Machine learning models for the estimation of monthly mean daily reference evapotranspiration based on cross-station and synthetic data
Lifeng Wu, Youwen Peng, Junliang Fan and Yicheng Wang

ABSTRACT

The estimation of reference evapotranspiration (ET₀) is important in hydrology research, irrigation scheduling design and water resources management. This study explored the capability of eight machine learning models, i.e., Artificial Neuron Network (ANN), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Extreme Gradient Boosting (XGBoost), Multivariate Adaptive Regression Spline (MARS), Support Vector Machine (SVM), Extreme Learning Machine and a novel Kernel-based Nonlinear Extension of Arps Decline (KNEA) Model, for modeling monthly mean daily ET₀ using only temperature data from local or cross stations. These machine learning models were also compared with the temperature-based Hargreaves–Samani equation. The results indicated that the estimation accuracy of these machine learning models differed in various scenarios. The tree-based models (RF, GBDT and XGBoost) exhibited higher estimation accuracy than the other models in the local application. When the station has only temperature data, the MARS and SVM models were slightly superior to the other models, while the ANN and HS models performed worse than the others. When there was no temperature data at the target station and the data from adjacent stations were used instead, MARS, SVM and KNEA were the suitable models. The results can provide a solution for ET₀ estimation in the absence of complete meteorological data.

Key words | cross-station, Kernel-based nonlinear extension of arps decline model, machine learning, reference evapotranspiration

INTRODUCTION

Evapotranspiration (ET) is the combination of two separate water loss processes: water evaporation from the soil and plant surfaces and plant transpiration by which water escapes from a plant’s body to the ambient air in the form of steam through its stomata (Ali Ghorbani et al. 2018; Moaenzzadeh et al. 2018). ET is one of the important components of hydrologic cycle. Reliable estimation of ET is the basis of developing precision irrigation system and improving water use efficiency. Although ET can be measured using eddy covariance, Bowen ratio system or lysimeters, their common problem is that they are expensive, time-consuming and require high professional knowledge, particularly in developing countries such as China. People usually use indirect methods to derive ET, that is, to use reference crop evapotranspiration (ET₀) and crop coefficient (K_c). According to the Food and Agriculture Organization of the United Nations (FAO) publication by Allen et al. (1998), ET₀ represents ‘the evapotranspiration from an actively growing virtual vegetated surface that is 0.12 m tall, completely shading the ground, with adequate water supply, and for daily time-step the aerodynamic resistance is 208/u² (wind speed at 2 meter) surface albedo is 0.23 and a bulk canopy
resistance is 70 s m\(^{-1}\). The most widely accepted methodology for ET\(_0\) estimation is the FAO 56 Penman–Monteith (PMF 56) formula, and it has been a standard method to test the other methods (Fan et al. 2016).

The main drawback of the PMF 56 formula is that it needs many high-quality meteorological data, e.g. solar radiation or sunshine duration, air temperature, wind speed and relative humidity, whereas these data are often unavailable in developing countries. For instance, the costs of observing solar radiation are very high. There are more than 2,000 meteorological stations in Iran. The results suggested that the HS model has a good performance worldwide, model parameter calibration is a crucial prerequisite for local applications (Samani 2000). Fourteen general parametric models were established based on the geographical, temperature and wind speed information by Martí et al. (2015) in eastern Spain. Feng et al. (2017) calibrated the HS model based on the Bayesian method in the Sichuan basin of Southwest China. Other regions have also reported the calibration of the HS model parameters (Gaviñán et al. 2006; Ravazzani et al. 2011; Shahidian et al. 2013; Heydari & Heydari 2014; Almorox & Grieser 2016; Cobaner et al. 2017; Shirí 2017; Valiantzas 2017).

In recent years, there has been more and more research focusing on the estimation and forecast of natural phenomenon (Yaseen et al. 2018, 2018c; Fan et al. 2018a; Ghorbani et al. 2018a, 2018b; Khosravi et al. 2018; Naganna et al. 2019; Xiao et al. 2019), including ET\(_0\) estimation by using machine learning models, e.g. Artificial Neural Network (ANN), Fuzzy Logic, Gene Expression Programming (GEP), Multivariate Adaptive Regression Splines (MARS), Decision Tree (DT), Random Forests (RFs), Support Vector Machine (SVM), Extreme Learning Machine (ELM) and Adaptive Neuro-fuzzy Inference System (ANFIS). Trajkovic (2005) compared the radial basis function neural network (RBFNN) model and three temperature-based empirical models (PMT, HS and Thornthwaite) for estimating ET\(_0\) at seven weather stations in Serbia. The results showed that the RBFNN model provided better ET\(_0\) estimates than the other models at most stations. Luo et al. (2015) evaluated four ANN models for ET\(_0\) prediction using forecasted temperature data. The results showed that the average values of RMSE ranged 0.87–1.36 mm d\(^{-1}\), and the prediction accuracy of maximum temperature was lower than that of minimum temperature. Yassin et al. (2016) compared the ANN and GEP models for estimating ET\(_0\) in Saudi Arabia. The results indicated that the ANN model performed slightly better than the GEP model under the same input
combination of meteorological data. Feng et al. (2016) compared three machine learning models and several empirical models for the estimation of ET<sub>0</sub> in the humid region of Southwest China, and the ELM and GANN models were recommended. Similar work has also been done in Iran (Mehdizadeh et al. 2017). It was found that the MARS and SVM models offered better ET<sub>0</sub> estimates than the GEP and empirical models. Mattar (2018) developed a GEP model for estimating ET<sub>0</sub> at 32 weather stations in Egypt. It was found that the GEP model had better estimation accuracy than the empirical models. Fan et al. (2018a) evaluated the M5 model tree (M5Tree), Gradient Boosting Decision Tree (GBDT), RF, XGBoost, SVM and ELM models for predicting daily ET<sub>0</sub> in different climates of China. They found that the ELM and SVM models performed slightly better than the XGBoost model in terms of estimation accuracy, while the XGBoost model had much less computational time than the ELM and SVM models.

In addition, machine learning models can be coupled with preprocessing or parameter optimization algorithms, and the hybrid models usually perform better than the traditional machine learning models (Feng et al. 2018; Yaseen et al. 2018a, 2018b; Wu et al. 2019). Tao et al. (2018a, 2018b) developed a coupled model based on the ANFIS model and firefly algorithm (FFA) for estimating ET<sub>0</sub> in Burkina Faso. The new ANFIS–FFA model (R² = 0.97, RMSE = 0.24 mm d<sup>-1</sup> and mean absolute percent error (MAPE) = 0.035) was superior to the ANFIS model (R² = 0.89, RMSE = 0.58 mm d<sup>-1</sup> and MAPE = 0.037). Shiri (2018) introduced a new hybrid model based on the RF model and wavelet transform (WT) to estimate ET<sub>0</sub> using air temperature and wind speed data. The results revealed that the new hybrid model improved the estimation accuracy of the RF model and was superior to the empirical models. The ANN and ELM models coupled with WT have also showed superiority to the ordinary ANN and ELM models (Kisi & Alizamir 2018).

After training and testing by local dataset, the application of machine learning models to other regions with similar climatic conditions may still have great uncertainty (Feng et al. 2019; Huang et al. 2019). To overcome this limitation, many scholars have tested the performance of machine learning models when using exogenous data (Martí et al. 2015; Landeras et al. 2018; Shiri 2019; Shiri et al. 2019). Martí and Gasque (2011) explored the use of continentality index to evaluate the station’s climate characteristics. The object station, which was selected based on this characteristic, was used to develop the ANN model to accomplish cross-station strategy. In addition, another new approach based on the geographical inputs has also been reported by Martí and Zarzo (2012). Karimi et al. (2017) evaluated the performance of the GEP and SVM models for ET<sub>0</sub> estimation in the humid region of South Korea. The model was developed and tested at each location in the first scenario, and the results showed that the machine learning models had superiority to the empirical models. In the second scenario, ET<sub>0</sub> was modeled using data from nearby stations and the generalized heuristic model was developed for the studied stations. They found that both the GEP and SVM models could fulfill these tasks, where the GEP model slightly outperformed the SVM model. Shiri et al. (2014) developed ANFIS models based on weather data from Spain and found that the model could successfully estimate ET<sub>0</sub> in both the arid and humid regions of Iran. Kisi (2016) found that the estimation accuracy of the LSSVM, MARS and M5Tree models in ET<sub>0</sub> modeling differed in various cross-station scenarios. The MARS model outperformed the other models when local input data were not available. However, the M5Tree model performed better than the others when both local input and output data were missing. Feng et al. (2017) applied the RF and GRNN models to estimate ET<sub>0</sub> in both local and cross-station scenarios, and found that both models could estimate ET<sub>0</sub> accurately in the Sichuan Province of China. Sanikhani et al. (2015) evaluated six temperature-based machine learning models (GRNN, RBFNN, ANFIS-GP, ANFIS-SC, GEP and MLP) and the HS model for the estimation of ET<sub>0</sub> at two stations in Turkey. The results indicated that the machine learning models, except the MLP model, were superior to the HS model in the cross-station scenario.

Jiangxi Province is located in South China, which experiences a subtropical monsoon climate. This region is a major producing area for double-cropping rice and citrus fruits in China. Variation in seasonal precipitation distribution is obvious, resulting in the frequent existence of seasonal drought in this region. In 2018, for instance, the region was hit by a severe drought, which affected more than 200,000 ha areas and more than 3 million people, and caused direct economic losses of 240 million US dollars. Therefore, the reliable estimation of ET<sub>0</sub> is of crucial
significance for the rational utilization of agricultural water resources in this region. To the best knowledge of the authors, comprehensive comparison of various types of machine learning models for ET₀ estimation has been very minimal, especially their performances with limited temperature data in local and cross-station applications. Machine learning models have different precision performance in various regions. The most suitable model in Jiangxi Province has not been reported yet and this is the first time to compare various types of models for the estimation of ET₀ in this region. In addition, an improved version of kernel-based learning model, i.e., Kernel-based Nonlinear Extension of Arps decline (KNEA) model (Ma 2019), has been recently developed and successfully applied in many other fields (Ma & Liu 2018a). However, the KNEA model has not yet been tested in ET₀ studies. Therefore, this study aims to evaluate and compare the performance of eight temperature-based machine learning models, i.e. ANN, RF, GBDT, XGBoost, MARS, SVM, ELM and KNEA models for: (1) locally estimating monthly mean daily ET₀ at 15 stations in the Jiangxi Province of China using only temperature data, and compare their performance with the empirical HS model; (2) evaluating the developed models for estimating monthly mean daily ET₀ with data from four stations; and (3) evaluating the model performance for estimating monthly mean daily ET₀ using a new synthetic dataset (local extraterrestrial radiation data and temperature data from other stations).

MATERIALS AND METHODS

Case study and data description

Jiangxi Province, covering an area of $1.67 \times 10^5$ km², is a major producing area for double-cropping paddy in China and yields paddy rice of 20.4 billion kg year⁻¹. The study area has a subtropical humid climate with the mean annual rainfall ranging 1,341–1,943 mm, which is largely influenced by the East Asian monsoon (Fan et al. 2018a). About 15 billion m³ year⁻¹ of water resources have been used for irrigation in this region. However, nearly 2 billion m³ year⁻¹ water shortage exists as a result of unreasonable use of water resources and uneven distribution of seasonal rainfall. In this study, monthly maximum and minimum temperature data and extraterrestrial solar radiation from 15 meteorological stations in Jiangxi Province of China (Figure 1) were selected for testing the machine learning models and the empirical HS model in monthly ET₀ modeling. The meteorological data were examined and shared by the National Meteorological Information Center (NMIC) of China Meteorological Administration. The extraterrestrial solar radiation ($R_{s}$) data were estimated on the basis of geographical, seasonal and solar information (Quej et al. 2017). It can be seen from Table 1 that there was no significant variation in the meteorological variables between the training and testing periods at all stations. In addition, the temperature of Station 58506 was much lower than that of the other stations. However, the average annual ET₀ was slightly lower compared with the others due to higher elevation. The values of meteorological variables of the other stations (except Station 58506) are very similar, indicating that the air temperature and ET₀ data in this area had fewer variations. This makes it possible to develop general models for monthly ET₀ estimation in the whole region.

Machine learning models for estimating reference evapotranspiration

Gradient Boosting Decision Tree

The DT is one of the most widely used classification algorithms, which can be represented as multiple if-else rules. Decision tree is actually a method to divide the space into hyper planes. Each time the space is divided, the current space is divided into two parts, such as the decision tree, which makes each leaf node an intersecting region of space. After getting the above decision tree learning, when entering a classification samples instance for decision-making, we can divide the sample into a leaf node according to the two characteristics of the sample $(x, y)$ values and classification results. This is the decision tree model of the classification process. The learning algorithm of decision tree has many subclasses, among which the ID3 algorithm, C4.5 and M5 model tree are the basic algorithms. The GBDT model is a hybridized algorithm that consists of an ensemble of decision trees. One single decision tree usually causes over-fitting issue, while the GBDT model is able to overcome this problem via integrating many weak decision trees.
Figure 1 | Geographical locations of the 15 weather stations in the Jiangxi Province of China used in this study.
with many leaf nodes. The GBDT model has many merits, such as the capability to identify nonlinear transformations, the capability to deal with a categorical variable, computational robustness and high scalability. GBDT had been used in web search (Mohan et al. 2011), subway ridership (Ding et al. 2016), global solar radiation (Fan et al. 2018b), pan evaporation (Lu et al. 2018) and ET₀ estimation (Fan et al. 2018a). More details can be found in Elith et al. (2008).

**Extreme Gradient Boosting**

The XGBoost model is proposed by Chen & Guestrin (2016), which is an improved version of Gradient Boosting Machines (GBMs) and in particular K Classification and Regression Trees (CART). This model is originated from the idea of ‘boosting’, which integrates all the predictions of a series of ‘weak’ learners to develop a ‘strong’ learner via an additive training process. The XGBoost model is supposed to prevent over-fitting issue and minimize the computational time. This is obtained by simplifying the objective functions that allow combining the predictive and regularization terms, while it maintains an optimal computational efficiency at the same time. Parallel calculations are also automatically executed for the functions in the XGBoost model in the training stage. More information about the XGBoost model refer to Chen & Guestrin (2016).

**Kernel-based Nonlinear Extension of Arps Decline Model**

KNEA is a newly nonlinear model initially proposed by Ma & Liu (2018a, 2018b) based on the Arps decline model (Ma et al. 2019a, 2019b) and kernel method (Vapnik 2013). KNEA can be described as:

\[
 f(x) = af(x - 1) + g(u(x)) + \mu
\]

where \( f(x) \) is the output at this time and \( f(x - 1) \) is the output at the last step time. \( u(x) \) is the factors which have effect on output, \( g(u(x)) \) can be interpreted as the relationship between \( u(x) \) and \( f(x) \). \( \mu \) is a bias. From this model, we can see that the output of this time is the result of joint action between the output from last time-step and the influencing factors are at this time. The nonlinear function \( g \) is hard to determine and can be translated to:

\[
 g(u(x)) = \omega^T \varphi(u(x))
\]
This means mapping the original influence factors into the new space. Formula (2) can be thus written as:

\[ f(x) = a f(x - 1) + \omega^T \varphi(u(x)) + \mu \]  

(3)

Although we still cannot solve Equation (3), we can find a very small value so that the difference between the left and right of the equation is as small as possible:

\[ e_x = f(x) - a f(x - 1) - \omega^T \varphi(u(x)) - \mu \]  

(4)

\[ \min_{a, \omega, e} \zeta(a, \omega, e) = \frac{1}{2} a^2 + \frac{1}{2} \|\omega\|^2 + \frac{\gamma}{2} \sum_{x=2}^{n} e_x^2 \]

s.t. \( f(x) = a f(x - 1) + \omega^T \varphi(u(x)) + \mu + e_x \)

(5)

where \( \gamma \) is called a regularization term, it can controls the smoothness of the model. Like SVM, this optimization problem can be used by the Lagrangian multiplier method:

\[ L(a, \omega, \mu, \lambda) = \zeta(a, \omega, e) \]

\[ - \sum_{x=2}^{n} \lambda_x (b f(x - 1) + \omega^T \varphi(u(x)) + \mu + e_x - f(x)) \]

where \( \lambda_x \) is the Lagrangian multiplier. The KKT conditions for optimality of the Lagrangian multiplier method are the following formulas:

\[ \frac{\partial L}{\partial a} = 0 \quad a = \sum_{x=2}^{n} \lambda_x f(x - 1) \]

\[ \frac{\partial L}{\partial \omega} = 0 \quad \omega = \sum_{x=2}^{n} \lambda_x \varphi(u(x)) \]

\[ \frac{\partial L}{\partial \mu} = 0 \quad \mu = 0 \]

\[ \frac{\partial L}{\partial e_x} = 0 \quad e_x = \lambda_x \gamma^{-1} \]

\[ \frac{\partial L}{\partial \lambda_x} = 0 \quad f(x) = a f(x - 1) + \omega^T \varphi(u(x)) + \mu + e_x \]

\[ \left( \begin{array}{c} 0 \\ 1_{n-1}^T \end{array} \right) \begin{pmatrix} \mu \\ \lambda \end{pmatrix} \left( \begin{array}{c} 0 \\ x_{2n} \end{array} \right) \]

(8)

where

\[ I_{n-1} = [1, 1, \ldots, 1]_n^T, \]

\[ \lambda_{n-1} = [\lambda_1, \lambda_2, \ldots, \lambda_{n-1}]_n^T, \]

\[ x_{2n} = [x(2), x(3), \ldots, x(n)]^T, \]

\[ \Omega_{ij} = \varphi(u(i)) \times \varphi(u(j)) = K(u(i), u(j)), \]

in which \( I_{n-1} \) is \( n-1 \) dimensional identity matrix with all the diagonal elements to be 1 and others to be 0. \( \lambda, \mu \) and \( a \) can be obtained by Equation (9). The \( \Omega_{ij} \) can be employed a kernel function \( K(\cdot, \cdot) \) which satisfies the Mercer’s theorem, and a RBF-type kernel function was selected in this study. More details about KNEA can be found in Ma & Liu (2018a).

\[ K(u(i), u(j)) = \exp \left[ \frac{-(u(i) - u(j))^2}{2\sigma^2} \right] \]  

(9)

In addition, Artificial Neuron Network (ANN), Support Vector Machine (SVM), RF, Multivariate Adaptive Regression Spline (MARS) and Extreme Learning Machine (ELM) were also used in this study, and the details of these models can be found in Friedman (1991), Breiman (2001), Huang et al. (2006) and Vapnik (2013).

**FAO 56 Penman–Monteith**

The PMF 56 equation suggested by Allen et al. (1998) was used to calculate monthly mean daily \( ET_0 \) (mm d\(^{-1}\)) and provide the reference data for testing of the empirical and machine learning models in this paper, which can be calculated as:

\[ ET_0 = \frac{0.408 (R_n - G) + \frac{900}{T_s + 273} U (e_s - e_a)}{\Delta + \gamma (1 + 0.34U)} \]  

(10)

where \( R_n \) is the net radiation at the crop surface, which is usually calculated by global solar radiation \( (R_s) \); \( G \) is the soil heat flux density; \( T_s \) is the mean daily air temperature at 2 m height, calculated as the mean of maximum \( (T_{\text{max}}) \) and minimum \( (T_{\text{min}}) \) air temperature; \( U \) is the wind speed at 2 m height; \( e_s \) and \( e_a \) are the saturation and actual vapor pressure; \( \Delta \) is the slope of vapor pressure curve; \( \gamma \) is the air psychrometric constant. As in daily time-step in this study, \( G \) can be
neglected. The details of the PMF 56 equation can be found in Allen et al. (1998).

### Hargreaves–Samani model

The temperature-based HS equation proposed by Hargreaves & Samani (1985) was used to estimate monthly mean daily $\text{ET}_0$ when only air temperature data are available:

$$\text{ET}_0 - \text{HS} = 0.0023R_a(T_a + 17.8)\sqrt{T_{\text{max}} - T_{\text{min}}} \quad (11)$$

As mentioned above, this model will cause underestimation or overestimation of $\text{ET}_0$ without parameter calibration. In this study, the model can be written as follows:

$$\text{ET}_0 - \text{HS} = aR_a(T_a + b)(T_{\text{max}} - T_{\text{min}})^c \quad (12)$$

where $a$, $b$ and $c$ are empirical coefficients. In this study, the PMF 56 equation with air temperature data only (PMT) has been tested for the estimation of monthly mean daily $\text{ET}_0$. However, it has been found to be inferior to the HS equation.

### Model scenarios

In the field of agricultural irrigation management, it is of great significance for decision-makers and planners to obtain the information of $\text{ET}_0$. In this study, eight machine learning models as well as the HS empirical model were developed and applied by using the temperature-based general model for the estimation of monthly mean daily $\text{ET}_0$ in the Poyang Lake Region of Jiangxi Province. The obtained results of the machine learning models were also compared with those estimated by the standard PMF 56 equation. Firstly, a general model for estimating $\text{ET}_0$ was established using data during 2001–2010 from 11 meteorological stations in the Poyang Lake Region (Figure 2). Secondly, the established model was tested in three cases: (1) comparing the eight machine learning models and the HS empirical model for the estimation of monthly mean daily $\text{ET}_0$ of the 11 stations using data from 2011 to 2015; (2) investigating the same predictive model and comparing their performance with the HS model based on input and output data from the other four stations (ID: 57896, 58509, 58608 and 58715) in the same region; and (3) investigating the same predictive model and comparing their performance with the HS model based on temperature data from the four neighboring stations (ID: 577793, 58606, 58813 and 58527) of the four target stations (ID: 57896, 58509, 58608 and 58715) and extraterrestrial radiation data from the four target stations (ID: 57896, 58509, 58608 and 58715), respectively. The second and third cases will be useful for regions lack of temperature data or with no local data at all. The coefficients of the empirical models were attained by the least-squares fitting method, while the parameters of machine learning models were optimized by the grid search technique.

### Statistical indicators

Four commonly used comparison statistics were employed to evaluate the proposed models in this study, including RMSE (Huffman 1997), $R^2$ (Hsu & Chen 1996), MBE and NRMSE (Fan et al. 2019a). RMSE and NRMSE can reflect the overall estimation accuracy of the models. $R^2$ indicates how much percentage of the data can be interpreted by the model, but there may also be a tendency to overestimate or underestimate. MBE is an indicator that reflects the overall overestimation or underestimation of the model. The four statistical indicators can be expressed as follows:

$$R^2 = \frac{\sum_{i=1}^{n} (Y_{i,m} - Y_{i,e})(Y_{i,e} - \bar{Y}_{i,e})^2}{\sum_{i=1}^{n} (Y_{i,m} - \bar{Y}_{i,m})^2 \sum_{i=1}^{n} (Y_{i,e} - \bar{Y}_{i,e})^2} \quad (13)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_{i,m} - Y_{i,e})^2} \quad (14)$$

$$\text{MBE} = \frac{1}{n} \sum_{i=1}^{n} (Y_{i,m} - Y_{i,e}) \quad (15)$$

$$\text{NRMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_{i,m} - Y_{i,e})^2} / \bar{Y}_{i,m} \times 100\% \quad (16)$$
where \( Y_{i,m}, Y_{i,e}, \bar{Y}_{i,m}, \) and \( \bar{Y}_{i,e} \) are the measured, estimated, mean of measured and mean of estimated ET\(_0\) by the PMF 56 model, respectively; \( n \) is the number of observations. Higher \( R^2 \) values indicate high simulation accuracy, whereas the lower absolute values of RMSE, MBE and NRMSE suggest better model performance. Considering the requirements of the MLP and KNEA models, the raw meteorological data were normalized between 0 and 1 as follows:

\[
z_n = \frac{z_i - z_{\text{min}}}{z_{\text{max}} - z_{\text{min}}} \tag{17}\]

where \( z_n \) and \( z_i \) represent the moralized and raw training and testing data; \( z_{\text{max}} \) and \( z_{\text{min}} \) are the minimum and maximum of the training and testing data.

**RESULTS AND DISCUSSION**

**Case 1**

The comparison of the eight machine learning models as well as the HS model for estimating monthly mean daily ET\(_0\) at the 11 stations in the Poyang Lake Region was performed. The statistical summary during training and testing are presented in Table 2. In general, MBE values were less than 0.05 mm d\(^{-1}\) during the training and testing periods. It means that there is no overall overestimation or underestimation by all the machine learning and empirical models. The tree-based models (RF, GBDT and XGBoost) had higher estimation accuracy during the testing stage. RMSE values of the RF, GBDT and XGBoost models were 0.276, 0.281 and 0.269 mm d\(^{-1}\) during testing. NRMSE
values of the RF, GBDT and XGBoost models were 0.116, 0.119 and 0.113 during the testing stage. The RMSE values of the kernel-based models (ELM, KNEA and SVM) and the MARS model were close to each other, which were 2.5–9.3% higher than those of the tree-based models. The ANN model performed worst among machine learning models during the testing period. Compared with the XGBoost model, RMSE was increased by 23% during testing. However, the accuracy of the ANN model was significantly higher than that of the HS model, with the RMSE and NRMSE values of 0.446 mm d\(^{-1}\) and 0.188 during testing, respectively. It is clear that the worst model (HS) can still produce results that are suitable for estimating monthly mean daily ET\(_0\) in this region. Overall, high estimation accuracy can be obtained by established models using only monthly mean daily maximum and minimum temperatures. This is because the global solar radiation in this area has a good relationship with daily maximum and minimum temperatures (Fan et al. 2012). Also, the relative humidity is very high over the year and the influence of wind is not as obvious as that in the arid areas. Thus, the information most closely related to ET\(_0\) can be described by temperature data alone. Similar results were also revealed in Southern China (Feng et al. 2017).

Figures 3 and 4 display the scatter plots of the PM-56 ET\(_0\) and those estimated by the machine learning models and the HS empirical model during the training and testing periods, respectively. It is clear from the figure that all the nine models had passed the significance test (\(P < 0.0001\)). However, scatter plots of different models showed various distributions. The RF model (\(R^2 = 0.987\)) gave the less discrete points during the training period. The scatter distribution of the ELM (\(R^2 = 0.987\)), GBDT (\(R^2 = 0.987\)) and XGBoost (\(R^2 = 0.987\)) models were very close to each other during training. The KNEA, MARS, SVM and ANN models had more discrete distribution during training and they were close to each other. The HS model showed a serious underestimation when the PMF56 ET\(_0\) > 5 mm d\(^{-1}\) during the training period. The ANN model produced more scattered estimates than the other machine learning models and the other eight machine learning models had a similar distribution of scatter points during testing. The scatter distribution was similar during testing and training for the HS model. This was not because the extreme values of the data were different between the two periods, but because the model itself did not capture useful information of temperature. In other words, diurnal temperature range and average temperature were not enough to describe the complex nonlinear relationship between temperature and ET\(_0\).

To evaluate the balance of different machine models and the empirical model, the radar chart of RMSE at the 11 stations is presented in Figure 5. It can be clearly seen that the RMSE of the HS model was higher than the machine learning models at each site. The main reason for the low accuracy of the ANN model was that the stations in the south and west had larger errors than the other

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE (mm d(^{-1}))</td>
<td>(R^2)</td>
</tr>
<tr>
<td>ELM</td>
<td>0.232</td>
<td>0.956</td>
</tr>
<tr>
<td>GBDT</td>
<td>0.233</td>
<td>0.958</td>
</tr>
<tr>
<td>KNEA</td>
<td>0.303</td>
<td>0.927</td>
</tr>
<tr>
<td>MARS</td>
<td>0.308</td>
<td>0.924</td>
</tr>
<tr>
<td>ANN</td>
<td>0.328</td>
<td>0.916</td>
</tr>
<tr>
<td>RF</td>
<td><strong>0.150</strong></td>
<td><strong>0.986</strong></td>
</tr>
<tr>
<td>SVM</td>
<td>0.308</td>
<td>0.925</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.199</td>
<td>0.968</td>
</tr>
<tr>
<td>HS</td>
<td>0.418</td>
<td>0.863</td>
</tr>
</tbody>
</table>

Note: The best statistical indicators among the models are marked in bold.
machine learning models. The accuracy of the SVM model was affected by the large errors of the three stations (ID: 57993, 58506 and 58527) in the north. The RF model ranked first at stations 59102 and 58813, but exhibited the moderate performance at the other stations. The GBDT and XGBoost models were very stable at each station and
ranked in the middle position. The above showed that different datasets had different impacts due to the various principles of model construction. However, due to the natural classification ability of tree-based models, different datasets can be converted into different decision trees, so the estimation accuracy of these models was higher.

Figure 4 | Scatter plots of the PMF 56 ET₀ and those estimated by the machine learning models and the HS empirical model during the testing period.
Case 2

The model established in the previous section can be applied in the areas where only temperature observations are available. The application potential of different models in this case was further assessed, in which four stations with independent datasets (ID: 57896, 58519, 58608 and 58715) were used to evaluate the performance of the nine models. To clearly see the rank of the statistical results, the top three models were highlighted with red, green and blue colours. The mean values of the statistical results of the four stations are also listed in Table 3. As seen from the table, different models behaved differently at various stations. Taking Station 57896 as an example, the ranks of the studied models were: ELM > SVM > XGBoost > MARS > KNEA > GBDT > RF > ANN > HS. However, the SVM, XGBoost, MARS and KNEA models were close to each other. The ANN and HS models were worse than the other models, with the increase in RMSE by 10.5–37.1% and 35.4–67.9% at Station 57896 compared with the other models. From the statistical results of the average value of the four stations, the MARS and SVM models performed slightly superior to the other models, while the ANN and HS models performed worse than the others. It can be seen that since the variations of temperatures and ET0 are small at each station, it is feasible to develop general models for the estimation of ET0 in this region.

Figure 6 presents the scatter plots of the PM-56 ET0 and those estimated by the same predictive machine learning models and the HS empirical model at the four stations. It is clear from the figure that all of the nine models had pass the significance test ($P < 0.0001$). All the machine learning models except the ANN model displayed relatively small scatter distribution. The HS model showed underestimation of monthly mean daily ET0 to some extent when ET0 < 1.5 mm d$^{-1}$ or > 5 mm d$^{-1}$.

Case 3

When a site lacks the basic temperature observation data, it can be replaced with temperature data from the other stations, which is commonly referred to as ‘cross-station application’. In this section, it was supposed that there were four stations having no maximum and minimum temperature data, but only calculated extraterrestrial radiation data. The temperature data from the nearest station were used to replace the missing temperature data for each of these stations. In short, temperature data from stations 57896, 58519, 58813 and 58715 were replaced with those from stations 57993, 58527, 58606 and 58608, respectively. The statistical results are shown in Table 4. The GBDT model performed best at Station 57896, with RMSE 11.8–65.8% less than the other models. The KNEA, ELM, MARS, SVM and XGBoost models were close to each
other, while the RF, ANN and HS models were obviously not as good as these five models. However, the MARS, KNEA and SVM models ranked the top three models at Stations 58519, 58608 and 58715. On this basis, the MARS, SVM and KNEA models performed superior to the other models at all the four stations.

Figure 7 presents the scatter plots of the PMF56 ET₀ and those estimated by the same predictive machine learning models and the HS empirical model performed at the other four stations in the cross-station applications. It is clear from the figure that all the nine models have passed the significance test (P < 0.0001). The scatter distribution of each model was not different from the performance of the models in the previous section. This indicates that it is feasible to use the adjacent meteorological data when local data are missing.

Kisi (2016) found that the LSSVM model was superior to the MARS model when local data were available, but the MARS model performed better than the LSSVM model when cross-station data were used. Similar results have been reported by Karimi et al. (2017) who found that the GEP model outperformed the SVM model in cross-station scenarios. In this study, the tree-based models performed better than the other models in local applications, while the MARS, SVM and KNEA models offered better ET₀ estimates than the others in the absence of local temperature data. This can be due to the differences in the dataset and the characteristics of various models. The tree-based models use greedy algorithms to explain every point as far as possible, but the dataset inevitably contain noise, which results in the over-fitting of the model to some extent. In addition, the tree-based models use many

### Table 3

<table>
<thead>
<tr>
<th>ID</th>
<th>Model</th>
<th>RMSE (mm d⁻¹)</th>
<th>R²</th>
<th>NRMSE</th>
<th>MBE (mm d⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>57896</td>
<td>ELM</td>
<td>0.237</td>
<td>0.968</td>
<td>0.093</td>
<td>0.102</td>
</tr>
<tr>
<td></td>
<td>GBDT</td>
<td>0.263</td>
<td>0.962</td>
<td>0.103</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>KNEA</td>
<td>0.246</td>
<td>0.96</td>
<td>0.096</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>MARS</td>
<td>0.246</td>
<td>0.96</td>
<td>0.095</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>0.325</td>
<td>0.945</td>
<td>0.128</td>
<td>0.091</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.294</td>
<td>0.945</td>
<td>0.116</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.244</td>
<td>0.962</td>
<td>0.096</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.244</td>
<td>0.962</td>
<td>0.096</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>HS</td>
<td>0.398</td>
<td>0.891</td>
<td>0.154</td>
<td>0.031</td>
</tr>
<tr>
<td>58519</td>
<td>ELM</td>
<td>0.367</td>
<td>0.906</td>
<td>0.146</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>GBDT</td>
<td>0.314</td>
<td>0.949</td>
<td>0.125</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>KNEA</td>
<td>0.337</td>
<td>0.958</td>
<td>0.133</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>MARS</td>
<td>0.294</td>
<td>0.958</td>
<td>0.116</td>
<td>0.148</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>0.454</td>
<td>0.939</td>
<td>0.173</td>
<td>0.263</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.34</td>
<td>0.931</td>
<td>0.135</td>
<td>0.134</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.332</td>
<td>0.962</td>
<td>0.132</td>
<td>0.215</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.335</td>
<td>0.947</td>
<td>0.133</td>
<td>0.177</td>
</tr>
<tr>
<td></td>
<td>HS</td>
<td>0.512</td>
<td>0.87</td>
<td>0.202</td>
<td>0.264</td>
</tr>
<tr>
<td>58608</td>
<td>ELM</td>
<td>0.282</td>
<td>0.955</td>
<td>0.12</td>
<td>−0.126</td>
</tr>
<tr>
<td></td>
<td>GBDT</td>
<td>0.242</td>
<td>0.968</td>
<td>0.103</td>
<td>−0.119</td>
</tr>
<tr>
<td></td>
<td>KNEA</td>
<td>0.237</td>
<td>0.966</td>
<td>0.099</td>
<td>−0.093</td>
</tr>
<tr>
<td></td>
<td>MARS</td>
<td>0.245</td>
<td>0.968</td>
<td>0.103</td>
<td>−0.127</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>0.257</td>
<td>0.955</td>
<td>0.109</td>
<td>−0.017</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.249</td>
<td>0.964</td>
<td>0.106</td>
<td>−0.12</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.216</td>
<td>0.97</td>
<td>0.092</td>
<td>−0.078</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.236</td>
<td>0.964</td>
<td>0.1</td>
<td>−0.091</td>
</tr>
<tr>
<td></td>
<td>HS</td>
<td>0.381</td>
<td>0.893</td>
<td>0.16</td>
<td>−0.018</td>
</tr>
<tr>
<td>58715</td>
<td>ELM</td>
<td>0.399</td>
<td>0.935</td>
<td>0.15</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>GBDT</td>
<td>0.427</td>
<td>0.951</td>
<td>0.161</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>KNEA</td>
<td>0.423</td>
<td>0.953</td>
<td>0.158</td>
<td>0.265</td>
</tr>
<tr>
<td></td>
<td>MARS</td>
<td>0.392</td>
<td>0.951</td>
<td>0.146</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>ANN</td>
<td>0.507</td>
<td>0.931</td>
<td>0.191</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>RF</td>
<td>0.43</td>
<td>0.953</td>
<td>0.162</td>
<td>0.295</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.417</td>
<td>0.953</td>
<td>0.157</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>XGBoost</td>
<td>0.445</td>
<td>0.955</td>
<td>0.168</td>
<td>0.307</td>
</tr>
<tr>
<td></td>
<td>HS</td>
<td>0.545</td>
<td>0.865</td>
<td>0.203</td>
<td>0.261</td>
</tr>
<tr>
<td>All</td>
<td>ELM</td>
<td>0.321</td>
<td>0.946</td>
<td>0.127</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>GBDT</td>
<td>0.312</td>
<td>0.958</td>
<td>0.123</td>
<td>0.078</td>
</tr>
<tr>
<td></td>
<td>KNEA</td>
<td>0.311</td>
<td>0.959</td>
<td>0.122</td>
<td>0.106</td>
</tr>
</tbody>
</table>

(continued)
weak classifiers to establish a sub-model (one weak classifier) for small samples independently. The weight of this sub-model is much higher than that of other models, which can subtly obtain some useful information when localizing the model. However, this sub-model model may not be applicable in other regions, and it may also cause the over-fitting of the model when cross-station data are applied. The inspiration of the MARS model also comes...
from the classification tree, but the largest difference between the model and the decision tree is that the base function can be coupled, which has the ability to describe the interaction. This may be the reason why the MARS model is more adaptable. On the other hand, the SVM and KNEA models adopt structural risk minimization and some noise can be artificially ignored by means of tuning parameters, which may explain the high stability of the SVM and KNEA models. Overall, the selection of suitable alternative sites for ET₀ estimation is a systematic project, not only depending on the distance of two sites, but also requiring the similarity of climate rather than the proximity of some individual values. In this study, only four groups of stations (eight stations) were selected to demonstrate the feasibility of switching stations for monthly mean daily ET₀ estimation. However, how to establish a more suitable model still needs to be further explored. In addition, only temperature data were switched in this study and the applicability of using more meteorological data from nearby stations for estimating monthly mean daily ET₀ in a target station is to be studied. Further study is also needed to assess the capability of the proposed models on various time scales (hourly or daily) or in different climatic zones.

**CONCLUSIONS**

This study compared the capability of eight machine learning models, i.e. ELM, GBDT, KNEA, MARS, ANN, RF, SVM and XGBoost, in modeling monthly mean daily ET₀.
using maximum and minimum air temperatures and extra-
terrestrial solar radiation data from 15 stations located in
the Jiangxi Province of China. These machine learning
models were also compared with the empirical Har-
greaves–Samani model. The results showed that the
tree-based models (RF, GBDT and XGBoost) had higher

Figure 7 | Scatter plots of the PMF 56 ET₀ and those estimated by the same predictive machine learning models and the HS empirical model performed at the other four stations in cross-
station applications.
estimation accuracy than the other models in local applications. When only temperature data were available, the MARS and SVM models performed slightly better than the other models, while the ANN and HS models performed worse than the others. When there was no temperature data at the target station and the temperature data from the adjacent station was used instead, the MARS, SVM and KNEA model outperformed the other models. This study can provide a solution for the estimation of $ET_0$ in the Jiangxi Province of China when lack of complete meteorological data and may provide a reference for other regions around the world with similar meteorological conditions.

ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Foundation of China (Nos 51879196, 51790533 and 51709143). Thanks to the National Meteorological Information Center of China Meteorological Administration for offering the meteorological data.

REFERENCES


Ding, C., Wang, D., Ma, X. & Li, H. 2016 Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees. Sustainability 8 (11), 1100.


Downloaded from https://iwaponline.com/hr/article-pdf/50/6/1730/635896/nh0501730.pdf by guest


