an artificial neural network algorithm. Hydrol Process 16:3189-3202.

Tabari H, Talaee PH, Abghari H (2012) Utility of coactive neuro-fuzzy inference system for pan evaporation modeling in comparison with multilayer perceptron. Meteorol Atmos Phys 116:147-154.

Tan SBK, Shuy EB, Chua LHC (2007) Modelling hourly and daily open-water evaporation rates in areas with an equatorial climate. Hydrol Process 21(4): 486-499.

Tang W, Yang K, He J, Qin J (2010) Quality control and estimation of global solar radiation in China. Sol. Energy 84(3): 466-475.

Tezel G, Buyukyildiz M (2016) Monthly evaporation forecasting using artificial neural networks and support vector machines. Theor Appl Climatol 124(1-2):1-12.

Vapnik V (1995) The Nature of Statistical Learning Theory.Springer, New York.

Vapnik V (1998) Statistical Learning Theory. Wiley, New York..

Vapnik V, Golowich SE, Smola AJ (1996) Support vector method for function approximation, regression estimation and signal processing, Advanced Neural Information Processing System 9: 281-287.

Wang L, Kisi O, Hu B, Bilal M, Zounemat-Kermani M, Li H (2017c) Evaporation modelling using different machine learning techniques. Int J Climatol 37:1076-1092.

Wang L, Kisi O, Zounemat-Kermani M, Li H (2017a) Pan evaporation modeling using six different heuristic computing methods in different climates of China. J Hydrol 544: 407- 427

Wang L, Niu Z, Kisi O, Li C, Yu D (2017b) Pan evaporation modeling using four different heuristic approaches. Comput Electron in Agr 140:203-213.

Xu XB, Yan T, Yang GS (2013) Environmental impact assessments of the Three Gorges Project in China: Issues and interventions. Earth-SciRev 124: 115-125.

Yang H, Xie P, Ni LY, Flower RJ (2012) Pollution in the Yangtze. Science 2012 (337): 410-410.

Yang H, Yang D (2012) Climatic factors influencing changing pan evaporation across China from 1961 to 2001. J Hydrol 414(2):184-193.

Yang Y, Wang D, Lv W, Mo Y, Ding L (2009) Solar radiation standard and its values transfer system in China.www.wmo.int/pages/prog/www/IMOP/publications/IOM-96 _TECO-2008/P1 (52)_Yang_China.pdf

Yao Y, Ling R, Li H, Zhou WJ, Yang D, Yu JH (2013) Changes of meteorological parameters and lightning current during water impounded in Three Gorges area. Atmos Res 134(12):150-160.

Yu TF, Si JH, Feng Q, Xi HY, Chu YW, Li K (2017) Simulation of Pan Evaporation and Application to Estimate the Evaporation of Juyan Lake, Northwest China under a Hyper-Arid Climate. Water 9: 952; doi:10.3390/w9120952

Table 1 Detail information of the studied meteorological sites

| Site ID | Site name | Latitude | Longitude | Altitude (m) | Annual pan evaporation (mm) |
|---|---|---|---|---|---|
| 1 | Chongqing | 29.58 | 106.47 | 259.1 | 1028.57 |
| 2 | Fengdu | 29.85 | 107.73 | 290.5 | 1165.61 |
| 3 | Wanzhou | 30.77 | 108.4 | 186.7 | 1054.40 |
| 4 | Fengjie | 31.05 | 109.5 | 607.3 | 1347.52 |
| 5 | Badong | 31.03 | 110.37 | 334 | 1655.02 |
| 6 | Yichang | 30.7 | 111.30 | 133.1 | 1337.18 |

Table 2 The empirical coefficients of the developed model

| Site ID | Site name | a | b | c | d | e | f | g | h |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Chongqing | 4.041 | 5.078 | 0.736 | 55.379 | -2.137 | -2.630 | -375.541 | 218.347 |
| 2 | Fengdu | 5.371 | 1.753 | 12.738 | 36.926 | 7.570 | 0.616 | -359.674 | -1014.673 |
| 3 | Wanzhou | 1.707 | 7.797 | 5.666 | 60.134 | 10.127 | -2.391 | -447.337 | -219.955 |
| 4 | Fengjie | 3.858 | 2.211 | -6.069 | 26.917 | 4.545 | 2.259 | -295.637 | 765.515 |
| 5 | Badong | 2.580 | 4.992 | -10.505 | 8.305 | 14.540 | 4.408 | -303.569 | 1214.355 |
| 6 | Yichang | 3.096 | 6.134 | -13.285 | 19.010 | 1.509 | 1.167 | -199.656 | 1481.532 |

Table 3 Error indicators of the developed models

| Site ID | Site name | Emperical model | | $SVM_{pol}$ | | $SVM_{rbf}$ | |
|---|---|---|---|---|---|---|---|
| | | RMSE | RRMSE | RMSE | RRMSE | RMSE | RRMSE |
| 1 | Chongqing | 10.073 | 11.77% | 4.808 | 5.62% | 4.820 | 5.63% |
| 2 | Fengdu | 21.535 | 22.22% | 13.844 | 14.28% | 14.333 | 14.79% |
| 3 | Wanzhou | 12.371 | 14.16% | 7.315 | 8.37% | 7.955 | 9.11% |
| 4 | Fengjie | 13.634 | 12.19% | 7.098 | 6.35% | 7.863 | 7.03% |
| 5 | Badong | 13.390 | 9.73% | 7.848 | 5.71% | 7.892 | 5.74% |
| 6 | Yichang | 11.735 | 10.56% | 6.893 | 6.20% | 6.967 | 6.27% |
| 7 | Average | 13.790 | 13.44% | 7.967 | 7.75% | 8.305 | 8.09% |

Table 4 Correlation coefficients between RMSE of SVM and meteorological variables

| Site | Solar radiation | Sunshine duration | Atmospheric pressure | Wind velocity | Air temperature | Water vapor | Relative humidity |
|---|---|---|---|---|---|---|---|
| Chongqing | 0.828* | 0.794* | -0.825* | 0.643* | 0.775* | 0.796* | -0.540 |
| Fengdu | 0.814* | 0.834** | -0.791* | 0.708* | 0.906** | 0.902** | -0.593 |
| Wanzhou | 0.799* | 0.799* | -0.769* | 0.704* | 0.760* | 0.759* | -0.477 |
| Fengjie | 0.834** | 0.713* | -0.906** | 0.680* | 0.823* | 0.788* | -0.032 |
| Badong | 0.861** | 0.814 | -0.846** | 0.072 | 0.750* | 0.741* | -0.174 |
| Yichang | 0.025 | 0.177 | -0.073 | -0.259 | 0.234 | 0.224 | 0.218 |

* Significant at 0.05 significance level. ** Significant at 0.01 significance level.

Table 5 Correlation coefficients between RRMSE of SVM and meteorological variables

| Site | Solar radiation | Sunshine duration | Atmospheric pressure | Wind velocity | Air temperature | Water vapor | Relative humidity |
|---|---|---|---|---|---|---|---|
| Chongqing | -0.667* | -0.678* | 0.656* | -0.746* | -0.685* | -0.627* | 0.472 |
| Fengdu | -0.593 | -0.569 | 0.571 | -0.652* | -0.385 | -0.340 | 0.697* |
| Wanzhou | -0.673* | -0.660* | 0.675* | -0.630* | -0.655* | -0.606* | 0.650* |
| Fengjie | -0.853** | -0.878** | 0.750* | -0.665* | -0.858** | -0.787* | 0.285 |
| Badong | -0.524 | -0.571 | 0.501 | 0.589 | -0.644* | -0.622* | -0.652* |
| Yichang | -0.919** | -0.814* | 0.894** | -0.758* | -0.851** | -0.808* | -0.353 |

* Significant at 0.05 significance level. ** Significant at 0.01 significance level.

Table 6 RMSE of SVM model developed at one site for estimation of pan evaporation at other sites[a]

| Site | Chongqing | Fengdu | Wanzhou | Fengjie | Badong | Yichang |
|---|---|---|---|---|---|---|
| Chongqing | 4.808 | 7.740 | 8.303 | 7.123 | 10.001 | 9.862 |
| Fengdu | 15.704 | 13.844 | 15.642 | 17.844 | 16.881 | 17.465 |
| Wanzhou | 7.661 | 7.321 | 7.315 | 7.900 | 7.726 | 7.901 |
| Fengjie | 8.684 | 8.635 | 13.864 | 7.098 | 11.633 | 19.256 |
| Badong | 11.683 | 14.097 | 11.276 | 12.792 | 7.848 | 13.921 |
| Yichang | 13.535 | 11.426 | 14.472 | 12.290 | 11.910 | 6.893 |
| Average | 10.346 | 10.511 | 11.812 | 10.841 | 11.000 | 12.550 |

[a] The RMSE represent for the accuracy of SVM developed using the data from the sites in the first row for estimation of pan evaporation at the sites in first column. The highlighted cells indicate that the SVM model developed using the data at row of the highlighted cells can be used to estimate the pan evaporation at the column of the highlighted cells.

Table 7 RRMSE of SVM model developed at one site for estimation of pan evaporation at other sites[a]

| Site | Chongqing | Fengdu | Wanzhou | Fengjie | Badong | Yichang |
|---|---|---|---|---|---|---|
| Chongqing | 5.62% | 9.05% | 9.70% | 8.33% | 11.69% | 11.53% |
| Fengdu | 16.20% | 14.28% | 16.14% | 18.41% | 17.42% | 18.02% |
| Wanzhou | 8.77% | 8.38% | 8.37% | 9.04% | 8.85% | 9.04% |
| Fengjie | 7.76% | 7.72% | 12.40% | 6.35% | 10.40% | 17.22% |
| Badong | 8.49% | 10.25% | 8.20% | 9.30% | 5.71% | 10.12% |
| Yichang | 12.18% | 10.28% | 13.02% | 11.06% | 10.71% | 6.20% |
| Average | 9.84% | 9.99% | 11.30% | 10.41% | 10.80% | 12.02% |

[a] The RRMSE represent for the accuracy of SVM developed using the data from the sites in the first row for estimation of pan evaporation at the sites in first column. The highlighted cells indicate that the SVM model developed using the data at row of the highlighted cells can be used to estimate the pan evaporation at the column of the highlighted cells.

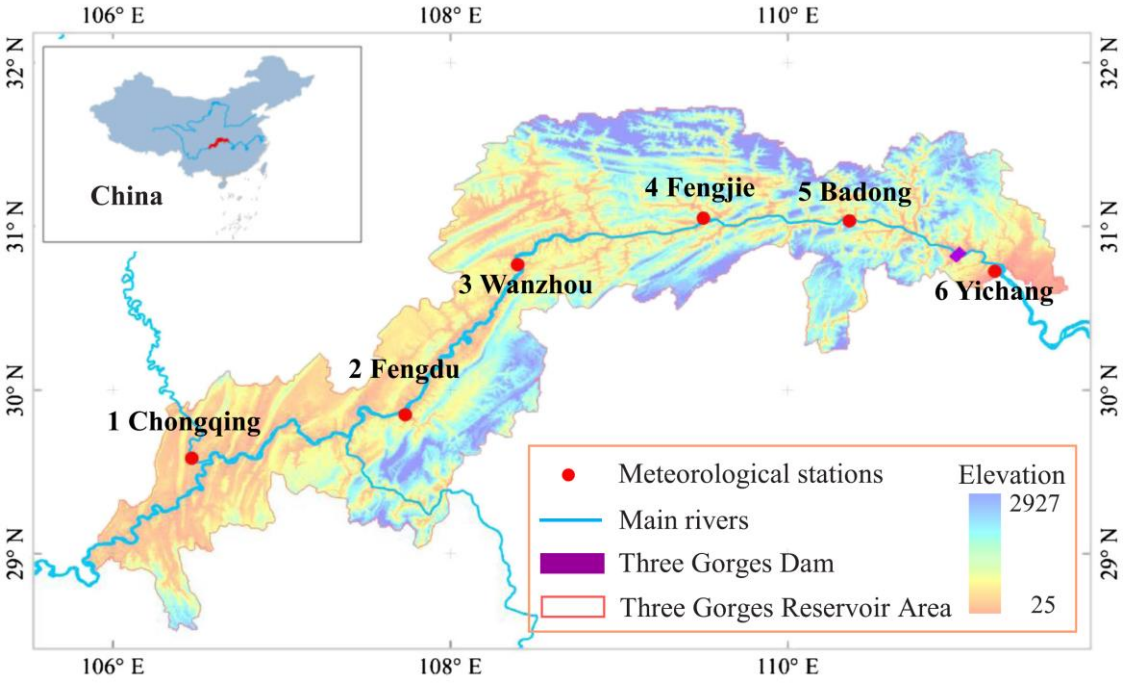


Figure 1 Location of Three Gorges Reservoir Area and the studied meteorological sites

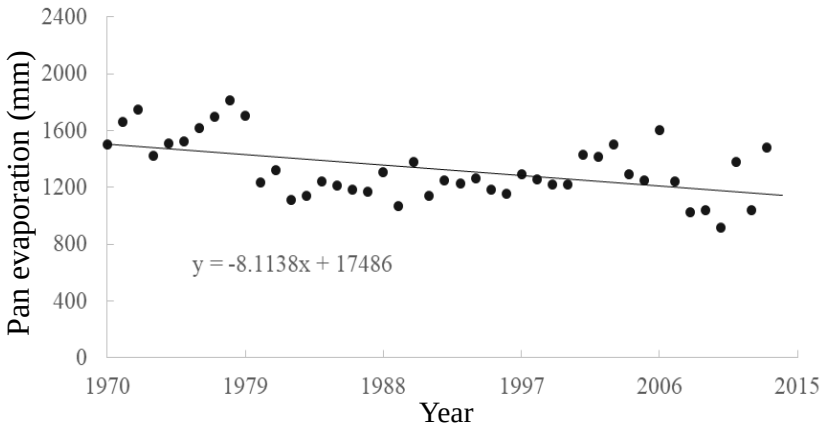


Figure 2 Interannual variation of pan evaporation in Three Gorges Reservoir Area

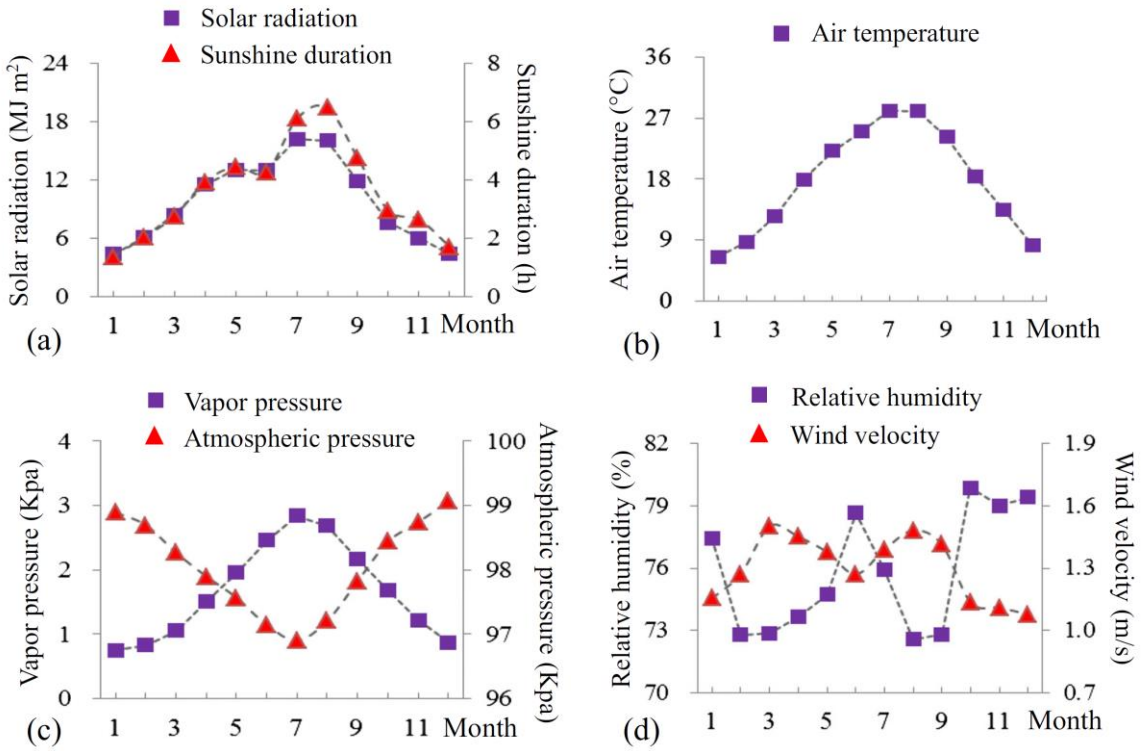


Figure 3 Temporal variations of the meteorological variables in Three Gorges Reservoir Area

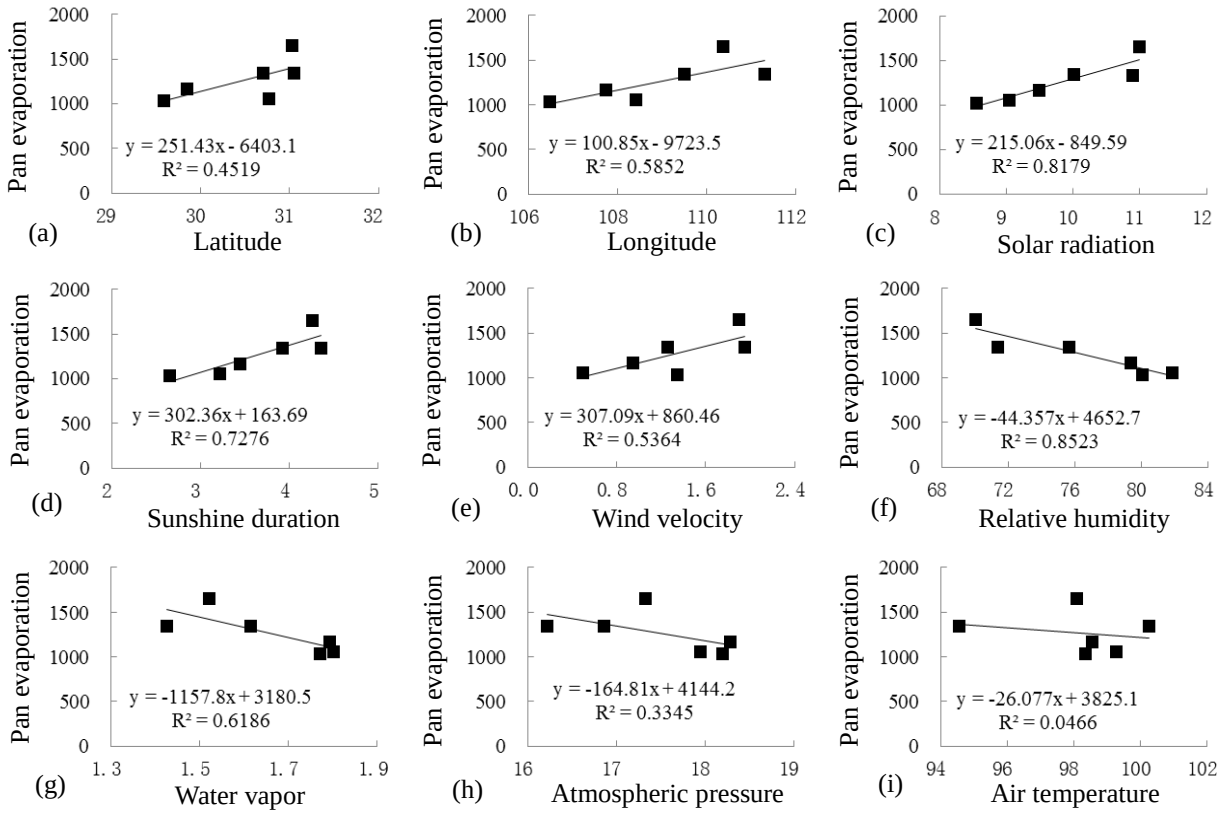


Figure 4 Relationships between pan evaporation (mm) and latitude ('), longitude ('), solar radiation (MJ $m^2$), sunshine duration (h), wind velocity (m/s), relative humidity (%), water vapor (Kpa), atmospheric pressure (Kpa) and air temperature (°C).

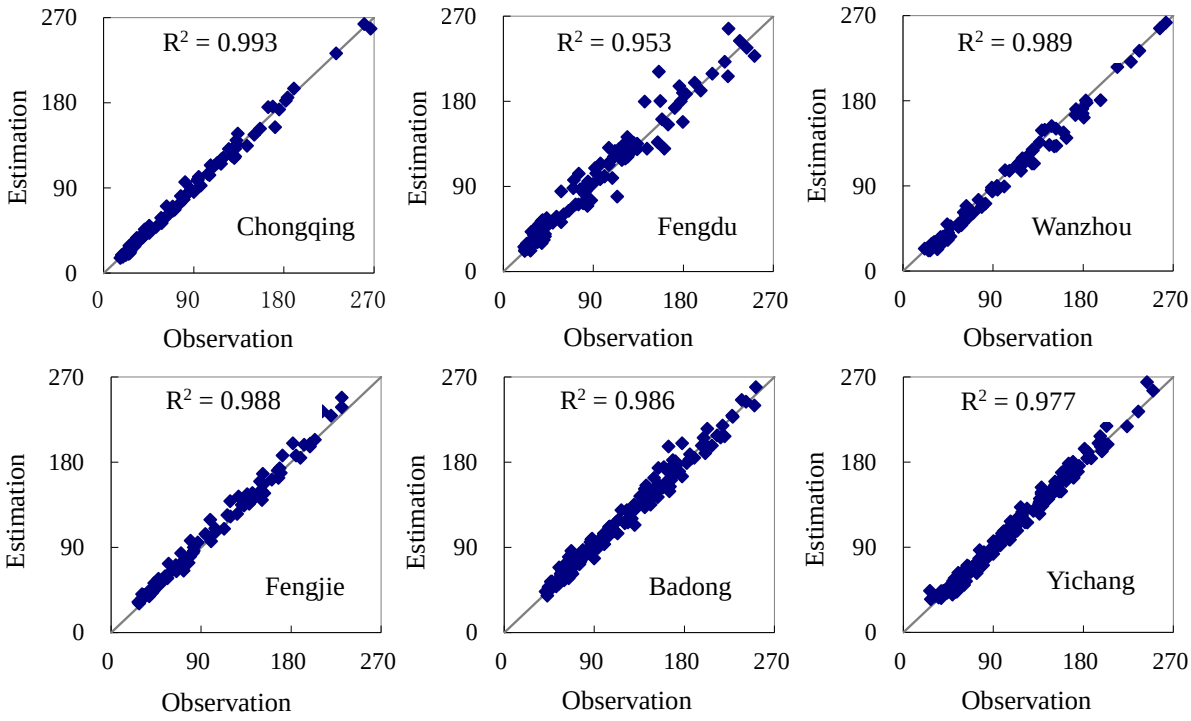


Figure 5 Scatter plots of the observations vs. estimations of pan evaporation (mm) by SVM model with polynomial kernel function

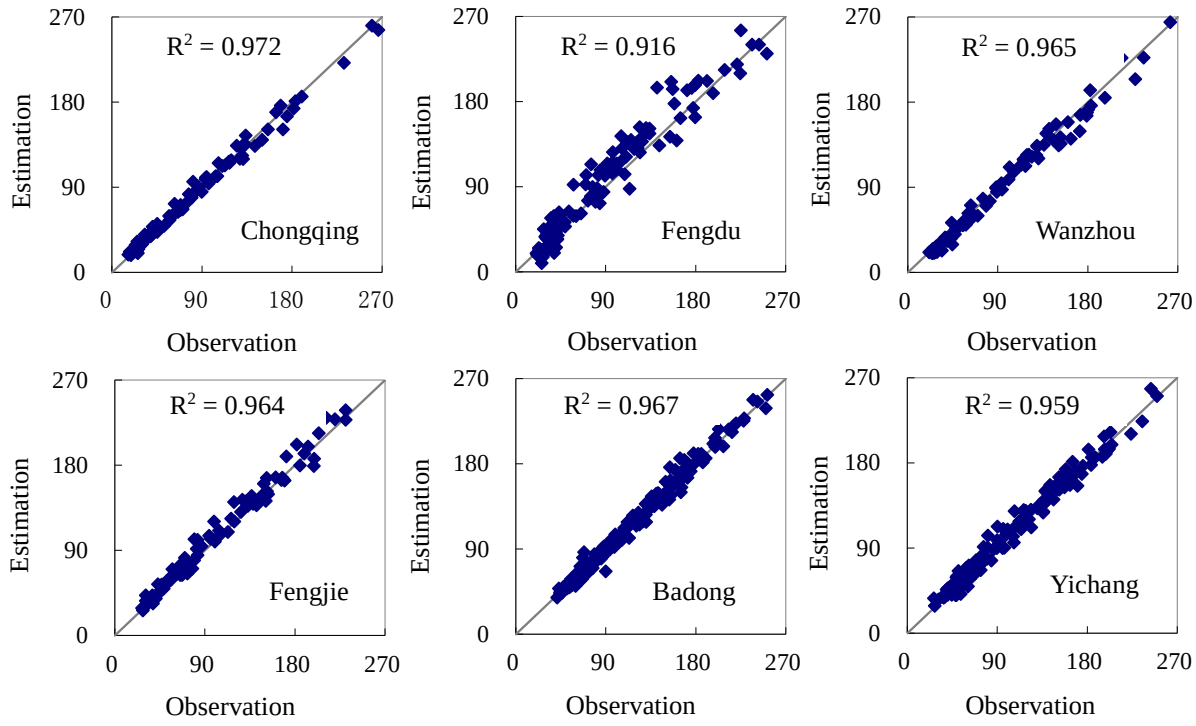


Figure 6 Scatter plots of the observations vs. estimations of pan evaporation (mm) by SVM model with radial basis function

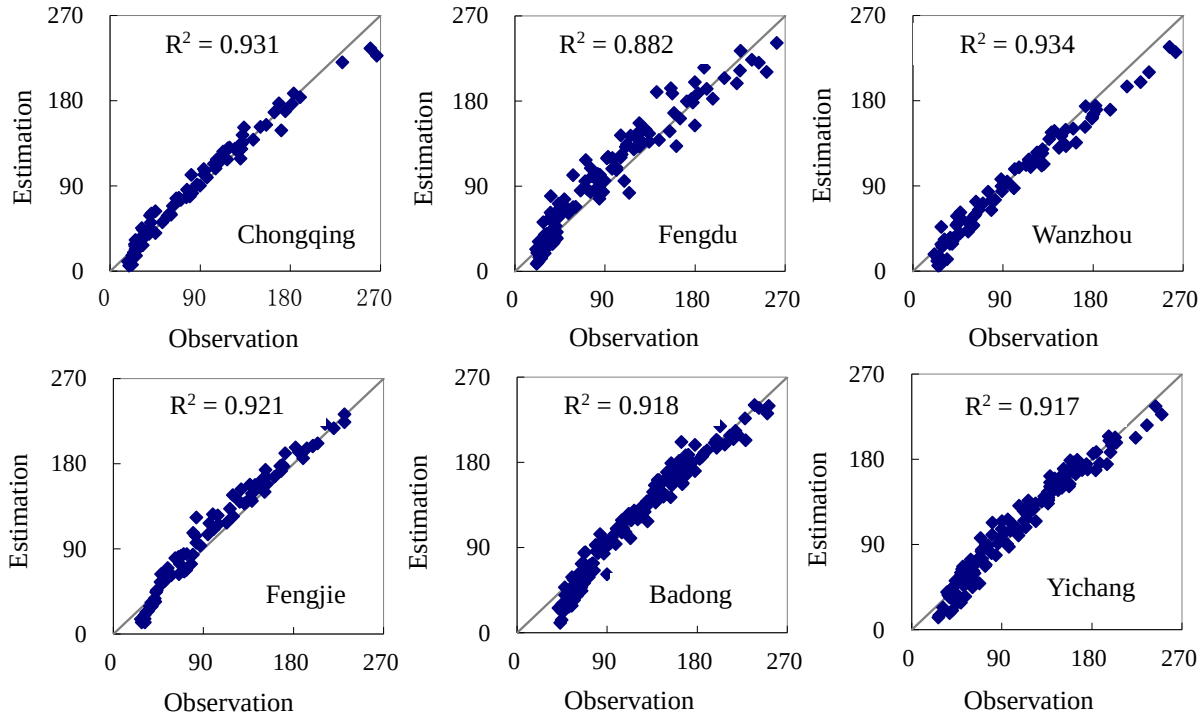


Figure 7 Scatter plots of the observations vs. estimations of pan evaporation (mm) by emperical model

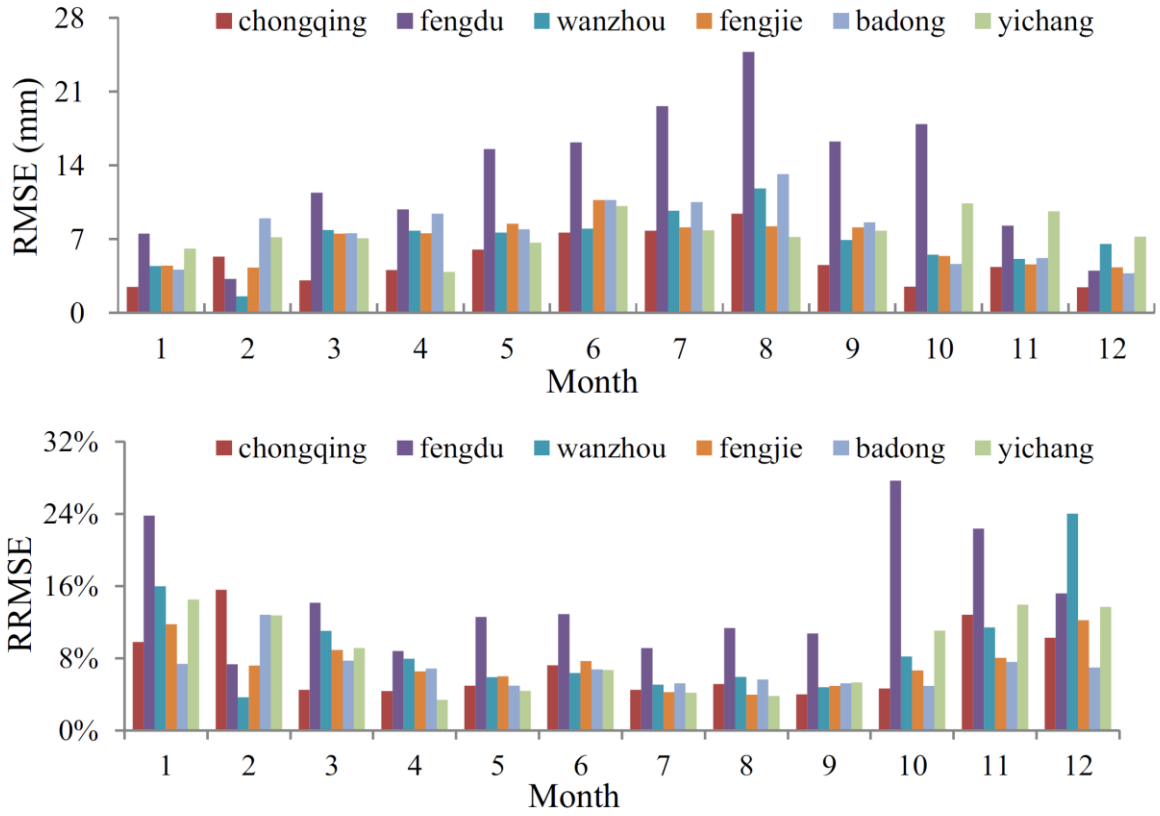


Figure 8 Error variation of the SVM model with polynomial kernel function