Agricultural and Forest Meteorology 316 (2022) 108864

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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

Using support vector machine to deal with the missing of solar radiation data in daily reference evapotranspiration estimation in China

Shang Chen [a,1], Chuan He [b,1], Zhuo Huang [a,c], Xijuan Xu [a,c], Tengcong Jiang [a,c], Zhihao He [a,c], Jiandong Liu [d], Baofeng Su [e,2], Hao Feng [c,f], Qiang Yu [f,g], Jianqiang He [a,c,g,2,*]

[a] Key Laboratory for Agricultural Soil and Water Engineering in Arid Area of Ministry of Education, Northwest A&F University, Yangling 712100, China
[b] PowerChina Beijing Engineering Corporation Limited, Beijing 100024, China
[c] Institute of Water-Saving Agriculture in Arid Areas of China, Northwest A&F University, Yangling, Shaanxi 712100, China
[d] State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing 100081, China
[e] College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China
[f] State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Water and Soil Conservation, Northwest A&F University, Yangling 712100, China
[g] Key Laboratory of Eco-Environment and Meteorology for the Qinling Mountains and Loess Plateau, Shaanxi Meteorological Bureau, Xian, Shaanxi 710014, China

ARTICLE INFO

Keywords:
Reference evapotranspiration
Penman-Monteith equation
Ångström-Prescott formula
Global solar radiation
Machine learning
Support vector machine

ABSTRACT

Accurate estimation of reference evapotranspiration (ET_0) is of great importance for regional water resources planning and irrigation scheduling. The FAO56 recommended Penman-Monteith (P-M) model is widely adopted as the standard method for ET_0 estimation, but its application is usually restricted by limited meteorological data worldwide, especially global solar radiation (R_s). This study provided two possible solutions to deal with the missing R_s data in ET_0 estimation in China mainland. In the first solution, R_s data were estimated with the Ångström-Prescott (A-P) formula and daily sunshine hours. The values of two A-P formula fundamental coefficients a and b were obtained through three ways: (1) estimated based on limited R_s measurements at 80 solar radiation measurement stations (or site-calibrated); (2) recommended by the FAO-56 manual (or FAO-recommended); and (3) estimated based on the altitude and latitude of each weather station through the support vector machine algorithm (or SVM-estimated). The second solution used the SVM algorithm and available weather variables without R_s. The results showed that the FAO-recommended coefficients a and b were separately overestimated and underestimated in China mainland, which generated the largest simulation errors of R_s. However, the transfer errors from R_s estimations to ET_0 estimations were reduced by using the P-M model for all of the three kinds of coefficients. Compared with the R_s-based models, the estimation accuracy of the SVM-ET_0 model yielded the highest accuracy both at the training stage (R^2 = 0.979; $RMSE$ = 0.273 mm d^{-1}) and the testing stage (R^2 = 0.973; $RMSE$ = 0.302 mm d^{-1}). Generally, both the P-M and the machine-learning-based methods could be used for the ET_0 estimation, when only R_s data were missing. However, considering the complexity in the programming, the P-M model combining with the A-P formula with the SVM-estimated A-P coefficients is recommended for daily ET_0 estimation in China.

1. Introduction

In current estimations of reference evapotranspiration (ET_0), the FAO56 recommended Penman-Monteith (P-M) model has been considered as the standard method and widely applied in diverse areas and climates around the world (Allen et al., 1998; Gavilán et al., 2007; López-Urrea et al., 2006). However, multiple weather variables are required when using the Penman-Monteith model to estimate ET_0, including global solar radiation (R_s), wind speed, air temperature, and vapor pressure. Due to the high costs of establishment and maintenance of solar radiation observation equipment, usually only a few weather stations could directly measure solar radiation. The sparse solar

* Corresponding author.
E-mail addresses: bfs@nwsuaf.edu.cn (B. Su), jianqiang_he@nwsuaf.edu.cn (J. He).
[1] These authors contributed equally to this work.
[2] These authors contributed equally as corresponding authors.

https://doi.org/10.1016/j.agrformet.2022.108864
Received 13 April 2021; Received in revised form 17 December 2021; Accepted 4 February 2022
Available online 12 February 2022

radiation observation stations, especially in developing countries, greatly hindered the application of the Penman-Monteith model (Wu et al., 2019). For instance, only about 130 stations have solar radiation measurement among the more than 2000 normal weather stations run by the China Meteorological Administration in China mainland (Fan et al., 2019a). Therefore, the applications of the Penman-Monteith model in larger area have to first deal with the scarcity of solar radiation data in China.

The estimation of R_s has been carried with various methods, including the linear interpolation of available limited R_s observations (Rivington et al., 2006), remote-sensing land surface temperature (Şenkal, 2010), process-based radiation transfer models (Gueymard, 2001), and statistical relations between R_s and temperature or sunshine durations (Bristow and Campbell, 1984; He et al., 2020). However, considering data requirement and computation cost, empirical models based on accessible meteorological variables have still been widely used in R_s estimations, where the involved meteorological variables mainly include sunshine (Ångström, 1924; Prescott, 1940), cloud cover (Iziomon and Mayer, 2002; Sabziparvar and Shetaee, 2007), temperature (Hassan et al., 2016; Richardson, 1981), and precipitation (Adaramola, 2012). Among the different empirical models, the Ångström-Prescott (A-P) formula has been widely adopted to estimate R_s due to its simple linear relationship and limited model inputs. The empirical coefficients a and b of the A-P formula, which separately represent the fraction of extraterrestrial radiation reaching the earth's surface on overcast days and the additional fraction of extraterrestrial radiation reaching the earth's surface on a clear day, must be obtained first before using the formula to estimate R_s.

If there were no enough solar radiation measurements available for the calibration of the A-P formula coefficients, some default values (e.g. $a = 0.25$ and $b = 0.50$) were usually suggested for most areas (Allen et al., 1998). However, previous studies have reported notable spatial variations in a and b coefficients of the A-P formula. Mousavi et al. (2013) calibrated the a and b coefficients 17 solar radiation observation sites in Iran and found that the coefficient a varied from 0.16 to 0.42 (mean value = 0.24) and the coefficient b varied from 0.42 to 0.69 (mean value = 0.53). However, obvious differences in a and b coefficients were found in another study in the same country. Mohammadi et al. (2016) obtained larger a values (mean value > 0.31) and smaller b values (mean value < 0.32) in two coastal cities in south Iran. Adaramola (2012) calibrated the a and b coefficients for eight cities in Nigeria and found that a ranged from 0.17 to 0.31 and b ranged from 0.31 to 0.75.

Great variations in A-P coefficients were also found in China. Liu et al. (2009) calibrated the a and b coefficients with daily R_s measurements in the Yellow River basin in northern China and their calibrated a ranged from 0.11 to 0.29 (mean value = 0.21) and b from 0.50 to 0.69 (mean value = 0.56). They further estimated the a and b coefficients for 80 weather stations with solar radiation measurement in China mainland and found the similar results in which a varied from 0.11 to 0.30 and b varied from 0.48 to 0.69, respectively (Liu et al., 2012). Wu et al. (2011) evaluated R_s estimation with the a and b formula in northeast China based on R_s measurements at 13 different weather stations. Their calibrated coefficients a and b varied from 0.16 to 0.35 and 0.40 to 0.62, respectively. In general, the uncertainties in the a and b coefficients might cause great errors in R_s estimation in a large area. Additionally, if the R_s values, which were estimated with the default a and b coefficients (e.g. $a = 0.25$ and $b = 0.50$) recommended by FAO-56 document, were further used to drive the Penman-Monteith model, error propagation would occur in ET_0 estimations. Therefore, correct re-estimation of the a and b coefficients according to the local specific weather conditions is essential for minimizing the errors in R_s estimations and further in ET_0 estimations that were based on the Penman-Monteith model and estimated R_s data (Sabziparvar et al., 2013; Yin et al., 2008).

The a and b coefficients were usually estimated through linear regressions based on a single factor (Hassan et al., 2016; Rietveld, 1978) or different factor combinations (Jin et al., 2005; Liu et al., 2009). However, traditional empirical regression models were unable to deal with the complex non-linear relationship among variables (Kisi and Parmar, 2016). Beside, the data requirements of empirical models could not always be met in model calibration and validation processes since meteorological data were always incomplete and solar radiation complicatedly distributed in big developing countries such as China (Ming et al., 2015). Recently, machine learning methods have been gradually employed in R_s and ET_0 estimations due to the flexible combination of input variables and high accuracy (Fan et al., 2020, 2019b; Wang et al., 2017). Among the various machine learning methods, support vector machine (SVM), which was developed based on statistical learning theory with the structural risk minimization principle (Vapnik, 1996), has been widely used in ET_0 estimations. Chen et al. (2011) applied SVM and several empirical models to estimate R_s and found the SVM had smaller errors. Wen et al. (2015) compared SVM, artificial neural network (ANN), and three empirical models (Priestley-Taylor, Hargreaves, and Ritchie) in daily ET_0 estimations and found SVM achieved the best performance. He et al. (2020) compared four different machine learning methods in R_s estimations in different climatic regions in China and found that SVM and extreme learning methods achieved more appropriate results.

In this study, we used the SVM method to deal with the missing of R_s measurements in daily ET_0 estimations in China. Two alternative solutions were provided. In the first one, daily R_s data were estimated with the A-P formula based on the differently derived a and b coefficients and then the estimated R_s data were used to estimated daily ET_0 with the Penman-Monteith model. The differently derived a and b coefficients were then evaluated for their accuracies in R_s estimations. In the second solution, the SVM-ET_0 model was established to directly estimate daily ET_0 based on the SVM method and normally accessible weather variables except for R_s. The main objectives were to (1) estimate the a and b coefficients of A-P formula through different ways in China mainland, (2) evaluate the errors in R_s and ET_0 estimations caused by the uncertainties in the a and b coefficients of A-P formula, and (3) explore the possibility of direct daily ET_0 estimation based on machine learning method and common meteorological data but without R_s measurements in China mainland.

2. Materials and methods

2.1. Study area and datasets

According to the climatic regionalization of China by Song et al. (2011), China mainland was divided into four different climatic regions according to local temperature, precipitation, latitude, and longitude (Song et al., 2011; Fig. 1), which were the mountain plateau zone (MPZ), the temperate continental zone (TCZ), the temperate monsoon zone (TMZ), and the subtropical monsoon zone (SMZ). The average elevations of the four climatic zones were 4236 m, 912 m, 288 m, and 611 m, for the MPZ, TCZ, TMZ, and SMZ, respectively. Large variations were found in average annual precipitation among the four climatic regions. The TCZ region, located in northwest China, has a dry climate and an average annual precipitation of 193 mm; the SMZ region is a humid region with an average annual precipitation of 1360 mm; and the MPZ and TMZ regions have average annual precipitations of 460 and 591 mm, respectively.

Two weather datasets were used in this study. This first dataset includes daily measured R_s and other accessible weather variable measurements from 80 solar radiation observation stations in China (Fig. 1 and Table S1). The second dataset only contains normal accessible weather variables that were observed at the rest 839 normal weather stations in China in 1960–2017 (Fig. 1). The accessible weather variables in the two datasets include maximum temperature (T_{max}, °C), minimum temperature (T_{min}, °C), precipitation (P, mm), wind speed (U, m s^{-1}), relative humidity (RH,%), and sunshine hours (n, h). All of the

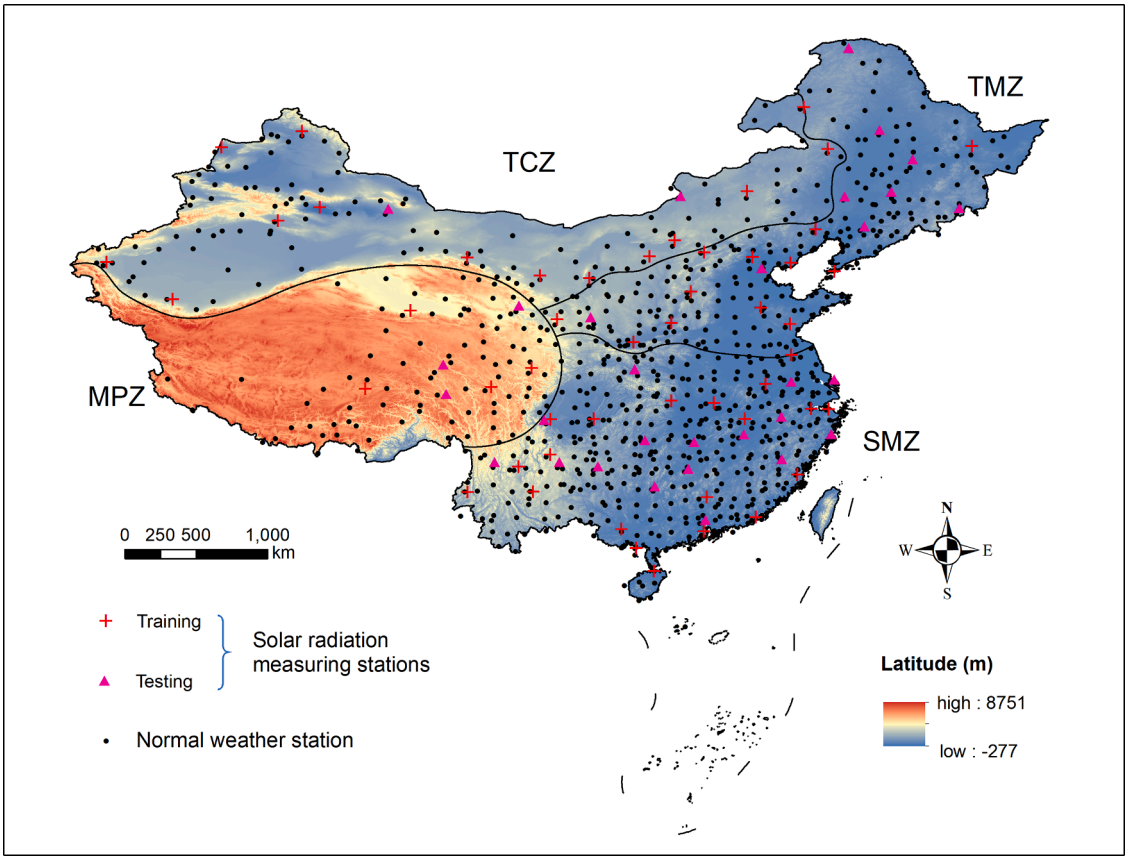

Fig. 1. Distribution of the 80 national meteorological stations with solar radiation observations (crosses for the training stage and tri-angles for the testing stage of the machine learning models) and 839 meteorological stations without solar radiation observations (black dots) but with long-term continuous normal meteorological observations in the mainland of China. The whole China mainland is divided into four different climatic zones: the mountain plateau zone (MPZ), the temperate continental zone (TCZ), the temperate monsoon zone (TMZ), and the subtropical monsoon zone (SMZ). The acronyms are the same below.

weather data involved were obtained from the China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/).

2.2. ET_0 estimation without R_s measurements

In this study, daily ET_0 was supposed to be estimated without readily measured R_s data. Two alternative solutions were established to estimate daily ET_0 with or without R_s input (Fig. 2). In the first or R_s-dependent solution, the indirectly estimated R_s values combined with the other weather variables were used to drive the Penman-Monteith model to estimate daily ET_0. Daily R_s was estimated through the A-P formula based on daily sunshine duration (n), daylength (N) and extraterrestrial radiation (R_a). In the second or R_s-independent solution, the SVM-ET_0 model was established to directly estimate daily ET_0 based on the SVM (support vector machine) machine learning method and several accessible weather variables except for R_s.

2.2.1. Estimation of the a and b coefficients of the Ångström-Prescott formula

Ångström-Prescott (A-P) formula: The original R_s estimation formula was proposed by Ångström (1924), which assumed a linear relation between the ration of R_s/R_{s0} and the ratio of n/N, where R_{s0} is clear-sky solar radiation. N could be estimated with Eq. (1).

$$N = 24 \times \omega_s/\pi \tag{1}$$

where ω_s is the sunset hour angle, rad. Since R_{s0} was difficult to obtain in most areas, Prescott (1940) suggested using R_a (extraterrestrial radiation) to replace R_{s0} and thus obtained the famous Ångström-Prescott (A-P) formula (Eq. (2)).

$$\frac{R_s}{R_a} = a + b\frac{n}{N} \tag{2}$$

where a and b are coefficients that varied between 0 and 1, and the sum of these two coefficients is the clear sky transmissivity Liu et al., 2010); extraterrestrial radiation R_a could be estimated according to the Eqs. (3)–((6).

$$R_a = (24 \times 60/\pi)G_{sc}d_r(\omega_s \sin\varphi \sin\delta + \cos\varphi\cos\delta\sin\omega_s) \tag{3}$$

$$d_r = 1 + 0.033\cos(2\pi \times J/365) \tag{4}$$

$$\delta = 0.409\sin(2\pi \times J/365 - 1.39) \tag{5}$$

$$\omega_s = \arccos(-\tan\varphi\tan\delta) \tag{6}$$

where G_{sc} is the solar constant of 0.082 MJ m^{-2} min^{-1}; d_r is the inverse square of the relative distance earth to sun; φ is the latitude, rad; δ is the solar declination, rad; and J is the day of the year.

FAO recommended a and b coefficients: Therefore, the coefficient a and b must be obtained first before applying the A-P formula to estimate R_s values. Depending on atmospheric conditions (humidity, dust) and solar declination (latitude and month), the coefficients will vary. However, where no actual solar radiation data are available and no calibration has been carried out for improved a and b coefficients. However, various studies have pointed out the importance of re-estimating the two coefficients in areas with significant differences in weather and geographical conditions.

Site-calibrated a and b coefficients: Based on available daily observed n and R_s values from the 80 national meteorological stations with solar

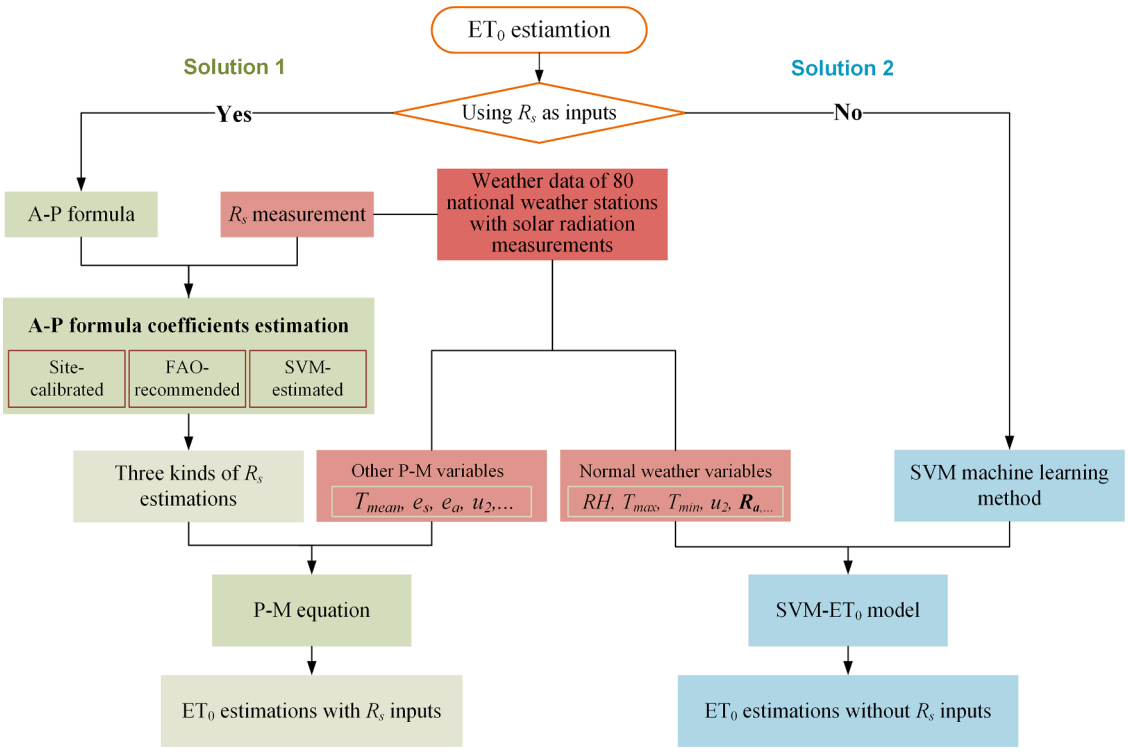

Fig. 2. Flowchart of the two alternative solutions for daily ET_0 estimation under the missing of R_s measurements. The Solution 1 used the Ångström-Prescott (A-P) formula to estimate R_s. Then, the estimated R_s values combined with the other weather variables were used to drive the Penman-Monteith model to estimate daily ET_0. Three kinds of *a* and *b* coefficients of A-P formula were obtained and compared, which were derived based on the site-calibrated, the FAO-recommended and the SVM-estimated methods, respectively. The Solution 2 established the SVM-ET_0 model to directly estimate daily ET_0 through the SVM (support vector machine) method based on common meteorological variables, such as relative humidity (*RH*), daily maximum air temperature (T_{max}), daily maximum air temperature (T_{min}), sunshine duration (*n*), wind speed at 2 m above ground (u_2), and extraterrestrial radiation (R_a).

radiation measurement in China mainland, site specific *a* and *b* coefficients could be calibrated using the least square regression between R_s/R_a and n/N. Some scholars had indicated that the *a* and *b* coefficients were time independent (Almorox and Hontoria, 2004; Liu et al., 2009). The *a* and *b* coefficients estimated at different time scales could not significantly improve R_s estimation accuracy. Thus, we just estimated the time-independent *a* and *b* coefficients in this study. Long-term R_s measurements under various weather conditions were needed to calibrate the A*a* and *b* coefficients. Therefore, the calibrated *a* and *b* coefficients were only available at the limited 80 national weather stations with direct R_s measurements.

SVM-estimated a and b coefficients: The SVM (support vector machine) model established by Vapnik (1996) is a supervised machine learning method for data analysis and pattern recognition (Fig. 3), and it has been widely employed for estimations of R_s (Chen and Li, 2013; Fan et al., 2018a; He et al., 2020). The SVM is based on the principle of structural risk minimization. Therefore, SVM can achieve small confidence interval and has a good generalization for future samples. SVM can more efficiently solve the problems of small samples, nonlinearity, and high dimensionality. In this study, the *'kernlab'* package in R language

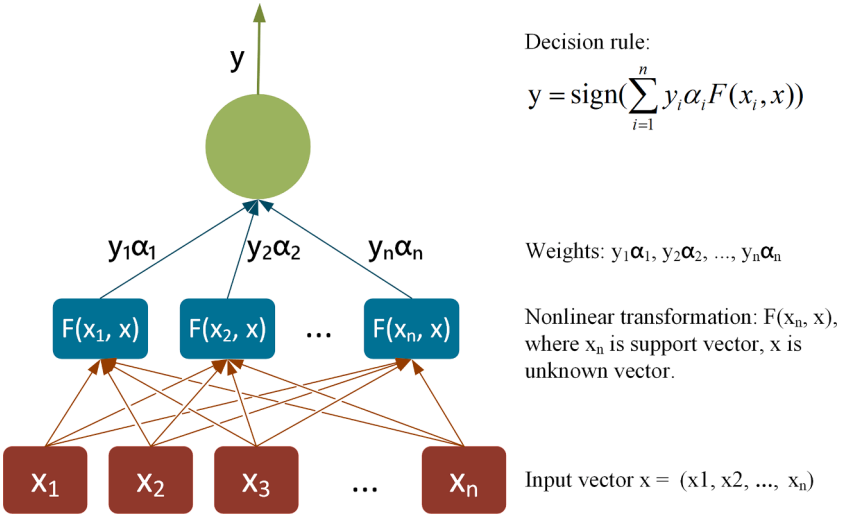

Fig. 3. General structure of the support vector machine (SVM) algorithm.

(Karatzoglou et al., 2004) was used to conduct the SVM-based solar radiation estimation. The main procedures of the estimation of the *a* and *b* coefficients with the SVM method were as follows.

(1) *The optimal combination of input factors for the SVM model.* Sabziparvar et al. (2013) reported a good relationship between the geographic variables (altitude and latitude) and the A-P coefficients. In China, Liu et al. (2014) established the regression functions between the A-P coefficients and altitude, latitude, and percent of sunshine. In this study, the combination of site-specific altitude and latitude was selected as input factors to establish SVM models for the estimation of the coefficient *a* and *b*.

(2) *Normalization of meteorological data at each weather station (**Eq. (7)**).*

$$x_n = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{7}$$

where x_n is the normalized data; x_i is the raw data; x_{max} and x_{min} are the maximum and minimum values of the raw data.

(1) *Determination of the ranges of key parameters in the SVM method.* The ranges of key parameters in SVM method were determined through the '*trial-and-error*' approach and then the optimal parameter values were selected with the grid search approach, while the remaining parameters were set to their default values.

(2) *Train and test the SVM models to estimate the a and b coefficients based on the site-calibrated coefficient values.* The site-calibrated *a* and *b* coefficient values from 50 randomly selected weathers stations with R_s measurements were used to establish the SVM model to estimate the *a* and *b* coefficients. Then, the site-calibrated *a* and *b* coefficient values from the rest 30 stations were used to test the established SVM model. To avoid the over-fitting problem at the training stages, we employed the 10-fold cross-validation method in the establishment of the estimation models.

Estimation of the a and b coefficients for the 839 normal meteorological stations through the SVM model established above. Then, the geographic distributions of the coefficient *a* and *b* were generated through interpolation of 839-site estimations for whole China mainland ((Fig. 3)).

2.2.2. R_s estimation through the A-P formula with three kinds of a and b coefficients

Daily R_s values were then estimated through the A-P formula with the site-calibrated, the FAO recommended, and the SVM-estimated *a* and *b* coefficients at the 50 train stations and 30 test stations, respectively. Then, the R_s values estimated with different A-P coefficients were compared with their corresponding observations so as to determine the accuracy of R_s estimation through the A-P formula with different kinds of *a* and *b* coefficients.

2.2.3. ET_0 estimations through the Penman-Monteith model with R_s input

The standard Penman-Monteith model (or P-M equation, Eq. (8)) has been recommended as a standard method for ET_0 estimation around the world by the Food and Agriculture Organization of the United Nations (FAO) (Allen et al., 1998).

$$ET_0 = \frac{0.408\Delta, R_n - G) + \gamma \frac{900}{T_{mean}+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \tag{8}$$

where R_n is the net radiation above canopy, MJ m^{-2} d^{-1};

G is the soil heat flux density, MJ m^{-2} d^{-1};
γ is the air psychrometric, kPa $^{\circ}C^{-1}$;
T_{mean} is the mean daily air temperature, °C;
u_2 is the wind speed at 2 m above ground, m/s;
e_s and e_a are the saturation and actual vapor pressures, kPa;
Δ is the slope of the vapor pressure curve, kPa $^{\circ}C^{-1}$;
R_n is calculated by the difference between R_{ns} and R_{nl} (Eq. (9)).

$$R_n = R_{ns} - R_{nl} \tag{9}$$

where R_{ns} is the net shortwave radiation (Eq. (10)), MJ m^{-2} d^{-1}; R_{nl} is the net outgoing longwave radiation (Eq. (11)), MJ m^{-2} d^{-1}.

$$R_{ns} = (1 - \alpha)R_s \tag{10}$$

$$R_{nl} = \sigma \left[\frac{T^4_{\max,K} + T^4_{\min,K}}{2}\right](0.34 - 0.14\sqrt{e_a})\left(1.35\frac{R_s}{R_{so}} - 0.35\right) \tag{11}$$

where α is albedo or canopy reflection coefficient, which is 0.23 for the hypothetical grass reference crop; σ is the Stefan-Boltzmann constant, 4.903×10^{-9} MJ K^{-4} m^{-2} day^{-1}; $T_{max,\ K}$ and $T_{min,\ K}$ are the maximum and minimum absolute temperature during the 24 h, respectively; R_{s0} is the clear-sky solar radiation (Eq. (12)), MJ m^{-2} day^{-1}.

$$R_{so} = (a + b)R_a \tag{12}$$

where *a* and *b* are the same coefficients in the A-P formula. In this way, daily ET_0 values can be estimated with the R_s values estimated through the A-P formula with the three kinds of *a* and *b* coefficients mentioned previously.

2.2.4. Direct ET_0 estimation through SVM method without R_s inputs

In this solution, we established the SVM-ET_0 estimation models without using R_s inputs in each climatic zone. Additionally, we also employed three additional machine learning algorithms to establish the ET_0 estimation models to assess the robustness of different machine learning methods. These three algorithms included the BP (back propagation) neural network, the Cubist model tree, and the ELM (extreme learning machine). However, we mainly showed the results of the SVM-ET_0 model for the sake of brevity. The results of ET_0 estimations based on the P-M model and R_s measurements were treated as the actual daily ET_0 values due to the lack of ET_0 observations. According to the number of stations located in each climatic region, the same proportional stations were randomly selected from each region to train (50/80) and to test (30/80) the SVM-ET_0 model. Then, similar meteorological variable combination (except for R_s) used in the P-M model was employed to establish the SVM-ET_0 model. To avoid the over-fitting problem at the training stages, we employed the 10-fold cross-validation method in the establishment of the SVM-ET_0 models. It was noteworthy that sunshine hour and extraterrestrial radiation were directly used in the establishment of SVM-ET_0 models to avoid the possible errors in R_s estimations. Therefore, the final input variables included wind speed, sunshine hour, extraterrestrial radiation, maximum temperature, minimum temperature, and relative humidity. Furthermore, daily ET_0 estimations based on the SVM-ET_0 models were compared with the estimations based on the Penman-Monteith equation at the same sites at both national and site scale. Two representative weather stations were randomly selected from each climatic zone for the comparison: Station 55299 and 56146 for the MPZ region; Station 51828 and 52681 for the TCZ region; Station 57461 and 58238 for the SMZ region; and Station 53963 and 54342 for the TMZ region. Finally, the average daily ET_0 was estimated through the established SVM-ET_0 model for the 839 normal weather stations. The results were further interpolated to generate a national average daily ET_0 distribution through the inverse-distance-weight interpolation method for whole China mainland.

2.3. Statistical indices

Five commonly used statistical indices were used to assess the estimation accuracy of different variables, including coefficient of determination (R^2, Eq. (13)), root mean squared error (*RMSE*, Eq. (14)), normalized root mean square error (*NRMSE*, Eq. (15)), mean absolute error (*MAE*, Eq. (16)), and coefficient of variation (*CV*, Eq. (17)):

$$R^2 = \frac{\left[\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\right]^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2\sum_{i=1}^{n}(Y_i-\overline{Y})^2} \quad (13)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-X_i)^2} \quad (14)$$

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-X_i)^2}}{\overline{X}} \times 100\% \quad (15)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i-X_i| \quad (16)$$

$$\mathrm{CV} = \frac{1}{\overline{X}}\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}} \quad (17)$$

where X_i and Y_i are the observation and estimation values at *i*th time step, respectively; $\overline{X}$ and $\overline{Y}$ are the average value of simulations and estimations, respectively;*n* is the number of samples.

3. Results

3.1. Evaluation of the a and b coefficients of A-P formula derived from different methods

The *a* and *b* coefficient were first calibrated through linear regression for each of the 80 weather stations in the four climatic regions with direct R_s measurement (Table 1). The regional average values of coefficient *a* varied from 0.16 to 0.23, while the values of coefficient *b* varied from 0.54 to 0.59. The CV (coefficient of variation) values of coefficient *a* were greater than coefficient *b* both at regional and national scales, which indicated larger variations in coefficient *a* in large areas.

Table 1
Comparison among the FAO-recommended, the site-calibrated, and the SVM-estimated *a* and *b* coefficients of Ångström-Prescott formula at the 80 national weather stations with solar radiation observation in China mainland.

Method	Region	Elevation(m)	Coefficients of the Ångström-Prescott formula			
			a		b	
			Mean	*CV* [a]	Mean	*CV*
FAO-recommended	Whole China[b]		0.25		0.50	
Site-calibrated	MPZ	4236	0.23	0.13	0.59	0.07
	TCZ	912	0.22	0.15	0.54	0.06
	SMZ	611	0.16	0.10	0.58	0.07
	TMZ	288	0.19	0.16	0.54	0.05
	Whole China		0.19	0.20	0.56	0.07
SVM-estimated	MPZ	4236	0.21	0.03	0.57	0.01
	TCZ	912	0.20	0.02	0.55	0.01
	SMZ	611	0.16	0.01	0.56	0.02
	TMZ	288	0.19	0.10	0.55	0.02
	Whole China		0.19	0.16	0.55	0.02

Notes:
[a] CV = coefficient variation
[b] The whole China mainland is divided into four different climatic zones: the mountain plateau zone (MPZ), the temperate continental zone (TCZ), the temperate monsoon zone (TMZ), and the subtropical monsoon zone (SMZ). The acronyms are the same below.

Compared with the site-calibrated *a* and *b* coefficients in this study, the FAO recommended a relatively larger *a* but a smaller *b* value, especially in the SMZ (subtropical monsoon zone) region where *a* was overestimated by 0.09 and *b* was underestimated by 0.08. Therefore, it was necessary to re-estimate the *a* and *b* coefficients of the A-P formula before applying it in R_s estimations in large area with complex geological and meteorological conditions. Beside, the SVM method estimated much better *a* and *b* coefficients for all of the four climatic zones. Differences between the site-calibrated and SVM-estimated coefficients were less than 0.02 for both coefficients in the four climatic zones. The *CV* values of SVM-estimated *a* and *b* coefficients were less than the site-calibrated *a* and *b* coefficients. Generally, the *CV* value of estimated *b* was less than the *CV* of estimated *a*. The results showed smaller spatial variations in the SVM-estimated *a* and *b* coefficients, particularly in the coefficient *b*.

The SVM-estimated *a* and *b* coefficients were compared with the site-estimated coefficients for the 50 national weather stations with solar radiation measurement in the train stage and 30 weather stations in the test stage in the support vector machine (SVM) algorithm (Fig. 4). Good agreement was found between the site-calibrated and SVM-estimated values of coefficient *a* since most points concentrated to the 1:1 line (Fig. 4a). The ranges of site-calibrated and SVM-estimated values of coefficient *a* were 0.13–0.23 and 0.11–0.29, respectively. The *RMSE* values of train and test stage were 0.028 and 0.021, which showed great accuracy and stability of the SVM method in the estimation of coefficient *a*. However, the estimation errors in coefficient *b* were greater than those in coefficient *a*, especially at the train stage (*RMSE* = 0.040). The range of the SVM-estimated *b* values (0.53–0.58) was narrower than the site-calibrated range (0.48–0.68). The *RMSE* values of all of 80 stations involved were 0.026 for *a* and 0.036 for *b*. In addition, the FAO had recommended a larger *a* but smaller *b* value than the SVM-estimated values.

The site-calibrated *a* and *b* coefficients for the 80 national weather stations with direct solar radiation measurements were then used to establish the SVM model to estimate site-specific *a* and *b* coefficients for the other 839 normal weather stations without direct solar radiation measurements and whole China (Fig. 5). Distribution of coefficient *a* values was obviously affected by geographic conditions (Fig. 5a). Larger site-calibrated and SVM-estimated *a* values were found in northern and western China, while smaller values were found in southern and eastern China (Fig. 5a and b). Compared with the coefficient *a*, distribution of coefficient *b* values was more fragmented and complicated with smaller values at high-latitude sites in northeast and northwest China. Generally, higher *a* and *b* values concentrated in the Qinghai-Tibetan Plateau, which was consistent with the rich solar resources in this area (Fig. 5d and e). The interpolation with the *Kriging* method generated the primary general distributions of *a* and *b* in whole China mainland (Fig. 5c and f), which could provide a useful reference for solar radiation estimation with the Ångström-Prescott formula in China.

3.2. R_s estimation using different A-P coefficients

Estimated R_s through the A-P formula with the SVM-simulated *a* and *b* coefficients were compared with those with the site-calibrated and FAO-recommended values at both the train stage (50 weather stations involved; Table 2) and the test stage (30 weather stations involved; Fig. 6), respectively. In the train stage, the estimated R_s with SVM-simulated coefficients achieved the highest R^2 (0.889) but lowest simulation errors (*RMSE* = 2.534 MJ m^{-2} d^{-1} and *MAE* = 1.853 MJ m^{-2} d^{-1}). The R_s estimation accuracy with SVM-estimated coefficients was better than those with the site-calibrated coefficients. Beside, the largest simulation errors were generated by the FAO-recommended *a* and *b* coefficients. At the test stage, the R_s estimation accuracy with SVM-estimated coefficients was less than those with site-calibrated coefficients. The R^2 values reduced from 0.889 to 0.875 while the *RMSE* values increased from 2.534 to 2.671 MJ m^{-2} d^{-1}. Additionally, the

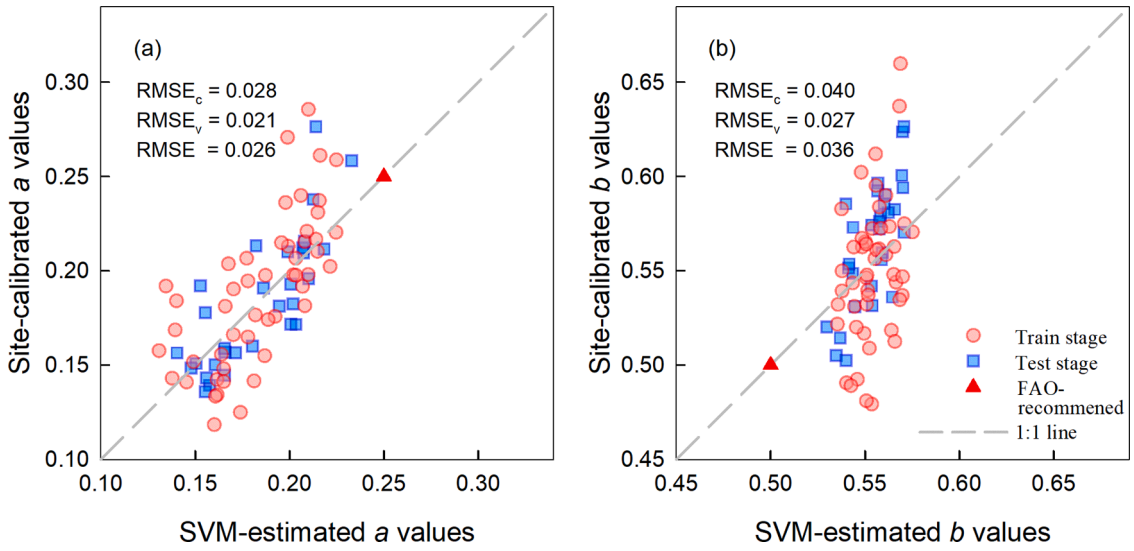

Fig. 4. Comparisons among the site-calibrated, the FAO-recommended, and the SVM-estimated *a* and *b* coefficients of the Ångström-Prescott formula at the train stage (50 weather stations involved, red circles) and the test stage (30 weather stations involved, blue squares) stages in the support vector machine (SVM) algorithm in China mainland. The red triangles are FAO recommended A-P coefficients ($a = 0.25$ and $b = 0.50$). $RMSE_c$ and $RMSE_v$ represents the root mean square error in the train and test stage, respectively. *RMSE* represents the general root mean square error of both train and test stage. The gray dashed line is the 1:1 line.

distributions of the estimated R_s with SVM-simulated *a* and *b* coefficients were similar to the estimated R_s with site-calibrated *a* and *b* coefficients. However, R_s estimated with the FAO-recommended *a* and *b* coefficients overestimated R_s in the range 5–10 MJ m^{-2} d^{-1} (the red area under 1:1 line in Fig. 6b). Compared with the FAO-recommended *a* and *b* coefficients, the SVM-estimated *a* and *b* coefficients can provide more reliable and accurate estimations of R_s through the A-P formula in China.

3.3. ET_0 estimations with or without R_s input

3.3.1. National scale evaluation of ET_0 estimation

The ET_0 values estimated with and without R_s input were compared at the training stage (Table 3) and the testing stage of the SVM-ET_0 model (Fig. 7). For the estimated ET_0 values through the P-M equation with three kinds of R_s estimations, the ET_0-estimations based on the site-calibrated *a* and *b* coefficients achieved the highest estimation accuracy estimation. Beside, ET_0 estimated based on the SVM-estimated *a* and *b* coefficients achieved a similar accuracy as the site-calibrated *a* and *b* coefficients, followed by the FAO-recommend *a* and *b* coefficients. Compared with ET_0 estimations through the P-M equation with R_s input, directly estimated ET_0 with the SVM method without R_s input (or SVM-ET_0 model) yielded the highest R^2 (0.979) but the lowest *RMSE* (0.282 mm d^{-1}) and *MAE* (0.180 mm d^{-1}) values. Hence, the SVM-ET_0 model provided the highest estimation accuracy at the training stage.

The machine-learning-based estimation model was further tested at the rest 30 national weather stations with solar radiation measurements and the results were also compared with those through the P-M equation with R_s input derived from three kinds of R_s estimation methods (Figs. 7 and S1, Table S1). Generally, ET_0 estimations with the R_s input achieved stable accuracy. The values of R^2 and *RMSE* of ET_0 estimations of the 30 weather stations were close to those of the 50 weather stations involved in the training stage of the SVM- ET_0 model. Similar to the results at the training stages, the highest estimation accuracy was obtained by the machine learning models (Table S1), followed by the P-M models with R_s inputs estimated with the site-calibrated, SVM-estimated, and FAO-recommended values of A-P coefficients of *a* and *b*. The R^2 and RMSE values of the SVM-ET_0 model were 0.973 and 0.302 mm d^{-1} at the testing stage (Fig. 7d). The data points were more concentrated than the ET_0 estimations through the P-M model with R_s inputs derived from the three different methods. It was noteworthy that there was small difference in ET_0 estimations among the four machine-learning-based models (Fig. S1). The RMSE values were 0.302, 0.309, 0.305, and 0.306 mm d-1 for the SVM-, BP-, Cubist-, and ELM-ET_0 models, respectively. The stable performance indicated the great potential of machine learning algorithms in daily ET_0 estimation.

3.3.2. Site-scale evaluation of ET_0 estimations

ET_0 estimations based on the SVM-estimated *a* and *b* coefficients were evaluated at eight randomly selected weather stations in the four different climatic zones in China mainland (Fig. 8). Good agreements were found between ET_0 estimations based on the observed R_s and estimated R_s through the A-P formula with the SVM-estimated *a* and *b* coefficients since all of the linear-fitting slopes were close to 1.0. The R^2 values of these selected stations were all greater than 0.95, expect for the Station 55299 in the MPZ region ($R^2 = 0.909$ and $RMSE = 0.350$ mm d^{-1}; Fig. 8a). However, the other station (Station 56146; Fig. 8b) in the same climatic zone achieved a much better simulation accuracy ($R^2 = 0.976$ and $RMSE = 0.202$ mm d^{-1}). The smallest estimation errors were obtained in the TCZ zone, where the *RMSE* values were both close to 0.21 mm d^{-1} and the R^2 values were equal to 0.989.

In addition, the SVM-ET_0 model was also evaluated at the same eight weather stations in the four different climatic zones in China mainland (Fig. 9). Compared with the ET_0 estimations through the P-M model based on R_s input estimated with SVM-estimated A-P coefficients, the SVM-ET_0 model provided slightly better estimations at six stations. Generally, the difference between these two methods was small. Similar to the ET_0 estimations with R_s inputs, the smallest R^2 was obtained at the Station 55299 in the MPZ zone ($R^2 = 0.914$ and $RMSE = 0.373$ mm d^{-1}; Fig. 9a). The performance of the SVM-ET_0 model was also poor at the other station in the MPZ (Station 56146; Fig. 9b). The values of R^2 and *RMSE* were 0.957 and 0.340 mm d^{-1}. The results indicated some considerable uncertainty in direct ET_0 estimation with the SVM-ET_0 model in the MPZ zone, where the topography and climate varied greatly. Generally, unbiased estimation errors were found in each climatic zone with the SVM-ET_0 model since the ratios of the fitting functions were all close to 1.0.

3.4. Estimated average annual ET_0 in China mainland

Average annual ET_0 values estimated with the P-M model with R_s input were compared with the ET_0 directly estimated with the SVM-ET_0 model without R_s input (Fig. 10). In this study, only the ET_0 estimation

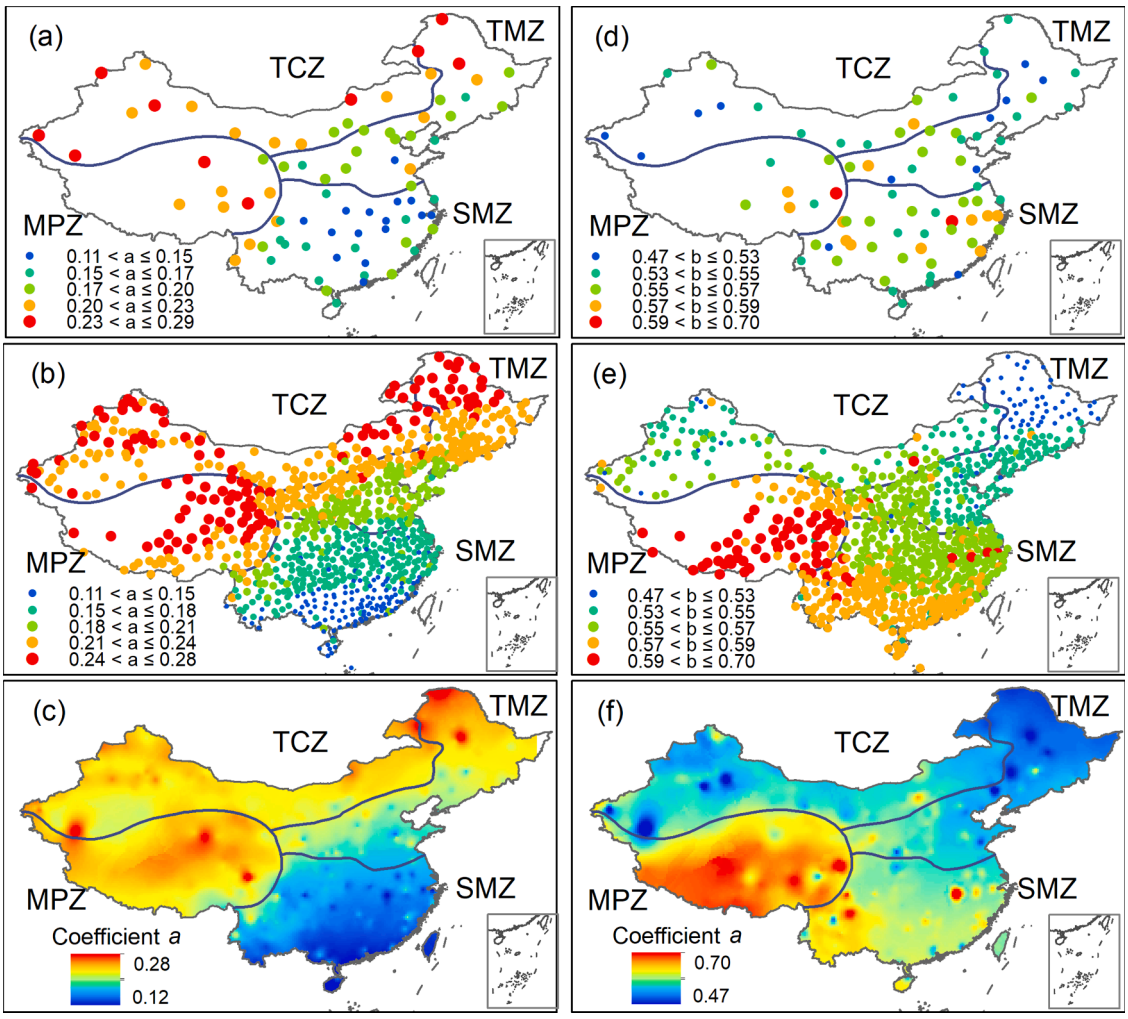

Fig. 5. Spatial distributions of coefficient *a* (a, b, c) and *b* (c, d, e) of the Ångström-Prescott formula in China. (a) and (d) are the site-calibrated values of the 80 national weather stations with direct solar radiation measurements, (b) and (e) are the SVM-estimated coefficient *a* and *b* values for the other 839 weather stations without direct solar radiation measurements through the support vector machine (SVM) method, and (c) and (f) are the interpolated coefficients of *a* and *b* through the *Kriging* interpolation method in whole China.

Table 2
Determination coefficient (R^2), root mean square error (*RMSE*, mm d^{-1}), and mean absolute error (*MAE*, mm d^{-1}) of R_s estimation though the Ångström-Prescott formula with three different kinds of coefficients for the 50weather station involved at the train stage of the SVM (support vector machine) machine learning algorithm.

Method	Statistical indices			
	R^2	*RMSE* (MJ m^{-2} d^{-1})	*NRMSE* (%)	*MAE* (MJ m^{-2} d^{-1})
Site-calibrated	0.886	2.622	18.040	1.935
FAO-recommended	0.860	3.071	21.802	2.261
SVM-estimated	0.889	2.534	18.647	1.853

with the R_s input based on the SVM-estimated *a* and *b* coefficients were selected from the three Penman-Monteith-dependent methods due to its satisfactory estimation accuracy and extrapolation ability. In general, these two methods obtained very similar pattern of ET_0 variations. The ranges of ET_0 values were 560–1838 mm $year^{-1}$ by the P-M equation with R_s input and 496–1826 mm $year^{-1}$ by the SVM- ET_0 model without R_s input. Beside, similar spatial distributions of ET_0 were also obtained by these two methods. The low- ET_0 areas were mainly in the TMZ and SMZ zones in eastern China. The high- ET_0 areas were mainly in the MPZ and TCZ zones in western China. However, the P-M model with R_s input slightly underestimated the ET_0 in some large areas in China, especially in the SMZ and western TCZ.

4. Discussion

4.1. Main factors that affecting the a and b coefficients of A-P formula

The variations in the *a* and *b* coefficients of the Ångström-Prescott formula resulted from both geographical location and weather conditions (Adaramola, 2012; Liu et al., 2014). The coefficient *a* represents the astronomical radiant fraction that reaches the earth's surface on a cloudy day and is affected by weather conditions (e.g. humidity, concentration of atmospheric particulates, and cloudiness) (Almorox and Hontoria, 2004). Previous studies have pointed out that the variations of coefficient *a* are mainly due to the altitude of target area (Chen et al., 2006; Liu et al., 2019). In this study, larger values of the site-calibrated coefficient *a* were found in the middle- and high- latitude regions in northern and western China, where the altitudes were also higher than the areas in eastern and southern China (Fig. 5a). Coefficient *b* reflects the transport properties (aerosol density) of a cloudless atmosphere, affected by altitude, and depends mainly on the total water content and turbidity of the atmosphere (Liu et al., 2009). As opposed to coefficient *a*, higher values of calibrated coefficient *b* mainly concentrated in the

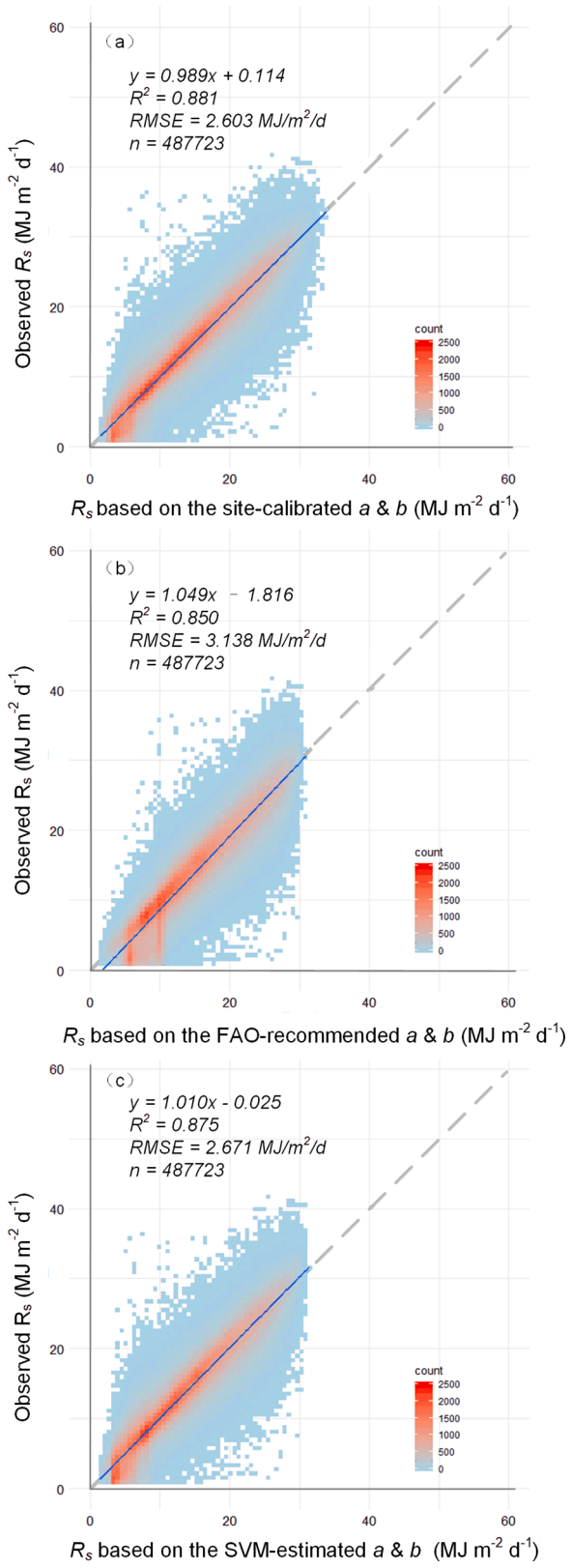

Fig. 6. Comparisons among R_s estimations through the Ångström-Prescott (A-P) formula with the site-calibrated (a), the FAO-recommended (b), and the SVM-estimated (c) a and b coefficients at the 30 national weather stations involved at the test stage of SVM (support vector machine) machine learning algorithm. The gray dashed lines are 1:1 line. The blue solid lines are linear regression lines.

Table 3
Determination coefficient (R^2), root mean square error (RMSE, mm d^{-1}), normalized root mean square error (NRMSE, %), and mean absolute error (MAE, mm d^{-1}) of the ET_0 estimation through the Penman-Monteith equation with differently estimated R_s inputs and the ET_0 estimation through the support vector machine (SVM) method without R_s inputs at the 50 weather stations involved at the train stage of the SVM-ET_0 model.

Method	A-P coefficients	Number of data used in train	Statistical indices			
			R^2	*RMSE* (mm d^{-1})	*NRMSE* (%)	*MAE* (mm d^{-1})
P-M equation with R_s input	Site-calibrated	819862	0.972	0.300	11.528	0.182
	FAO-recommend	0	0.959	0.383	14.671	0.250
	SVM-estimated	819862	0.969	0.317	12.136	0.200
SVM model without R_s input		819862	0.979	0.282	9.702	0.180

middle- and low-latitude regions in China, where the climate was relatively humid (Fig. 5d). Generally, coefficient b would be affected by more variables and therefore the distribution of coefficient b was not as regular as coefficient a. Compared with the site-calibrated a and b coefficients based on the R_s measurements at the 80 national weather stations in China, the values of a and b recommended by the FAO were relatively higher and lower, respectively.

Previous works have found that the variations in the a and b coefficients were mainly due to latitude and altitude of selected sites (De Souza et al., 2016; Paulescu et al., 2016). Beside, Liu et al. (2014) pointed out that the estimation accuracy could not always be improved with the increment of variable number used to establish the predictive models. To simplify the data requirement and increase the applicability of the machine learning models in most areas, we only used site altitude and latitude information to establish the SVM models to estimate the a and b coefficients in this study. Good agreement was found between the site-calibrated and SVM-estimated values of coefficient a, but larger estimation errors were found in the coefficient b. This was mainly because coefficient b was affected by more complex interactions among meteorological variables. Thus, some new machine learning models, which established with more variables (e.g. atmosphere water and particulates content, and etc.), might be needed to improve the estimation accuracy for coefficient b.

4.2. *Performance of different kinds of a and b coefficients in R_s estimations*

In this study, the a and b coefficients of the Ångström-Prescott formula were calibrated at 80 national weather stations with solar radiation measurements. Then, the 80-site calibrated a and b coefficients were used to establish the SVM model to estimate the a and b coefficient for the 839 national weather stations without R_s measurements. Compared with the site-calibrated and the SVM-estimated a and b coefficients, the FAO had recommended larger and smaller values for the coefficients of a and b, respectively. In addition, great variations were found in the two coefficients across the 80 weather stations, which were usually ignored in actual application of the A-P formula since a pair of default values were usually set for the coefficients of a and b. The results also indicated the largest estimation errors of R_s with FAO-recommended a and b coefficient for the A-P formula (*RMSE* = 3.138 MJ m^{-2} d^{-1}, R^2 = 0.850). However, the spatial variations of the a and b coefficients could be captured by the SVM models. Good R_s estimation accuracy was found for the SVM-estimated a and b coefficients at both the train and the test stages of the SVM model, since the *RMSE* values of the two stages were 2.534 and 2.671 MJ m^{-2} d^{-1}, respectively. It was notable that the R_s estimation accuracy with the SVM-estimated a and b

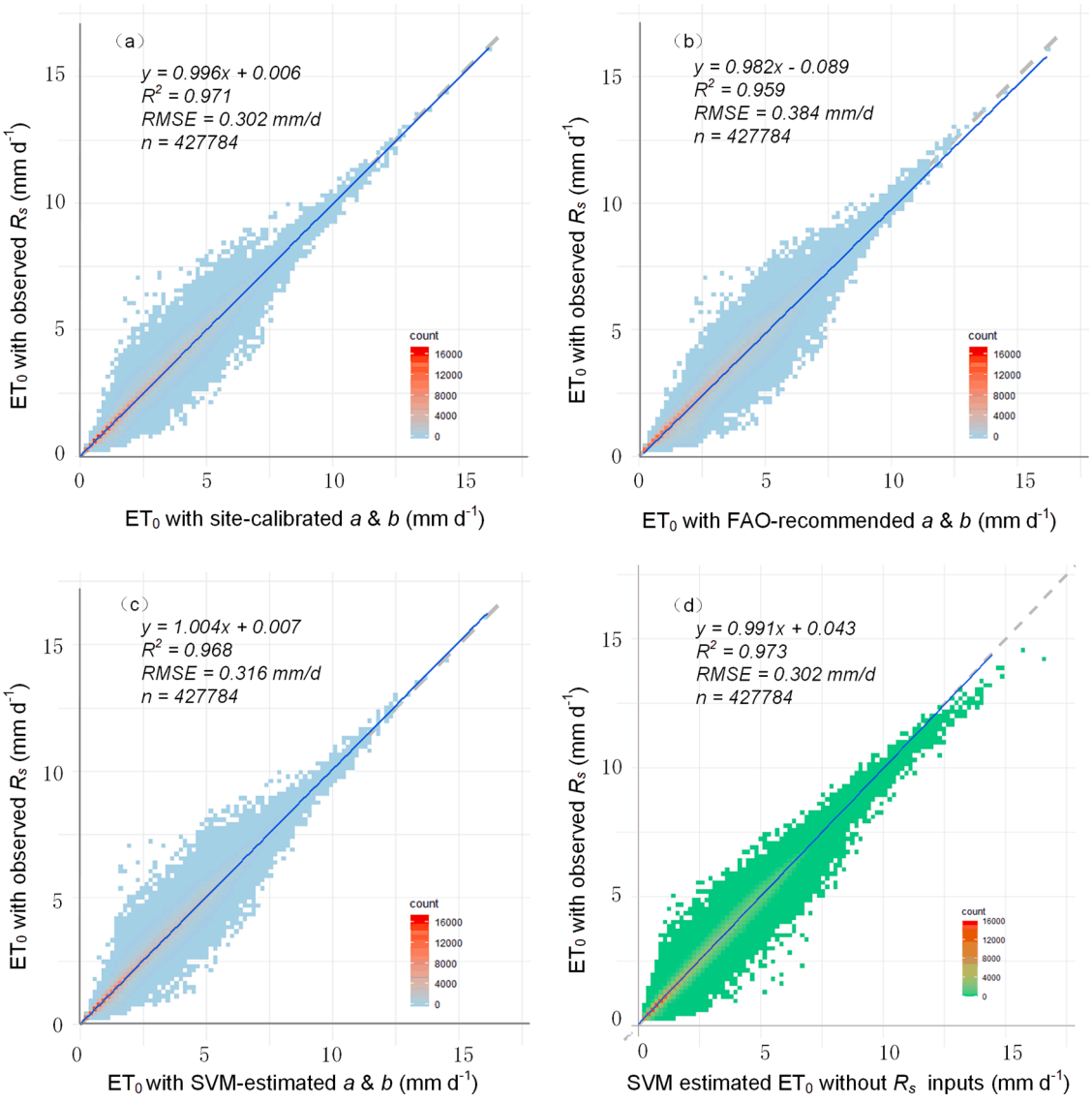

Fig. 7. Comparisons among the ET_0 estimated through the Penman-Monteith equation with R_s inputs derived from the Ångström-Prescott formula and the site-calibrated coefficients (a), the FAO-recommended coefficients (b), and SVM-estimated coefficients (c) and the ET_0 directly estimated through the SVM-ET_0 model (d) at the 30 weather stations with solar radiation measurements at the test stage of the SVM- ET_0 model. R^2 is the determination coefficient; *RMSE* is the root mean square error; and *n* is the number of data points. The gray dashed lines are 1:1 lines and the blue lines are linear regression lines. Red color shows higher data density.

coefficients was even better than those of the site-calibrated *a* and *b* coefficients in the train stage of the SVM model (with larger R^2 but smaller *RMSE* and *MAE* values).

Previous R_s-estimation studies based on machine learning algorithm directly estimated daily R_s with various combinations of meteorological variables. For instance, Chen et al. (2014) used 20 different combinations of input variables to establish SVM models for R_s estimations at 15 cities in China and found the *RMSE* values were all less than 2.3 MJ m^{-2} d^{-1} and the average *RMSE* was about 1.097 MJ m^{-2} d^{-1}. The smaller errors in their study were mainly because the SVM models were established and tested at each individual site. In contrast, our SVM models were established at 50 different national weather stations with solar radiation measurement and then tested at the rest 30 weather stations in whole China mainland. Fan et al. (2019) compared 12 empirical models and 12 machine learning methods in daily R_s estimations at 50 sites in China. Similar *RMSE* values (2.055–2.751 MJ m^{-2} d^{-1}) were also achieved as our works (2.073–2.573 MJ m^{-2} d^{-1}). He et al. (2020) established the SVM models with different combinations of input variables at the same sites as our study. The RMSE values in their study varied from 1.957 to 4.057 MJ m^{-2} d^{-1}. In this study, different from the above studies, the SVM algorithm was used to establish a common model for estimation of the *a* and *b* coefficients of Ångström-Prescott formula across China. Then, with the SVM-estimated *a* and *b* coefficients, R_s was estimated just with inputs of *n*, *N*, R_a and the A-P formula. The results indicated no obvious difference between accuracies of direct R_s estimation through machine learning method independent of the A-P formula and indirect R_s estimation through the A-P formula with machine-learning-estimated *a* and *b* coefficients. Generally, R_s could be accurately estimated just with daily sunshine hour and the SVM-estimated *a* and *b* coefficients for the A-P formula in China mainland. Hence, this indirect method and the SVM-estimated *a* and *b* coefficients values were recommended for users who are not good at computer programming and not familiar with machine learning methods. However, since only the SVM method and one kind of input variable combination (latitude and altitude) were used in the estimation of *a* and *b* coefficients, more machine learning methods and variable combinations should be taken into account to improve the estimation accuracy of *a* and *b* coefficients in future study.

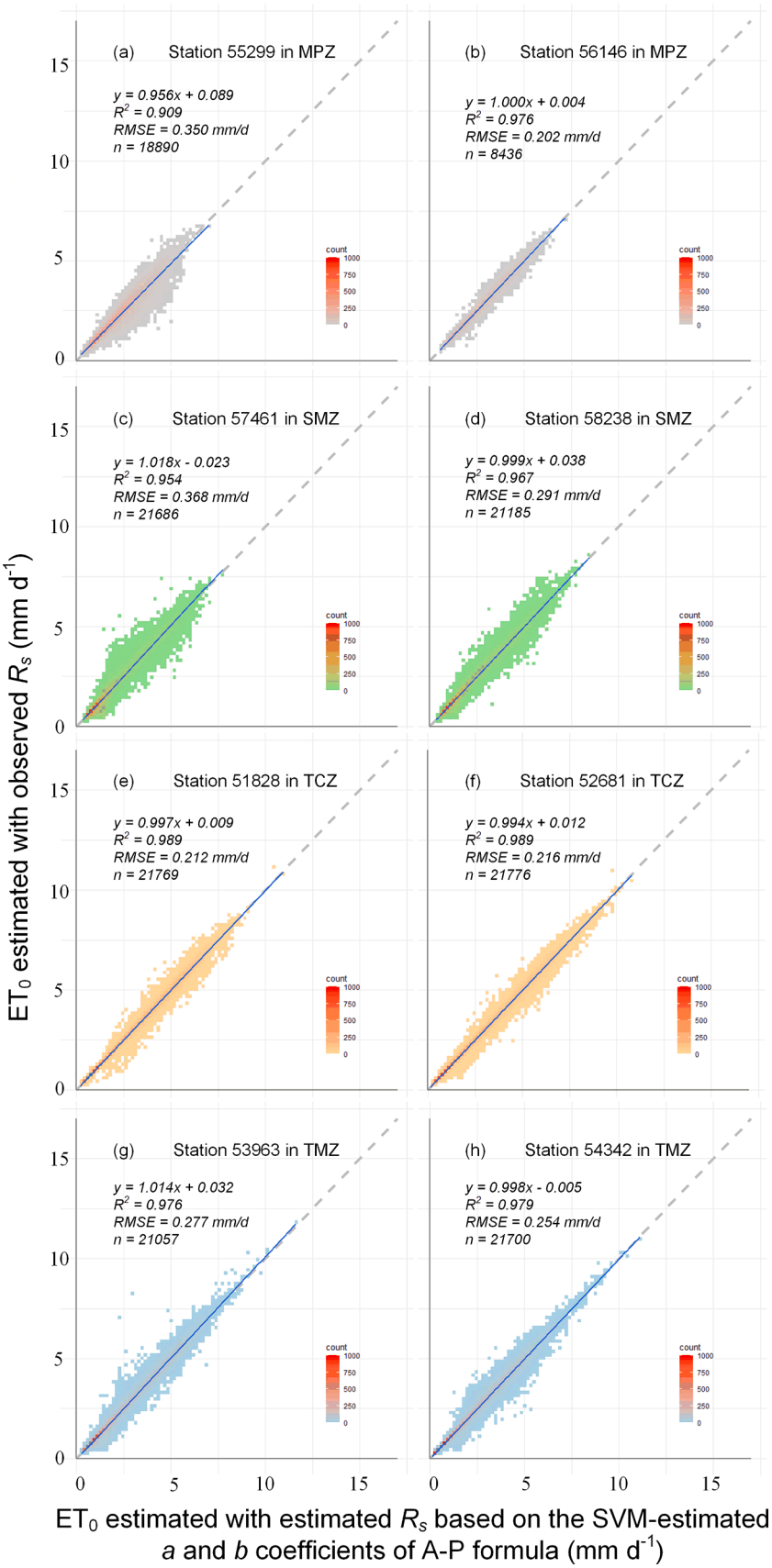

Fig. 8. ET_0 estimations with R_s input estimated based on the SVM-estimated a and b coefficients of the Ångström-Prescott at eight randomly selected representative weather stations in the climatic zones of MPZ (a, b), SMZ (c, d), TCZ (e, f), and TMZ (g, h) at the test stage. R^2 is the determination coefficient, *RMSE* is the root mean square error, and n is the number of data points. The gray dashed lines are 1:1 lines and the blue lines are linear regression lines. Color bars show the data density.

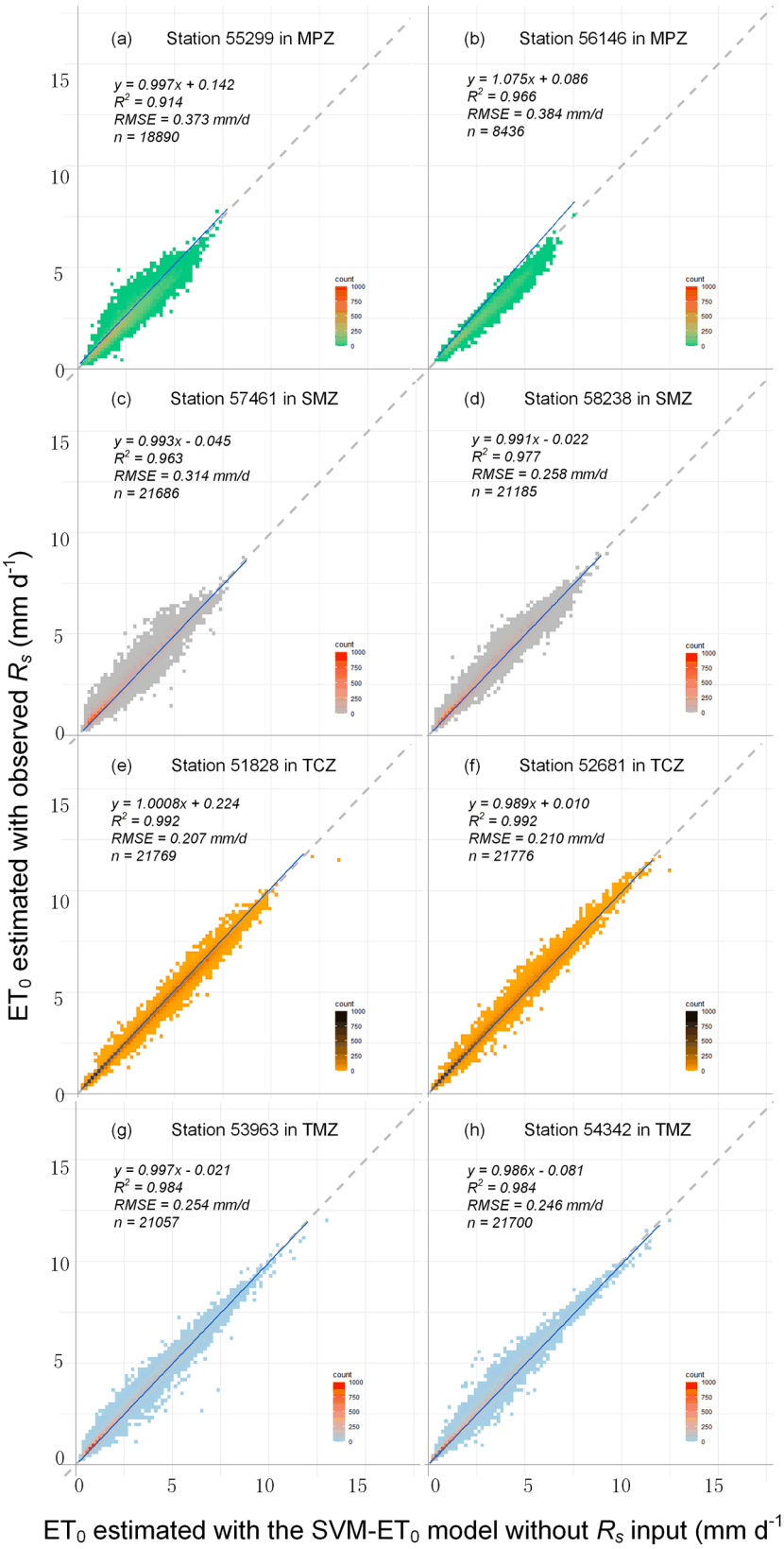

Fig. 9. ET_0 estimations through the SVM- ET_0 model at eight randomly selected representative weather stations in the climatic zones of MPZ (a, b), SMZ (c, d), TCZ (e, f), and TMZ (g, h) at the test stage of the SVM algorithm. R^2 is the determination coefficient, *RMSE* is the root mean square error, *n* is the number of data points. The gray dashed lines are 1:1 lines and the blue lines are linear regression lines. Color bars show the data density.

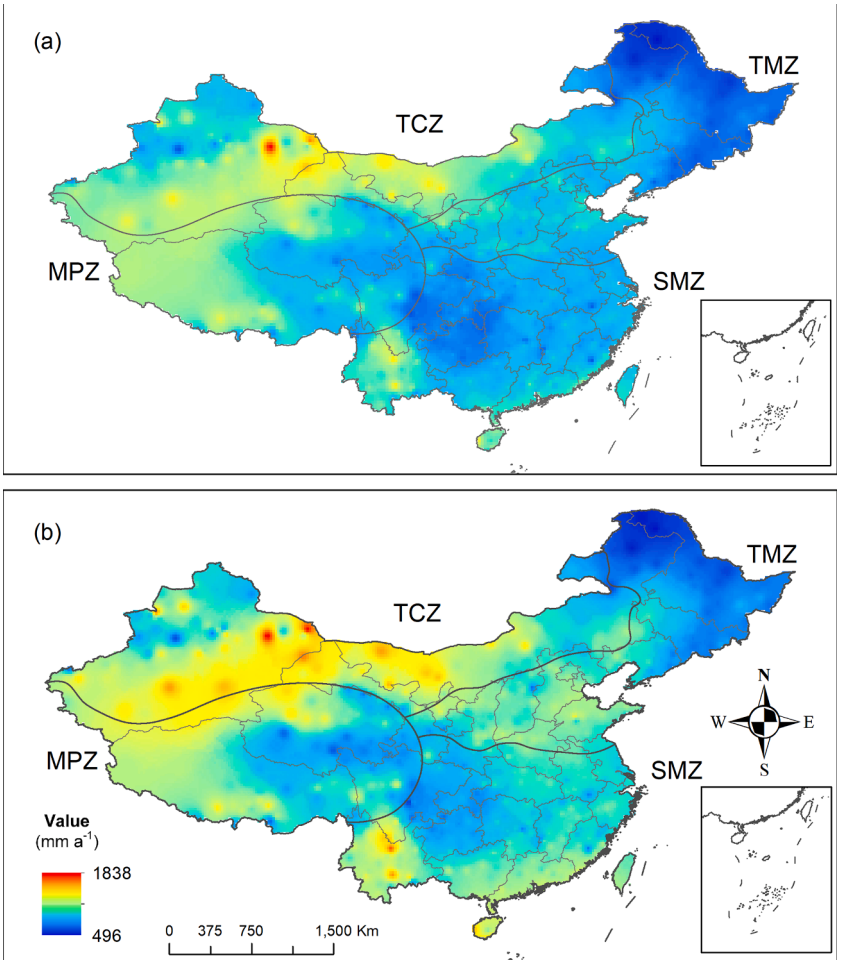

Fig. 10. Average annual ET_0 estimated with the Penman-Monteith (P-M) equation with R_s input (a) and the SVM-ET_0 model without R_s input (b). Daily R_s values used in the P-M equation were estimated through the Ångström-Prescott (A-P) formula with the SVM-estimated *a* and *b* coefficients. The ET_0 values in the two figures were both created through the inverse-distance-weight interpolation of the ET_0 values estimated at the 839 national weather stations without direct R_s measurements.

4.3. Performance of the ET_0 estimation solutions with or without R_s inputs

4.3.1. ET_0 estimation through Penman–Monteith equation with R_s inputs

In this study, ET_0 was supposed to be estimated under the scenario of missing R_s observations with two alternative solutions. In the first solution, ET_0 was estimated through the P-M formula with differently estimated R_s data. And in the second solution, ET_0 was directly estimated based on SVM method and similar weather variables used the P-M model except for R_s. Since the A-P formula was recommended by the FAO for R_s estimation in areas without solar radiation measurements, the first solution thus used the A-P formula to estimate R_s. Three kinds of A-P coefficients (the site-calibrated, the FAO-recommended, and the SVM-estimated) were all used and compared in R_s estimations and further in ET_0 estimations. Errors of ET_0 estimations based on the FAO-recommended *a* and *b* coefficients were the largest in both the train and test stages of the SVM method, while the site-calibrated *a* and *b* coefficients obtained the highest accuracy. Beside, there was no great difference between ET_0 estimations with the site-calibrated and the SVM-estimated *a* and *b* coefficients since the differences of R^2, *RMSE*, and *MAE* values were 0.003, 0.015 mm $d^{-1,}$ and 0.018 mm d^{-1}, respectively. This indicated a great potential for the SVM-estimated *a* and *b* coefficients to be used in ET_0 estimation in China. Mousavi et al. (2014) used two empirical (cloud- and temperature-based) models in the calibrations of the *a* and *b* coefficients. Then, the calibrated *a* and *b* coefficients were used to estimate R_s, which was further used in the ET_0 estimations in Iran. Their findings suggested that the errors generated by the calibrated *a* and *b* coefficients in R_s estimation were smaller than those generated by the FAO-recommended *a* and *b* coefficients.

Evaluation of error propagation became important in the R_s dependent ET_0 estimations since new errors were introduced into the estimation of daily R_s. It was noteworthy that the accuracy in R_s dependent ET_0 estimations was higher than the R_s estimation accuracy due to higher R^2 and smaller *NRMSE* for all of the three kinds of *a* and *b* coefficients. Sabziparvar et al. (2013) used the P-M model to estimate ET_0 with the FAO recommended and locally calibrated *a* and *b* coefficients of A-P formula in Iran. They also found that the deviations in ET_0 estimations were smaller than those in R_s estimations. ET_0 estimated with the P-M model was both determined by the energy term and the dynamic term (Allen et al., 1998). Thus, R_s could only explain a part of the variations in ET_0 estimation and led to smaller variations in ET_0 estimations.

4.3.2. ET_0 estimation through machine learning methods without R_s input

Estimation of ET_0 can be treated as complex non-linear regressions that relying on huge climatic variables (Zhang et al., 2020). The estimation accuracy was usually unsatisfactory and data requirements were difficult to meet in most ET_0 estimation cases based on empirical models (Luo et al., 2014). With the development of computing algorithms and hardware equipment, machine learning methods have been gradually used in ET_0 estimations (Fan et al., 2018b; Ferreira et al., 2019). Machine learning methods required no prior knowledge about the non-linear processes in ET_0 estimations and could simplify the

establishment of estimation models with flexible combinations of input climatic variables. In this study, the SVM method was employed to establish the ET_0 estimation models (or SVM-ET_0) in different climatic zones in China. To exclude the influences of different input combinations of meteorological variables on ET_0 estimation accuracy, ET_0 estimation without R_s inputs employed the similar weather variables as the P-M equation.

Because of the lack of direct R_s measurements, the variable R_a or extraterrestrial solar radiation was used to establish the SVM-ET_0 model. Compared with the ET_0 estimation through the Penman-Monteith equation with differently estimated R_s data, the ET_0 directly estimated with the SVM-ET_0 model obtained the highest R^2 (0.979 and 0.973) but the smallest RMSE (0.282 and 0.302 mm d^{-1}) values both at the training and testing stages of the SVM models. Generally, the simulation accuracy of the SVM-ET_0 model was acceptable comparing with the other ET_0 estimation using machine learning methods in China. Wen et al. (2015) used SVM to estimate daily ET_0 with limited climatic data in the Ejina basin under extreme arid weather conditions in northwestern China. The R^2 was 0.772–0.950, *RMSE* was 0.262–0.539 mm $d^{-1,}$ and *MAE* was 0.07–0.446 in test stage. Feng et al. (2017) evaluated the machine learning methods of RF, WNN, and GRNN for daily ET_0 estimation in southwest China. The results indicated that the RF method obtained slightly better accuracy with R^2 of 0.894–0.988 and *MAE* of 0.1–0.3 mm d^{-1}. Fan et al. (2019a) compared the machine learning methods of LightGBM, M5Tree, and RF in ET_0 estimation with different combinations of climatic variables in humid subtropical region of southeastern China. The results indicated that LightGBM outperformed M5Tree and RF in almost all input combinations in the test stage with R^2 of 0.85–0.97, *RMSE* of 0.27–0.58 mm d^{-1} and *NRMSE* of 0.11–0.24. Zhang et al. (2020) compared machine learning methods of CatBoost, RF, and GRNN models in daily ET_0 estimations at 15 weather stations covering arid and semi-arid regions of northern China. The estimation results of their most accurate model were also similar to our results with average R^2 of 0.846–0.999, *RMSE* of 0.096–0.821 mm $d^{-1,}$ and *MAE* of 0.067–0.603 mm d^{-1}. Limited by the scarcity of R_s measurements, the above studies were mainly carried out in some selected regions of China. Thus, the SVM-ET_0 model without R_s input could be applied in daily ET_0 estimations with acceptable accuracy in China.

4.3.3. Comparison between ET_0 estimation solutions with and without R_s inputs

In this study, the ET_0 estimations were carried out with two alternative solutions in China mainland when direct R_s measurements were missing: based on P-M formula and differently estimated R_s (Solution 1) and based on machine learning method of SVM and relevant meteorological variables except for R_s (Solution 2). In Solution 1, R_s was estimated and compared through the Ångström-Prescott formula approach with the site-calibrated, the FAO-recommended, and the SVM-estimated a and b coefficients. In the Solution 2, similar meteorological variables involved in the Penman-Monteith equation were used to establish the SVM-ET_0 model to directly estimate daily ET_0. In general, ET_0 could be more accurately and stably estimated through the Penman-Monteith formula with estimated R_s data, especially R_s data estimated through the Ångström-Prescott formula and the SVM-estimated a and b coefficients. In addition, the R^2 and RMSE values were similar between the two solutions, except for the P-M model with R_s input estimated with the A-P formula and FAO recommended values for coefficients a and b.

The Penman-Monteith equation in Solution 1 required several input meteorological variables. However, some simplified empirical ET_0 estimation models (e.g. Blaney-Criddle model, Hargreaves-Samani model), in which fewer input variables are required, always showed unsatisfactory accuracy in comparison with the machine learning models (Wen et al., 2015). Fan et al. (2019a) compared four empirical models (the Hargreaves-Samani model with input variable of T_{mean}, the Tabari model with input variable of e_s, e_a, and U_2, the Makkink model with input variable of T_{mean} and R_s, and the Trabert model with input variable of T_{max}, T_{mean} and R_s) and three machine learning methods (LightGBM, M5Tree, and RF) in ET_0 estimations in the humid subtropical region of China. Their results showed that all of the three machine learning models yielded better daily ET_0 estimations than the empirical models with corresponding combinations of input meteorological variables. In addition, Wen et al. (2015) concluded that machine learning models established with T_{max}, $T_{mean,}$ and R_s were enough to accurately estimate daily ET_0.

Generally, the SVM-ET_0 model outperformed the three P-M models. However, the result did not indicate the failure of the P-M models, especially the one with the SVM-estimated A-P coefficients. Considering the similar estimation accuracy as the SVM-ET_0 model, the P-M model combining with R_s input estimated through the Ångström-Prescott formula with SVM-estimated a and b coefficients are recommended for the users who are not good at computer programing. However, the machine learning methods are also recommended when even fewer meteorological variables are available. In this study, only similar input meteorological variables involved in the Penman-Monteith equation to establish two solutions for daily ET_0 estimation when R_s measurements were missing. In further studies, new machine learning algorithms (Mohammadi and Mehdizadeh, 2020; Salam and Islam, 2020) and different combinations of input meteorological variables (Paredes et al., 2020; Xiang et al., 2020) needed to be explored to improve ET_0 estimation accuracy and efficiency under different conditions of available weather data.

5. Conclusions

In this study, we assessed the performances of two alternative solutions for daily ET_0 estimation when short-wave sloar radiation (R_s) measurements are missing or lacking in China mainland. In the first solution with R_s input, R_s was estimated through the Ångström-Prescott (A-P) formula and was then used in the standard Penman-Monteith equation to estimate daily ET_0. In this soltion, the variations in the a and b coefficients of A-P formula should not be ignored in countries with complex topography and climate conditions. In the second solution without R_s input, daily ET_0 was estimated based on the machine learning method of SVM (suppot vector machine) and the similar meteorological variables involved in the P-M equation except for R_s. Some main conclusions have been drawn as follows:

The FAO recommended larger coefficient a (0.25) but smaller coefficient b (0.5) for the A-P formula in China mainland, which could result in large simulation errors in daily R_s estimations. The a and b coefficients of the A-P formula, which were estimated through the SVM method with input combination of site-specific altitude and latitude, achieved an accuracy of R_s estimation close to that of the site-calibrated a and b coefficients, which indicated a great potential for the SVM-estimated a and b coefficients in R_s estimations with the A-P formula in regions where R_s measurements are missing or scarce. However, the error propagation from R_s estimation to ET_0 estimation decreased when using the estimated R_s to drive the P-M model to estimated daily ET_0.

Compared with the method based on the P-M equation with differently derived R_s input, the machine-learning-based ET_0 model established based on input meteorological variables of T_{max}, T_{min}, RH, $U_{2,}$ and R_a obtained better ET_0 estimation accuracy both at the training and testing stages. Generally, both the two solutions with and without R_s input could be used in ET_0 estimations in China mainland. However, if only R_s measurements are missing or scarce, the P-M equation with R_s inputs estimated through the A-P formula with SVM-estimated a and b coefficients is recommended for daily ET_0 estimations in China mainland for the users who are not good at coumpter programing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Acknowledgments

This research was partially supported by the National Key R&D Program of China (No. 2018YFB1500901), the Natural Science Foundation of China (No. 52079115, 41961124006), the Key Research and Development Program of Shaanxi (No. 2019ZDLNY07-03), the Open Project Fund from the Key Laboratory of Eco-Environment and Meteorology for the Qinling Mountains and Loess Plateau, Shaanxi Provincial Meteorological Bureau (No. 2019Z-5) , and the “111 Project” (No. B12007) of China. The author would like to thank Chinese Meteorological Administration (CMA) for providing the meteorological data.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.agrformet.2022.108864.

References

Adaramola, M.S., 2012. Estimating global solar radiation using common meteorological data in Akure, Nigeria. Renew. Energy 47, 38–44.
Allen, R.G., Pereira, L.S., Raes, D., Smith, M., 1998. Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements-FAO Irrigation and Drainage Paper 56, 300. FAO, Rome, p. D05109.
Almorox, J., Hontoria, C., 2004. Global solar radiation estimation using sunshine duration in Spain. Energy Convers. Manag. 45 (9–10), 1529–1535.
Ångström, A., 1924. Solar and terrestrial radiation. Report to the international commission for solar research on actinometric nvestigations of solar and atmospheric radiation. Q. J. R. Meteorol. Soc. 50 (210), 121–126.
Bristow, K.L., Campbell, G.S., 1984. On the relationship between incoming solar radiation and daily maximum and minimum temperature. Agric. For. Meteorol. 31, 159–166.
Chen, J.L., Li, G.S., 2013. Evaluation of support vector machine for estimation of solar radiation from measured meteorological variables. Theor. Appl. Climatol. 115 (3–4), 627–638.
Chen, J.L., Liu, H.B., Wu, W., Xie, D.T., 2011. Estimation of monthly solar radiation from measured temperatures using support vector machines-a case study. Renew. Energy 36 (1), 413–420.
Chen, R., Lu, S., Kang, E., Yang, J., Ji, X., 2006. Estimating daily global radiation using two types of revised models in China. Energy Convers. Manag. 47 (7–8), 865–878.
De Souza, J.L., Lyra, G.B., Dos Santos, C.M., Ferreira Junior, R.A., Tiba, C., Lyra, G.B., Lemes, M.A.M., 2016. Empirical models of daily and monthly global solar irradiation using sunshine duration for Alagoas State, Northeastern Brazil. Sustain. Energy Technol. Assess. 14, 35–45.
Fan, J., Ma, X., Wu, L., Zhang, F., Yu, X., Zeng, W., 2019a. Light gradient boosting machine: an efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agric. Water Manag. 225, 105758.
Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., Xiang, Y., 2018a. Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. Energy Convers. Manag. 164, 102–111.
Fan, J., Wu, L., Ma, X., Zhou, H., Zhang, F., 2020. Hybrid support vector machines with heuristic algorithms for estimation of daily diffuse solar radiation in air-polluted regions. Renew. Energy 145, 2034–2045.
Fan, J., Wu, L., Zhang, F., Cai, H., Zeng, W., Wang, X., Zou, H., 2019b. Empirical and machine learning models for predicting daily global solar radiation from sunshine duration-A review and case study in China. Renew. Sustain. Energy Rev. 100, 186–212.
Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X., Xiang, Y., 2018b. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. Agric. For. Meteorol. 263, 225–241.
Feng, Y., Cui, N., Gong, D., Zhang, Q., Zhao, L., 2017. Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling. Agric. Water Manag. 193, 163–173.
Ferreira, L.B., da Cunha, F.F., de Oliveira, R.A., Fernandes Filho, E.I., 2019. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM-a new approach. J. Hydrol. 572, 556–570.
Gavilán, P., Berengena, J., Allen, R.G., 2007. Measuring versus estimating net radiation and soil heat flux: impact on Penman-Monteith reference ET estimates in semiarid regions. Agric. Water Manag. 89 (3), 275–286.
Gueymard, C., 2001. Parameterized transmittance model for direct beam and circumsolar spectral irradiance. Sol. Energy 71, 325–346.
Hassan, G.E., Youssef, M.E., Mohamed, Z.E., Ali, M.A., Hanafy, A.A., 2016. New temperature-based models for predicting global solar radiation. Appl. Energy 179, 437–450.
He, C., Liu, J., Xu, F., Zhang, T., Chen, S., Sun, Z., Zheng, W., Wang, R., He, L., Feng, H., Yu, Q., He, J., 2020. Improving solar radiation estimation in China based on regional optimal combination of meteorological factors with machine learning methods. Energy Convers. Manag. 220, 113111.
Iziomon, M.G., Mayer, H., 2002. Assessment of some global solar radiation parameterizations. J. Atmos. Sol. Terr. Phys. 64 (15), 1631–1643.
Jin, Z., Yezheng, W., Gang, Y., 2005. General formula for estimation of monthly average daily global solar radiation in China. Energy Convers. Manag. 46 (2), 257–268.
Karatzoglou, A., Smola, A., Hornik, K., 2004. Kernlab-an S4 package for Kernel methods in R. J. Stat. Softw. 1–20.
Kisi, O., Parmar, K.S., 2016. Application of least square support vector machine and multivariate adaptive regression spline models in long term estimation of river water pollution. J. Hydrol. 534, 104–112.
Liu, X., Li, Y., Zhong, X., Zhao, C., Jensen, J.R., Zhao, Y., 2014. Towards increasing availability of the Ångström-Prescott radiation parameters across China: spatial trend and modeling. Energy Convers. Manag. 87, 975–989.
Liu, X., Mei, X., Li, Y., Porter, J.R., Wang, Q., Zhang, Y., 2010. Choice of the Ångström–Prescott coefficients: are time-dependent ones better than fixed ones in modeling global solar irradiance? Energy Convers. Manag. 51 (12), 2565–2574.
Liu, X., Mei, X., Li, Y., Zhang, Y., Wang, Q., Jensen, J.R., Porter, J.R., 2009. Calibration of the Ångström-Prescott coefficients (a, b) under different time scales and their impacts in estimating global solar radiation in the Yellow River basin. Agric. For. Meteorol. 149 (3–4), 697–710.
Liu, X., Xu, Y., Zhong, X., Zhang, W., Porter, J.R., Liu, W., 2012. Assessing models for parameters of the Ångström-Prescott formula in China. Appl. Energy 96, 327–338.
Liu, Y., Tan, Q., Pan, T., 2019. Determining the parameters of theÅngström-Prescott model forestimating solar radiation in differentregions of China: calibration andmodeling. Earth Sp. Sci. 6, 1976–1986.
López-Urrea, R., Olalla, F.M.S., Fabeiro, C., Moratalla, A., 2006. An evaluation of two hourly reference evapotranspiration equations for semiarid conditions. Agric. Water Manag. 86 (3), 277–282.
Luo, Y., Chang, X., Peng, S., Khan, S., Wang, W., Zheng, Q., Cai, X., 2014. Short-term forecasting of daily reference evapotranspiration using the Hargreaves-Samani model and temperature forecasts. Agric. Water Manag. 136, 42–51.
Ming, Z., Shaojie, O., Hui, S., Yujian, G., 2015. Is the “Sun” still hot in China? The study of the present situation, problems and trends of the photovoltaic industry in China. Renew. Sustain. Energy Rev. 43, 1224–1237.
Mohammadi, B., Mehdizadeh, S., 2020. Modeling daily reference evapotranspiration via a novel approach based on support vector regression coupled with whale optimization algorithm. Agric. Water Manag. 237, 106145.
Mohammadi, K., Khorasanizadeh, H., Shamshirband, S., Tong, C.W., 2016. Influence of introducing various meteorological parameters to the Angström-Prescott model for estimation of global solar radiation. Environ. Earth Sci. 75 (3).
Mousavi, R., Sabziparvar, A.A., Marofi, S., Ebrahimi Pak, N.A., Heydari, M., 2014. Calibration of the Angström-Prescott solar radiation model for accurate estimation of reference evapotranspiration in the absence of observed solar radiation. Theor. Appl. Climatol. 119 (1–2), 43–54.
Paredes, P., Pereira, L.S., Almorox, J., Darouich, H., 2020. Reference grass evapotranspiration with reduced data sets: parameterization of the FAO P-M temperature approach and the Hargeaves-Samani equation using local climatic variables. Agric. Water Manag. 240, 106210.
Paulescu, M., Stefu, N., Calinoiu, D., Paulescu, E., Pop, N., Boata, R., Mares, O., 2016. Ångström–Prescott equation: physical basis, empirical models and sensitivity analysis. Renew. Sustain. Energy Rev. 62, 495–506.
Prescott, J.A., 1940. Evaporation from a water surface in relation to solar radiation. Trans. R. Soc. S. Aust. 46, 114–118.
Richardson, C.W., 1981. Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour. Res. 17 (1), 182–190.
Rietveld, M.R., 1978. A new method for estimating the regression coefficients in the formula relating solar radiation to sunshine. Agric. Meteorol. 19 (2), 243–252.
Rivington, M., Matthews, K.B., Bellocchi, G., Buchan, K., 2006. Evaluating uncertainty introduced to process-based simulation model estimates by alternative sources of meteorological data. Agric. Syst. 88 (2–3), 451–471.
Sabziparvar, A., Shetaee, H., 2007. Estimation of global solar radiation in arid and semi-arid climates of East and West Iran. Energy 32 (5), 649–655.
Sabziparvar, A.A., Mousavi, R., Marofi, S., Ebrahimipak, N.A., Heidari, M., 2013. An improved estimation of the Angstrom-Prescott radiation coefficients for the FAO56 Penman-Monteith evapotranspiration method. Water Resour. Manag. 27 (8), 2839–2854.
Salam, R., Islam, A.R.M., 2020. Potential of RT, bagging and RS ensemble learning algorithms for reference evapotranspiration estimation using climatic data-limited humid region in Bangladesh. J. Hydrol. 590, 125241.
Şenkal, O., 2010. Modeling of solar radiation using remote sensing and artificial neural network in Turkey. Energy 35 (12), 4795–4801.
Song, Y., Achberger, C., Linderholm, H.W., 2011. Rain-season trends in precipitation and their effect in different climate regions of China during 1961-2008. Environ. Res. Lett. 6 (3), 034025.
Vapnik, V.N., 1996. The nature of statistical learning theory. Technometrics 38 (4), 409.
Wang, L., Kisi, O., Zounemat-Kermani, M., Zhu, Z., Gong, W., Niu, Z., Liu, H., Liu, Z., 2017. Estimation of solar radiation in China using different adaptive neuro-fuzzy methods and M5 model tree. Int. J. Climatol. 37 (3), 1141–1155.
Wen, X., Si, J., He, Z., Wu, J., Shao, H., Yu, H., 2015. Support-vector-machine-based models for modeling daily reference evapotranspiration with limited climatic data in extreme arid regions. Water Resour. Manag. 29 (9), 3195–3209.
Wu, L., Zhou, H., Ma, X., Fan, J., Zhang, F., 2019. Daily reference evapotranspiration estimation based on hybridized extreme learning machine model with bio-inspired

optimization algorithms: application in contrasting climates of China. J. Hydrol. 577, 123960.
Wu, Z., Du, H., Zhao, D., Li, M., Meng, X., Zong, S., 2011. Estimating daily global solar radiation during the growing season in Northeast China using the Ångström-Prescott model. Theor. Appl. Climatol. 108 (3–4), 495–503.
Xiang, K., Li, Y., Horton, R., Feng, H., 2020. Similarity and difference of potential evapotranspiration and reference crop evapotranspiration-a review. Agric. Water Manag. 232, 106043.
Yin, Y., Wu, S., Zheng, D., Yang, Q., 2008. Radiation calibration of FAO56 Penman-Monteith model to estimate reference crop evapotranspiration in China. Agric. Water Manag. 95 (1), 77–84.
Zhang, Y., Zhao, Z., Zheng, J., 2020. CatBoost: a new approach for estimating daily reference crop evapotranspiration in arid and semi-arid regions of Northern China. J. Hydrol. 588, 125087.