



RESEARCH ARTICLE

WILEY

Parameterization of the Ångström–Prescott formula based on machine learning benefit estimation of reference crop evapotranspiration with missing solar radiation data

Shang Chen^{1,2,3}  | Wenzhe Feng³  | Liang He⁴ | Wei Xiao^{1,2} | Hao Feng^{3,5} | Qiang Yu^{5,6} | Jiandong Liu⁷ | Jianqiang He^{3,5}

¹Jiangsu Key Laboratory of Agricultural Meteorology, Nanjing University of Information Science and Technology, Nanjing, China

²Yale-NUIST Center on Atmospheric Environment, International Joint Laboratory on Climate and Environment Change (ILCEC), Nanjing University of Information Science and Technology, Nanjing, China

³Key Laboratory for Agricultural Soil and Water Engineering in Arid Area of Ministry of Education, Northwest A&F University, Yangling, China

⁴National Meteorological Center, China Meteorological Administration, Beijing, China

⁵State Key Laboratory of Soil Erosion and Dryland Farming on the Loess Plateau, Institute of Water and Soil Conservation, Northwest A&F University, Yangling, China

⁶Key Laboratory of Eco-Environment and Meteorology for the Qinling Mountains and Loess Plateau, Shaanxi Meteorological Bureau, Xi'an, China

⁷State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

Correspondence

Jianqiang He, Key Laboratory for Agricultural Soil and Water Engineering in Arid Area of Ministry of Education, Northwest A&F University, Yangling 712100, China.
Email: jianqiang_he@nwsuaf.edu.cn

Funding information

Natural Science Foundation of China, Grant/Award Numbers: 41975143, 52079115, 42021004; Open Foundation of Jiangsu Key Laboratory of Agricultural Meteorology, Grant/Award Number: JKLAM2305; National Key R&D Program of China, Grant/Award Number: 2018YFB1500901; Key Research and Development Program of Shaanxi, Grant/Award Number: 2019ZDLNY07-03;

Abstract

Accurately estimated reference evapotranspiration (ET_0) is essential to regional water management. The FAO recommends coupling the Penman–Monteith (P-M) model with the Ångström–Prescott (A-P) formula as the standard method for ET_0 estimation with missing R_s measurements. However, its application is usually restricted by the two fundamental coefficients (a and b) of the A-P formula. This paper proposes a new method for estimating ET_0 with missing R_s by combining machine learning with physical-based P-M models (PM- ET_0). The benchmark values of the A-P coefficients were first determined at the daily, monthly, and yearly scales, and further evaluated in R_s and ET_0 estimates at 80 national R_s measuring stations. Then, three empirical models and four machine-learning methods were evaluated in estimating the A-P coefficients. Machine learning methods were also used to estimate ET_0 (ML- ET_0) to compare with the PM- ET_0 . Finally, the optimal estimation method was used to estimate the A-P coefficients for the 839 regular weather stations for ET_0 estimation without R_s measurement for China. The results demonstrated a descending trend for coefficient a from northwest to southeast China, with larger values in cold seasons. However, coefficient b showed the opposite distribution as the coefficient a . The FAO has recommended a larger a but a smaller b for southeast China, which produced the region's largest R_s and ET_0 estimation errors. Additionally, the A-P coefficients calibrated at the daily scale obtained the best estimation accuracy for both R_s and ET_0 , and slightly outperformed the monthly and yearly coefficients without significant difference in most cases. The machine learning methods outperformed the empirical methods for estimating the A-P coefficients, especially for the sites with extreme values. Further, ML- ET_0 outperformed the PM- ET_0 with yearly A-P coefficients but underperformed those with daily and monthly ones. This study indicates an exciting potential for combining machine learning with physical models for estimating ET_0 . However, we found that using the A-P coefficients with finer time scales is unnecessary to deal with the missing R_s measurements.

Wei Xiao, Jiandong Liu and Jianqiang He contributed equally to this study.

Startup Foundation for Introducing Talent of NUIST, Grant/Award Number: 2023r101

KEYWORDS

Ångström–Prescott formula, global solar radiation, machine learning, penman–Monteith model, reference evapotranspiration

1 | INTRODUCTION

Estimating reference crop evapotranspiration (ET_0) accurately is essential for regional water resource planning and irrigation scheduling (Shiri et al., 2014). However, measuring ET_0 directly through experiments is usually restricted by high labour and time consumption (Xing et al., 2022). The FAO56–recommended Penman–Monteith (P-M) model is widely adopted as the standard method for ET_0 estimation (Allen et al., 1998). Still, limited meteorological data usually restrict its application worldwide, especially global solar radiation (R_s) (Xing et al., 2023). Because of the cost of expensive measuring instruments, weather stations measuring R_s directly cover a limited number of locations worldwide, especially in developing countries (Jahani et al., 2017). Various weather datasets containing R_s values have been produced, including the NASA Prediction of Worldwide Energy Resource (NASA POWER) (Chandler et al., 2013), the ERA-Interim reanalysis products (Dee et al., 2011), and the Japanese Meteorological Agency (JRA-55) (Kobayashi et al., 2015). These datasets provide valuable references for the relevant research driven by R_s but usually have lag time and unsatisfactory performance for the R_s products. Hence, real-time R_s estimation is always needed to meet various application requirements.

Previous studies have established various estimation models for R_s , such as (1) the interpolation model with limited R_s observations (Rivington et al., 2006), (2) the empirical regression model between R_s and meteorological and geographic variables (Bristow & Campbell, 1984; Makadea & Jamil, 2018), (3) the satellite-based model with direct measurement of surface shortwave radiation (He et al., 2015), and (4) the radiation transfer models (Pawlak et al., 2004). Empirical methods for R_s estimation have been widely established based on different input variables considering the low data requirements and computation. These statistical models are divided into sunshine- (Ångström, 1924; Naserpour et al., 2020), cloud- (Badescu & Dumitrescu, 2015; Ehnberg & Bollen, 2005), and temperature-based models (Feng et al., 2019; Yacef et al., 2014). Studies have shown that sunshine-based models outperformed other models (Abdul-Aziz et al., 1993; De, Souza, et al., 2016; Trnka et al., 2005), especially the Ångström–Prescott (A-P) method (Prescott, 1940). The A-P formula is developed based on limited inputs and a simple linear relationship and thus has been widely used in ET_0 estimation with missing R_s data (Allen et al., 1998; Chen et al., 2022).

The default values recommended by the FAO-56 document (e.g., $a = 0.25$ and $b = 0.50$) have been widely used since data were always inadequate for calibrating these coefficients (Allen et al., 1998). Unfortunately, Sabziparvar et al. (2013) and Liu et al. (2014) have reported remarkable spatial variations in the A-P coefficients that can cause significant errors in regional and national R_s and

ET_0 estimates. Thus, calibrating the A-P coefficients with actual local R_s measurements is essential for minimizing R_s errors (Peng et al., 2022). In addition to the spatial variation, temporal fluctuation of the A-P coefficients has also attracted the attention of researchers (Liu, Mei, Li, Zhang, Wang, et al., 2009). No consensus on time-dependent or fixed A-P coefficients for R_s estimation has been reached. Soler (1990) found that monthly a and b coefficients outperformed yearly ones, and Tymvios et al. (2005) drew a similar conclusion. Further, other studies report that time-dependent coefficients obtained similar accuracy (Almorox & Hontoria, 2004; Ertekin & Evrendilek, 2007) or worse (Hussain et al., 1999) R_s estimation.

Additionally, the A-P coefficients are usually estimated through single- (Hassan et al., 2016) or multi-factor (Liu, Mei, Li, Wang, Jensen, et al., 2009) empirical regression models. These traditional models are usually inefficient in dealing with the complex nonlinear relationships between input factors and target variables (Kisi & Parmar, 2016). Further, input variables are usually set for a specific empirical model, which cannot always be done (Ming et al., 2015). Machine learning is widely adopted for R_s and ET_0 estimates due to the flexible combination of predictors and satisfactory accuracy in dealing with nonlinear problems (Fan et al., 2020; Gürel et al., 2020). However, considerable uncertainty can be generated in scenarios different from those in the training processes since machine learning usually performs as “black box” without a physical basis (Dong et al., 2022; Zhang et al., 2018).

Therefore, this study evaluated several empirical regression models and machine learning methods to estimate the A-P coefficients for ET_0 at different time scales with missing R_s data. The main objectives of this study were to (1) evaluate the spatiotemporal variation of the A-P coefficients, (2) assess the performance of time-dependent (daily and monthly) and fixed (yearly and FAO-recommended) A-P coefficients in daily R_s and ET_0 estimation, and (3) evaluate different empirical regression models and machine learning methods in estimating the A-P coefficients across all of China. The findings were expected to facilitate the application of the P-M and A-P models in ET_0 estimation for sites without R_s measurements in China.

2 | MATERIALS AND METHODS

2.1 | Study area

The four main climatic zones are the mountain plateau zone (MPZ), the temperate continental zone (TCZ), the temperate monsoon zone (TMZ), and the subtropical monsoon zone (SMZ) (Figure 1). The average elevations of the MPZ, TCZ, TMZ, and SMZ are 4236, 912, 288, and 611 m, respectively. There is a noticeable difference in mean

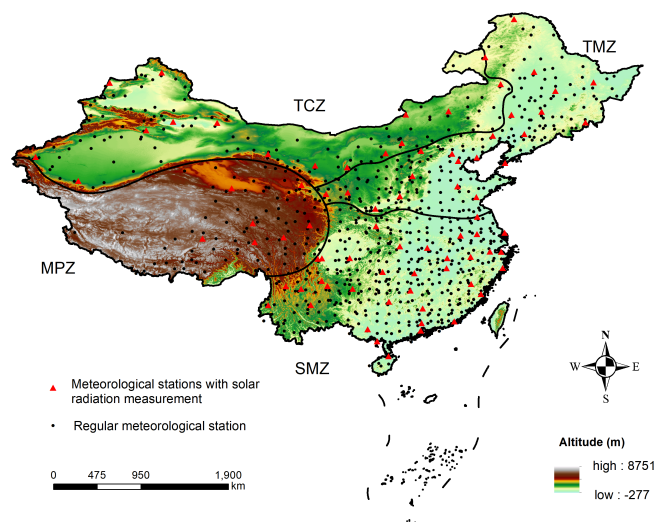


FIGURE 1 Study area and distribution of the 80 national solar radiation measuring stations (red crosses) and 839 weather stations without solar radiation measurements (black dots) in the four climatic zones of MPZ (the mountain plateau), TCZ (the temperate continental zone), TMZ (the temperate monsoon zone), and SMZ (the subtropical monsoon zone) in China. The acronyms are the same below.

annual precipitation for these four climatic zones. The most arid and humid zones are TCZ (193 mm) and the SMZ (1360 mm) in northwest and southeast China, while the TMZ (460 mm) and MPZ (591 mm) share a similar annual precipitation.

2.2 | Datasets

Daily measurements of R_s were collected from 80 R_s national measuring stations in China (Figure 1 and Table S1). There were 7, 16, 23, and 34 R_s stations in the MPZ, TCZ, TMZ, and SMZ zones, respectively. Daily meteorological measurements in this dataset including daily global solar radiation (R_s , MJ m^{-2}), sunshine hours (n , h), maximum (T_{\max} , $^{\circ}\text{C}$) and minimum air temperature (T_{\min} , $^{\circ}\text{C}$), average air temperature (T_{mean} , $^{\circ}\text{C}$), wind speed (U , m s^{-1}), precipitation (P , mm), and relative humidity (RH , %). Additionally, except for R_s , regular measurements were collected from 839 weather stations. The two datasets also contained each site's latitude, longitude, and altitude information and were obtained from the China Meteorological Administration (CMA). Weather measurements will be omitted for the day with the following two conditions: (1) missing measurements for either R_s or n ; (2) $R_s/R_a \geq 1$ or $n/N \geq 1$ (Persaud et al., 1997).

2.3 | Description of the Ångström–Prescott (A-P) formula and penman–Monteith model

Ångström (1924) established a linear relationship between R_s/R_{s0} and n/N in 1924, where R_{s0} represents the solar radiation on a sunny day, n represents the sunshine hour, and N represents the day length. We use Equation (1) to estimate the value of N :

$$N = 24 \times \omega_s / \pi \quad (1)$$

where ω_s is the sunset hour angle in rad. Prescott (1940) suggested replacing R_{s0} with R_a (extra-terrestrial radiation) due to the difficulty in measuring R_{s0} . In this way, the Ångström formula was translated into the Ångström–Prescott (A-P) format (Equation 2):

$$\frac{R_s}{R_a} = a + b \frac{n}{N} \quad (2)$$

where a (0–1) and b (0–1) are fundamental coefficients, and the sum of a and b is the transmissivity of clear sky. We further adopt Equation (3) to estimate R_a :

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

where G_{sc} represents the solar constant, $0.0820 \text{ MJ m}^{-2} \text{ min}^{-1}$; d_r is the inverse square of the Earth–Sun relative distance; φ is the latitude, rad; δ is the solar declination, rad.

The standard Penman–Monteith model (or P-M equation, Equation 4) has been recommended as the standard method for estimating ET_0 by the Food and Agriculture Organization of the United Nations (FAO) (Allen et al., 1998):

$$ET_0 = \frac{0.408 \Delta (R_n - G) + \gamma \frac{900}{T_{\text{mean}} + 273} u_2 (e_s - e_a)}{\Delta + \gamma (1 + 0.34 u_2)} \quad (4)$$

where R_n is the net radiation above the canopy, $\text{MJ m}^{-2} \text{ d}^{-1}$; G represents the soil heat flux density, $\text{MJ m}^{-2} \text{ d}^{-1}$; γ is the air psychrometric, $\text{kPa } ^{\circ}\text{C}^{-1}$; T_{mean} is the mean daily air temperature, $^{\circ}\text{C}$; u_2 represents the wind speed 2 m above ground in m s^{-1} ; e_s and e_a are the saturation and actual vapour pressures, kPa; Δ is the slope of the vapour pressure curve, $\text{kPa } ^{\circ}\text{C}^{-1}$.

2.4 | Generating the optimal time-scale A-P coefficients for R_s and ET_0

Previous studies have indicated the great spatiotemporal variations in the two fundamental coefficients (a and b) of the A-P formula. Hence, using fixed A-P coefficients can cause large errors in R_s estimation and further generate significant errors in ET_0 with the P-M model (PM- ET_0). However, local calibration of the coefficients a and b requires mass meteorological data, including R_s and other variables. Unfortunately, long-term measurements of R_s are always scarce for most areas worldwide. This study established a framework to generate the A-P coefficients with the optimal time scale for R_s and ET_0 estimation over all of China (Figure 2). First, the temporal (daily, monthly, and yearly) and spatial variations of the A-P coefficients were assessed based on the benchmark values of 80 solar radiation measuring stations. Then, the performance of the three different time-scale A-P coefficients was evaluated in calculating R_s and further estimating ET_0 . Finally, the empirical regression functions and machine learning algorithms were employed to estimate the A-P coefficients with

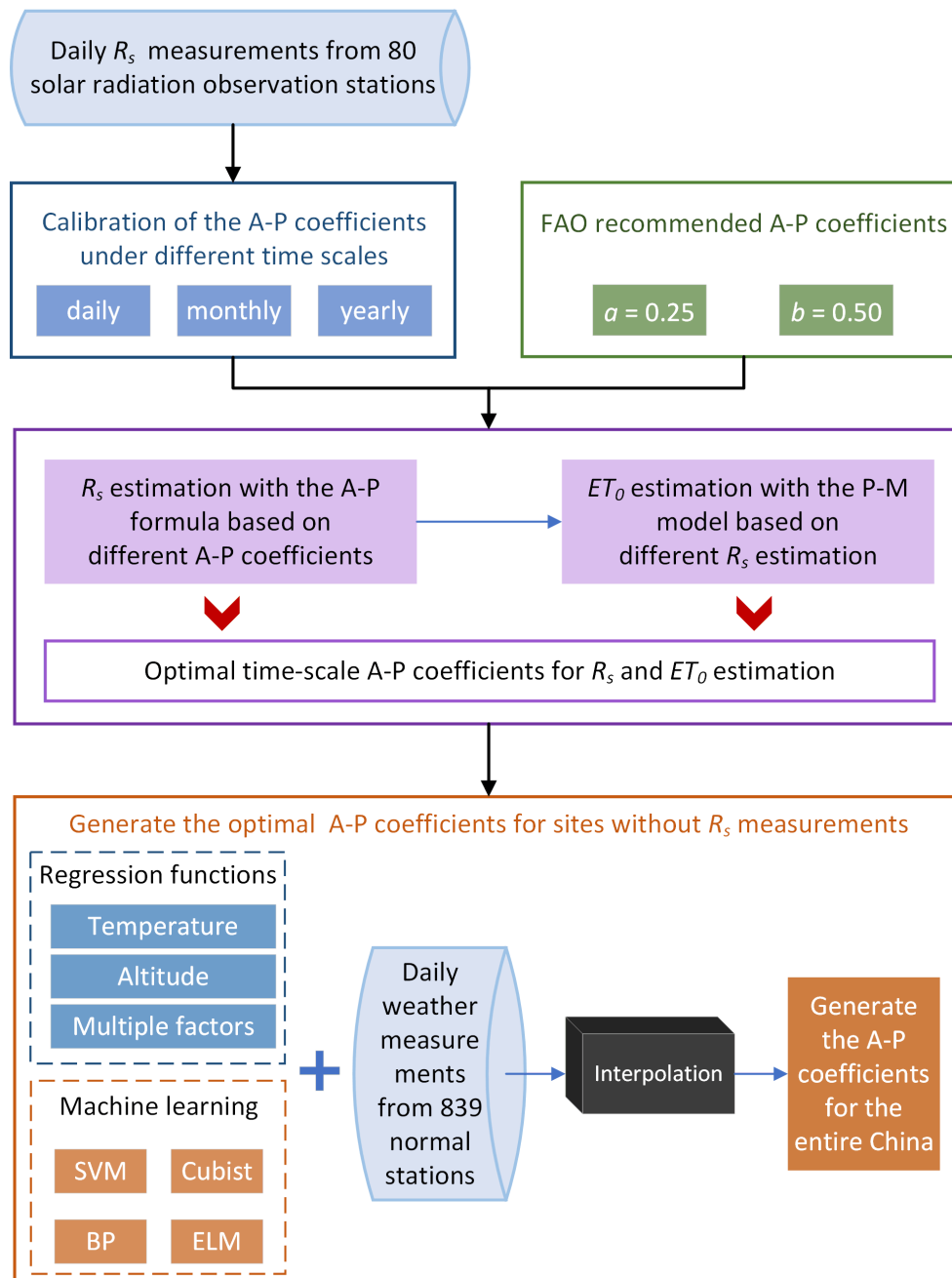


FIGURE 2 Flowchart of the generation of the optimal time-scale fundamental coefficients (a and b) of the Ångström-PreScott (A-P) formula in R_s and ET_0 estimation for the entire China. Estimation of ET_0 is conducted based on the Penman-Monteith (P-M) model. The SVM, Cubist, BP, and ELM represent four different machine learning methods of the support vector machine, Cubist model tree, back propagation neural network, and extreme learning machine, respectively.

common meteorological and geographic variables. The optimal method was selected to estimate the A-P coefficients of 839 regular weather stations without direct R_s measurements.

2.4.1 | Benchmark values of the A-P coefficients in different time scales

This study focused on the variations of yearly, monthly, and daily values of the A-P coefficients. First, the benchmark values of coefficients a and b were obtained for the three time scales based on least squares regression for each of the 80 R_s measuring stations. All of the R_s measurements were used to fit the regression function for the yearly A-P coefficients. For the monthly A-P coefficients, the R_s

measurements in the target month were used to fit the given month. The target-day R_s measurements were used to fit the A-P coefficients for the given day. Taking a station with 50 years of R_s measurements as an example, the number of data involved in the calibration of a and b were 18 250 (approx. $365 \text{ d} \times 50 \text{ a}$), 1500 (approx. $30 \text{ d} \times 50 \text{ a}$ for each month), and 50 (50 a for each day) for the yearly, monthly, and daily scale, respectively.

2.4.2 | Evaluating time-scale A-P coefficients for R_s and ET_0 estimation

The benchmark (daily, monthly, and yearly) and FAO-recommended values of the A-P coefficients were evaluated in the estimations of R_s

and ET_0 at the 80 R_s measuring stations. The estimates based on the four types of A-P coefficients above were compared directly with the daily measurements for R_s . For ET_0 , calculations based on the measured R_s and the P-M models (i.e., ET_0 benchmark) were accepted as the standard ET_0 values due to a lack of direct measurement. Then, the performance of the four A-P coefficients for ET_0 estimation was assessed according to the deviation between the estimates and benchmarks. Finally, the time scale of A-P coefficients with the most minor errors in R_s and ET_0 estimations was taken as the optimal time scale. The A-P coefficients with the relevant time scale were then estimated for all 839 regular weather stations.

2.4.3 | Estimating A-P coefficients with empirical regression functions

This study employed three empirical regression models and four machine-learning methods to estimate the A-P coefficients. Since the three empirical models can only estimate yearly (or time-invariant) A-P coefficients, the comparisons between empirical models and machine learning methods were only conducted at a yearly scale. For monthly and daily a and b , we evaluated the performances of different machine learning methods.

Empirical model I based on mean annual air temperature

Liu, Mei, Li, Wang, Zhang, and Porter (2009) established the estimation formula for the A-P coefficients based on the R_s records from 30 R_s measuring stations in the Yellow River Basin in northern China. This model employed only mean annual air temperature in the two-step model (Equations 5 and 6):

$$b = 4.33 \times 10^{-4} \cdot T^2 - 0.0126 \cdot T + 0.6289 \quad (5)$$

$$(a + b) = -9.66 \times 10^{-3} \cdot T + 0.8424 \quad (6)$$

where T represents the mean annual air temperature of the target site, °C.

Empirical model II based on altitude

Liu, Mei, Li, Wang, Zhang, and Porter (2009) also established another estimation function for a and b based only on site's altitude through a two-step method (Equations 7 and 8):

$$a = 1.57 \times 10^{-5} \cdot Z + 0.1705 \quad (7)$$

$$(a + b) = 3.58 \times 10^{-5} \cdot Z + 0.7121 \quad (8)$$

where Z is the altitude of the target site, m.

Empirical model III based on multiple variables

Liu et al. (2014) further modified the above formulas by introducing additional variables (Equations 9 and 10). The new functions were established and evaluated based on the R_s records from 80 R_s measuring stations:

$$a = 1.04 \times 10^{-5} \cdot Z + 0.1094 \cdot \left(\frac{n}{N}\right) - 7.64 \times 10^{-4} \cdot \phi + 0.1917 \quad (9)$$

$$(a + b) = 3.39 \times 10^{-5} \cdot Z - 0.13922 \cdot \cos \varphi + 0.0241 \cdot \cos \phi + 0.8349 \quad (10)$$

where Z is altitude, m; n is sunshine hours, n; N is day length, n; ϕ is longitude; φ is latitude.

2.4.4 | Estimating A-P coefficients with machine learning models

The steps for establishing machine-learning models for the estimating A-P coefficients include normalizing of input data, determining key parameters of machine learning models, training and testing models, and applying the models (Chen et al., 2022). Four machine learning methods were employed to estimate the A-P coefficients at daily, monthly, and yearly scales based on the variables same as the Empirical model III. The four machine learning algorithms were a back propagation (BP) neural network, a Cubist model tree, a support vector machine (SVM), and an extreme learning machine (ELM). The optimal machine-learning algorithm was then used to estimate the relevant time-scale A-P coefficients at the 839 weather stations without R_s measurements. The Kriging method was selected to interpolate the 839-site A-P coefficients (Peng et al., 2022). Additionally, the four machine learning algorithms were also directly used to establish the ET_0 estimation (ML- ET_0) models based on the above variables and been compared with the PM- ET_0 . The algorithms were coded in the R language (R Core Team, 2013).

BP neural network

The BP (backpropagation) neural network is one kind of feed-forward network for supervised learning (Rumelhart et al., 1986). This algorithm minimizes the error between the actual and expected outputs through backpropagation to modify the key parameters of the network (Tian et al., 2020). This study used the *nnet* package in the R language to establish the BP model. The key parameters of this algorithm include the number of hidden nodes (*size*) and the weight attenuation parameter (*decay*).

Cubist model tree

The Cubist model tree was established based on the M5 tree algorithm (Quinlan, 1992). Predictions with the Cubist were conducted based on the combinations of several successive piecewise models. This study used the *Cubist* package in the R language to establish the Cubist model. The critical parameters of this algorithm include the number of boosting iteration models (*committees*) and the number of instances used to correct the rule-based prediction (*neighbours*).

SVM algorithm

The SVM (support vector machine) was established based on the Vapnik (1996) construct of support vectors. The SVM can minimize the error by adding the hyperplane to maximize the margin between

prediction and observation (Drucker et al., 1997). The SVM is good at dealing with problems with small samples, nonlinearity, and high dimensionality and is extensively used in R_s estimation (Chen et al., 2013; He et al., 2020). This study adopted the *e1071* package (Karatzoglou et al., 2004) in the R language to establish the SVM model. The kernel function was set as *radial*. The critical parameters of this algorithm include the effect of a single sample on the entire classification hyperplane (*gamma*) and the cost of constraint violation (*cost*).

ELM algorithm

The ELM method is also a feed-forward neuron network algorithm but much faster due to its better generalization capability (Huang et al., 2006). This study used the *elmNNRcpp* package in the R language to establish the ELM model. The critical parameters of this algorithm include the number of hidden neurons (*nhid*), the type of activation function (*actfun*), and the distribution from which the input weights and the bias were initialized (*init_weights*).

2.5 | Statistical analysis

This study used three statistical indicators to evaluate the A-P coefficients estimated with different methods, including the coefficient of determination (R^2 , Equation 11), the root mean square error (RMSE, Equation 12), and the normalized root mean square error (*nRMSE*, Equation 13). A larger R^2 shows better model fitness and a smaller RMSE shows slight model deviation. The statistical *nRMSE*, originally the ratio between RMSE and the mean value of the relevant observations, was used to evaluate the estimation errors of different variables. A smaller *nRMSE* value usually indicates higher accuracy:

$$R^2 = \frac{\left[\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} \quad (11)$$

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (x_i - y_i)^2} \quad (12)$$

$$nRMSE = \frac{RMSE}{\bar{y}} \quad (13)$$

where m is the data sample size; x_i is the i th estimation value; y_i is the i th observed value; \bar{x} is the mean value of x_i ; \bar{y} is the mean value of y_i .

3 | RESULTS

3.1 | Benchmark values of the A-P coefficients for different time scales

The benchmark values of the a and b for the three time scales were obtained through linear regression at the 80 R_s measuring stations (Table 1) and Tables S1 and S2. Generally, the mean benchmark values of A-P coefficients were similar at the different time scales. The mean value ranges of a and b coefficients were 0.16–0.23 and 0.53–0.59, respectively. Compared with the FAO-recommended values ($a = 0.25$, $b = 0.50$), a large a and a small b are recommended for China. For coefficient a , the yearly benchmarks ranged from 0.12 to 0.29, while the daily benchmarks ranged from 0.01 to 0.61. For coefficient b , the yearly benchmarks ranged from 0.48 to 0.72, while the daily benchmarks ranged from 0.05 to 0.96. The largest mean values of the benchmarks were found in the MPZ zone for both the a and b . The smallest mean benchmark value was found in the SMZ zone for a and the TCZ zone for b .

Periodic variations were found in the benchmark values of a and b at daily and monthly scales at the four randomly selected representative stations (Station 55 299 in MPZ; Station 51 567 in TCZ; Station 54 511 in TMZ; Station 59 287 in SMZ) in China (Figure 3). For coefficient a , larger benchmark values were found in December, January,

TABLE 1 Maximum (Max), minimum (Min), mean values, and standard errors (SE) of the benchmark values of coefficient a and b of the Ångström-PreScott (A-P) formula in the four climatic zones in China.

Coefficient	Region	Site number	Yearly			Monthly			Daily		
			Max	Min	Mean ± SE	Max	Min	Mean ± SE	Max	Min	Mean ± SE
a	MPZ	7	0.29	0.20	0.23 ± 0.03	0.31	0.17	0.23 ± 0.08	0.47	0.01	0.23 ± 0.09
	TCZ	16	0.27	0.18	0.22 ± 0.02	0.38	0.14	0.22 ± 0.05	0.61	0.01	0.23 ± 0.07
	TMZ	23	0.28	0.14	0.19 ± 0.03	0.39	0.12	0.19 ± 0.04	0.57	0.02	0.20 ± 0.06
	SMZ	34	0.21	0.12	0.16 ± 0.02	0.27	0.10	0.16 ± 0.03	0.36	0.01	0.16 ± 0.04
	Total	80	0.29	0.12	0.19 ± 0.04	0.39	0.10	0.19 ± 0.05	0.61	0.01	0.19 ± 0.06
b	MPZ	7	0.66	0.54	0.59 ± 0.04	0.70	0.48	0.58 ± 0.05	0.93	0.29	0.58 ± 0.08
	TCZ	16	0.60	0.48	0.54 ± 0.04	0.64	0.35	0.53 ± 0.06	0.86	0.08	0.53 ± 0.09
	TMZ	23	0.59	0.49	0.54 ± 0.03	0.64	0.34	0.54 ± 0.06	0.86	0.05	0.53 ± 0.08
	SMZ	34	0.72	0.51	0.58 ± 0.04	0.78	0.45	0.57 ± 0.05	0.96	0.24	0.57 ± 0.07
	Total	80	0.72	0.48	0.56 ± 0.04	0.78	0.34	0.56 ± 0.06	0.96	0.05	0.55 ± 0.08

and February (winter) in the MPZ, TCZ, and TMZ zones. In contrast, smaller ones were found in June, July, and August (summer). Additionally, the temporal trends of b were contrary to a . Distributions of benchmark values of the two coefficients also varied among sites. Compared with the sites in the other three climatic zones, no noticeable temporal trend was found at Site 57 461 in the SMZ zone.

3.2 | Estimating R_s and ET_0 based on different time-scale A-P coefficients

3.2.1 | Estimating R_s

Superior performance in R_s estimation was obtained through the A-P formula with different values of the fundamental coefficients (Figure 4). The RMSE values of the four methods were all less than $3.1 \text{ MJ m}^{-2} \text{ d}^{-1}$, and R^2 values were more significant than 0.85. Additionally, all of the four methods overestimated the R_s in the interval of $1\text{--}10 \text{ MJ m}^{-2} \text{ d}^{-1}$. However, R_s was underestimated with the FAO-recommended A-P coefficients, which generated the most errors, especially under the high R_s conditions (Figure 4d). The maximum estimation of the FAO-recommended coefficients was $31.3 \text{ MJ m}^{-2} \text{ d}^{-1}$ while the maximum observation was $41.6 \text{ MJ m}^{-2} \text{ d}^{-1}$. Compared with the R_s estimated with the FAO-recommended coefficients, a noticeable improvement was generated using the site-calibrated A-P coefficients. Meanwhile, the estimation errors of R_s were decreased with the finer temporal resolution of the A-P coefficients. The best R_s estimations were produced with the daily A-P coefficients (Figure 4c).

3.2.2 | Estimating ET_0

The site-calibrated coefficients provided significant improvement in ET_0 estimation was over the FAO recommended coefficients of the A-P formula (Figure 5a). The mean RMSEs were 0.34, 0.28, 0.29, and 0.30 mm d^{-1} in ET_0 estimation using the FAO recommended, daily, monthly, and yearly calibrated A-P coefficients, respectively. The site-calibrated coefficients improved the estimation accuracy significantly compared to the FAO-recommended coefficients, especially in the SMZ. Additionally, there was an insignificant decreasing trend in errors for the site-calibrated A-P coefficients from yearly to daily scales. Further, an apparent spatial gradient of estimation errors increased from the north to the south (Figure 5d,e). The most remarkable errors were found in the SMZ for all four groups of coefficients, especially the FAO-recommended coefficients. Seven out of the 80 stations had RMSEs greater than 0.5 mm d^{-1} ; the largest RMSE was greater than 0.65 mm d^{-1} . These three site-calibrated coefficients provided better estimations for most sites than the FAO-recommended coefficients did. The daily A-P coefficients achieved the most considerable improvement in ET_0 estimation in the SMZ with all sites' RMSE less than 0.5 mm d^{-1} .

The four different A-P coefficients in ET_0 estimation were evaluated in different months (Figure 6). Generally, the FAO-recommended site-calibrated coefficients produced the largest errors in ET_0 in all

months. Compared with the FAO-recommended coefficients, the site-calibrated coefficients significantly improved estimation accuracy throughout the year, especially from April to September. The best estimates were provided through the R_s obtained with the daily A-P coefficients. However, these three site-calibrated coefficients in ET_0 estimation had almost no significant difference. Additionally, larger estimation errors were found in hot seasons (the mean RMSE $>0.3 \text{ mm d}^{-1}$) than in cold seasons (the mean RMSE $<0.2 \text{ mm d}^{-1}$).

Generally, the ET_0 estimated with the PM- ET_0 models showed greater variations (Table 2). Using the P-M model combining daily A-P coefficients provided the best ET_0 estimations among these eight models. However, the ET_0 estimations based on the FAO recommended A-P coefficients obtained the largest estimation errors. Compared with the PM-based methods, the four ML- ET_0 methods obtained similar accuracies in ET_0 estimation, the RMSEs were all greater than 0.97, RMSEs less than 0.31 mm d^{-1} , and $nRMSEs$ less than 12%. Notably, the PM- ET_0 model with the yearly A-P coefficients obtained similar estimation accuracy as the machine-learning models. The difference of RMSE and $nRMSE$ were less than 0.003 mm d^{-1} and 0.3% between the PM- ET_0 with yearly coefficients and the ML- ET_0 model with SVM. The results proved that the P-M model combined with site-special A-P coefficients was potential for ET_0 estimation due to acceptable accuracy and convenience.

3.3 | Optimal methods for estimating the A-P coefficients

3.3.1 | Yearly A-P coefficients

Generally, the machine learning methods outperformed the empirical models in estimating both a and b (Figure 7). The $nRMSEs$ were 16.8%–23.6% for a and 6.5%–8.8% for b . The empirical models underestimated a but overestimated b at most sites (Figure 7a,c). Empirical model III provided similar estimate distribution as benchmarks and produced the smallest errors. However, the three empirical models performed poorly at the sites with extreme benchmark values. Compared with the empirical models, all four machine learning methods obtained good estimates of the a and b coefficients, especially for a (Figure 7b,d). The $nRMSEs$ were 10.2%–11.3% for a and 4.8%–5.6% for b . Hence, the machine learning methods were more accurate and promising for estimating the a and b coefficients of the A-P formula. Among the four machine learning methods, the SVM and Cubist provided the most negligible errors for the a and b , respectively.

3.3.2 | Monthly A-P coefficients

Compared with the yearly values of the a and b coefficients, large errors were obtained in the estimates of monthly values at the training stages (Table 3). The $nRMSEs$ were 3.8%–8.4% for a and 3.5%–6.8% for b . The highest estimation accuracy was provided in the MPZ, which was the same as the yearly coefficients. Larger errors arose in the TCZ and TMZ zones. The optimal machine learning methods were

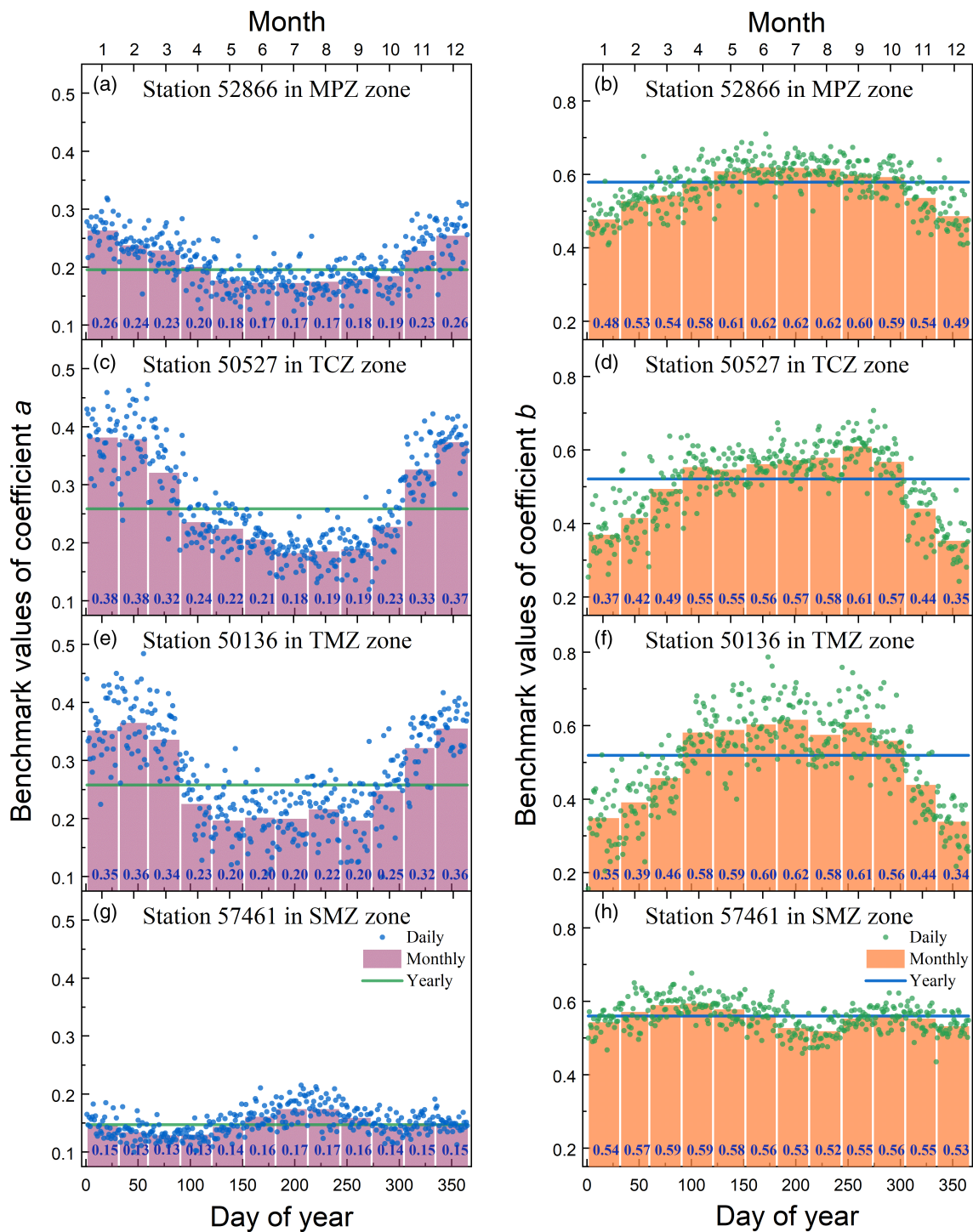


FIGURE 3 Benchmark values of the a and b coefficients of the Ångström-Prescott (A-P) formula at the daily (dots), monthly (bars), and yearly (horizontal lines) scales at four representative sites of each climatic zones (Station 52 866 in MPZ; Station 50 527 in TCZ; Station 50 136 in TMZ; Station 57 461 in SMZ).

also different in the estimating the A-P coefficients. For a , the Cubist method outperformed the other three methods in two climatic zones and all of China. The corresponding $RMSEs$ and $nRMSEs$ were 0.008–0.014 and 3.8%–8.6%. However, for b , the BP method outperformed the other three methods at three zones and all of China. The corresponding $RMSEs$ and $nRMSEs$ values were 0.018–0.026 and 3.1%–5.0%.

3.4 | Estimating the A-P coefficient without R_s measurement

3.4.1 | Generating the yearly A-P coefficients

The yearly values of the A-P coefficients were estimated for the 839 regular weather stations with the selected optimal machine

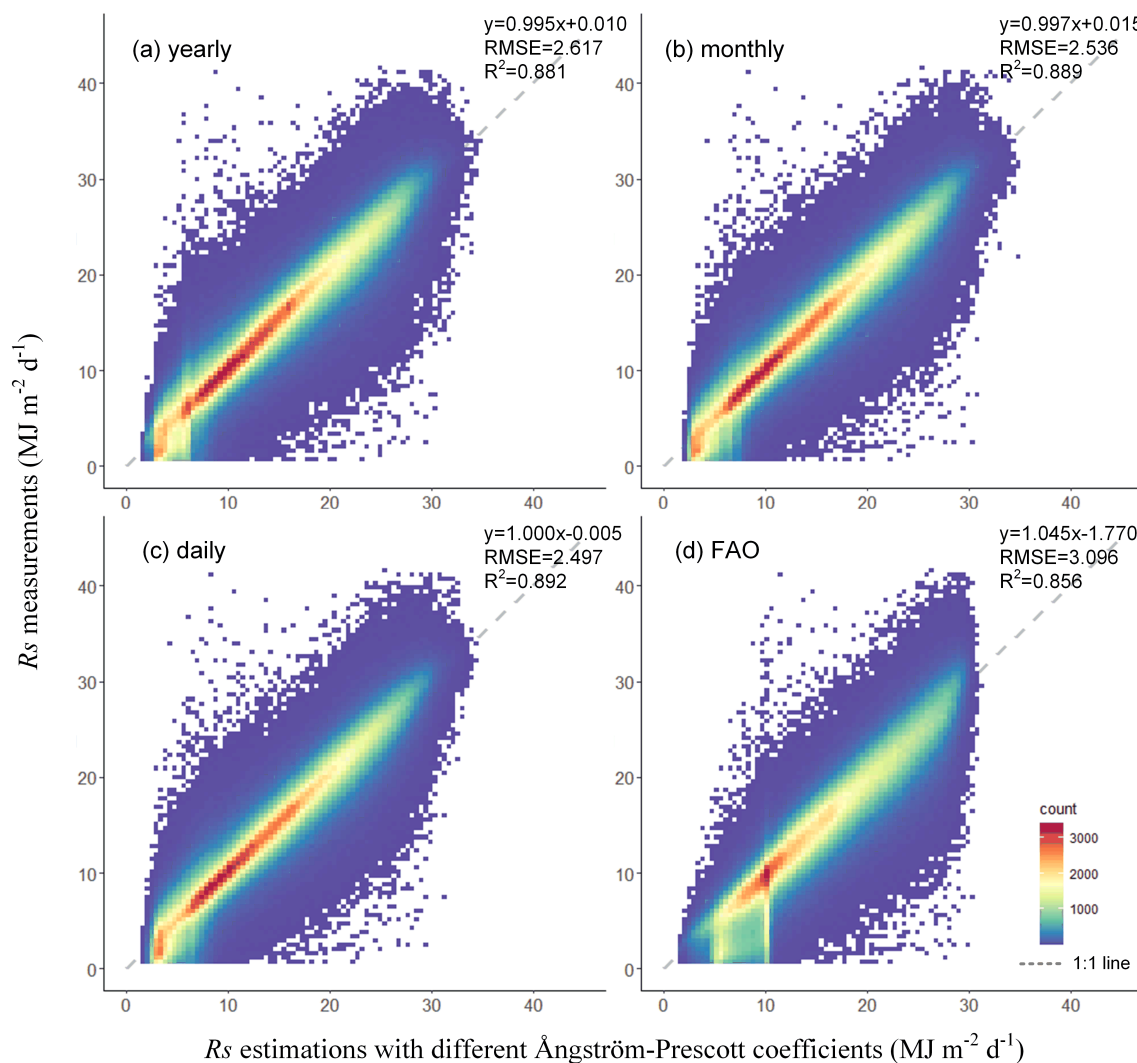


FIGURE 4 Estimation of daily global solar radiation (R_s) with the yearly (a), month (b), and daily (c) scale calibration and the FAO recommendation (d) of the two fundamental coefficients of the Ångström-Prescott formula at the 80 R_s measuring stations in China.

learning method of SVM (Figure 8 and Tables S1 and S2). Compared with the site-calibrated values (Table 1), the machine learning methods provided similar ranges for a and b . Generally, remarkable spatial variations were found for the two coefficients in the four different climatic zones. Larger coefficient values were mainly found in the MPZ and TCM, while smaller values were obtained in the SMZ and TMZ. The estimate distributions of a revealed a positive correlation with altitude. The coefficient b showed an opposite distribution to a . The largest estimated b was found in the MPZ, followed by the SMZ, TMZ, and TCZ.

3.4.2 | Generating the monthly A-P coefficients

The monthly coefficient of a was estimated with the optimal machine learning method of Cubist for the 839 regular weather stations and further interpolated over all of China (Figure 9; Table S3). For a , larger values were found in northern and western China. In comparison, smaller values were found in southeast China (especially in the SMZ). There

were also noticeable temporal variations of the coefficient a in different months. Extreme values were in the cold season (Figure 9a,b,i), which resulted in the largest ranges of a . Further, the smallest spatial variation for a was in September and October (Figure 9i,j).

The values of monthly b were generated with the optimal machine learning method of BP (Figure 10; Table S4). Generally, b showed the opposite spatiotemporal distributions to a . Larger values were found in south eastern China during cold seasons (Figure 10a,b,i). However, the spatial patterns of b were similar to a in September and October since the coefficient varied over a smaller range than in other months.

3.5 | ET_0 estimation without R_s measurements

Daily R_s values, calculated through the A-P formula based on the FAO-recommended and the machine learning-based coefficients, were compared at the 839 regular weather stations without direct R_s measurement (Figure 11). In general, both types of A-P coefficients produced similar distributions of ET_0 in different meteorological

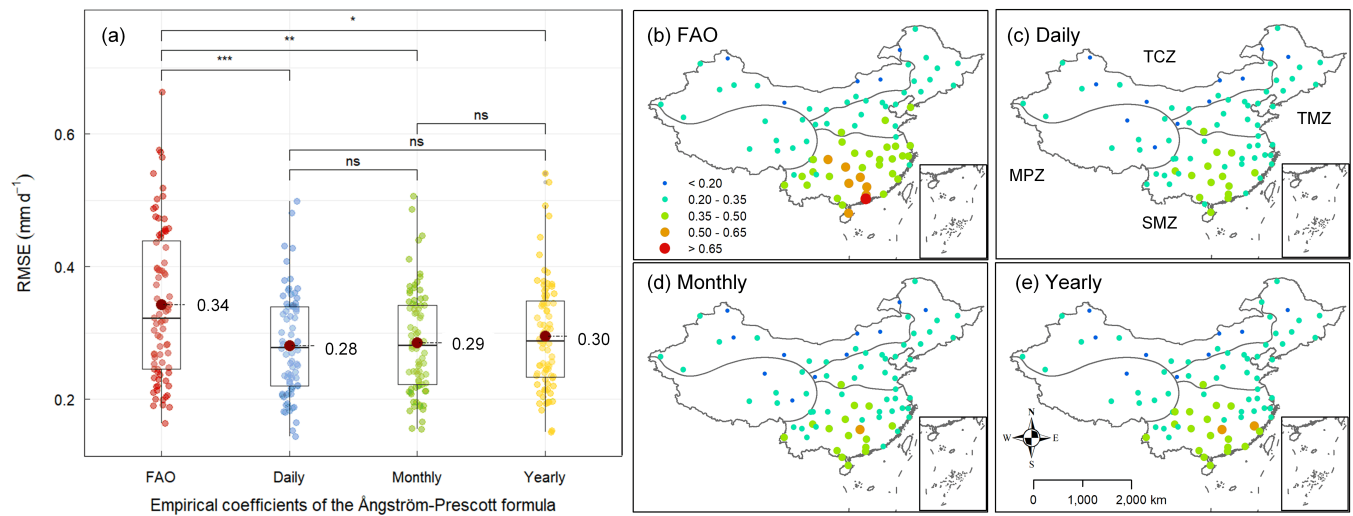


FIGURE 5 RMSE for ET_0 estimation based on the site-calibrated and FAO-recommended values of the fundamental coefficients of the Ångström-Prešcott (A-P) formula. The A-P coefficients were calibrated at the daily, monthly, and yearly scales. The dark red dot shows the mean value of each box. The symbol * shows the significance level, and ns represents no significance.

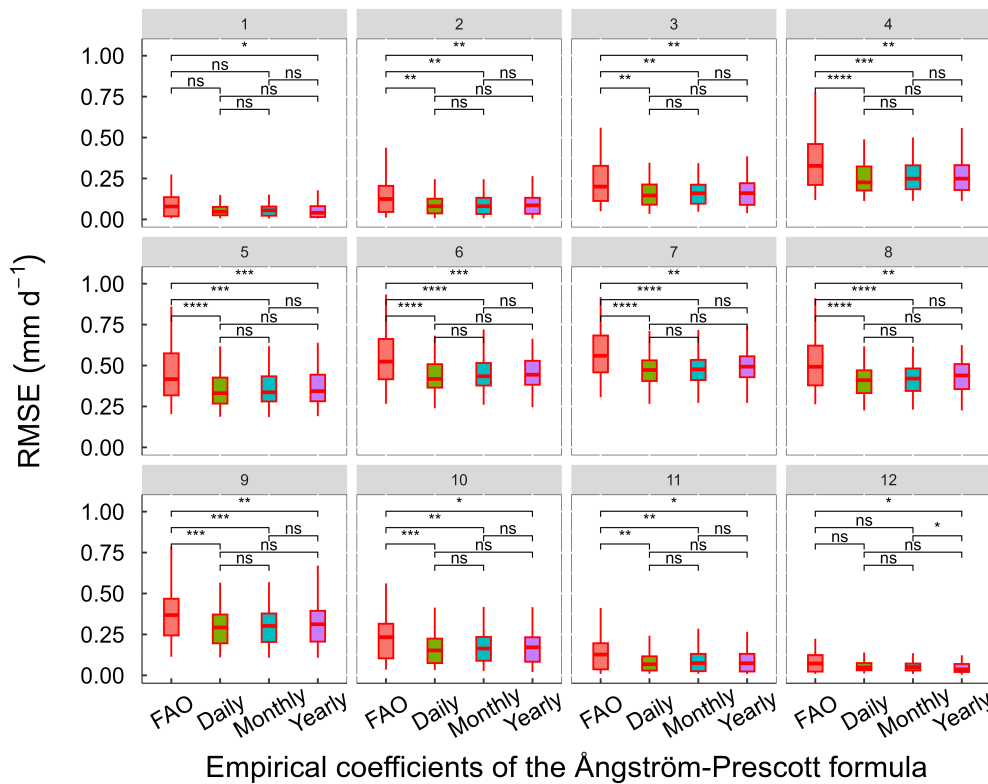


FIGURE 6 Temporal variation of RMSE for ET_0 estimations based on the estimated R_s and the Penman-Monteith model. The R_s is estimated through the Ångström-Prešcott formula with the daily, monthly, and yearly coefficients. The number above each subfigure shows the month index. The symbol * shows the significance level, and ns represents no significance.

zones. Difference of ET_0 estimated with these A-P coefficients in the TCZ and SMZ was more significant than those in the other two meteorological zones. However, there were still differences between these two kinds of A-P coefficients in ET_0 estimation in the four zones. Compared with the machine learning-based coefficients, ET_0 was slightly overestimated in MPZ and TCZ while being significantly underestimated in TMZ and SMZ by the FAO-recommended coefficients. Notably, there was almost no difference between the two groups of ET_0 estimated in the TCZ.

4 | DISCUSSION

4.1 | Spatiotemporal variations of A-P coefficients

The A-P formula has been extensively applied for daily R_s estimation because of the simple data requirements and acceptable performance (Li et al., 2012). However, some earlier researchers found that the two fundamental coefficients of the A-P formula were site-dependent (Chen et al., 2006; Jin et al., 2005). Hence, these coefficients must be calibrated

locally to guarantee satisfactory accuracy for daily R_s estimation. In this study, the linear regression method calibrated different time-scale benchmarks of the A-P coefficients at 80 national R_s measuring stations. Larger

TABLE 2 Coefficient of determination (R^2), root mean square error (RMSE, mm d^{-1}), and normalized root mean square error (nRMSE, mm d^{-1}) of ET_0 estimations with BP, Cubist, ELM, and SVM. The grey-shaded cells show the approach obtains the highest estimation accuracy.

ET_0 category	Methods	R^2	RMSE (mm d^{-1})	nRMSE (%)
PM- ET_0	Daily	0.974	0.287	10.9
	Monthly	0.972	0.293	11.1
	Yearly	0.970	0.305	11.6
	FAO	0.961	0.375	14.2
ML- ET_0	BP	0.972	0.309	11.6
	Cubist	0.973	0.305	11.4
	ELM	0.972	0.306	11.4
	SVM	0.973	0.302	11.3

Note: The PM- ET_0 represents the Penman-Monteith model derived by the R_s estimated with the Ångström-Prescott formula. The ML- ET_0 represents the ET_0 estimated directly with machine learning.

mean a benchmarks were found in the MPZ and TCZ zones with a dry climate and high altitude, while smaller mean a benchmarks were found in the SMZ zone with a wet environment and low altitude.

Coefficient a varies with the season. Larger values of a were found in winter (dry), while b showed the opposite trend in China. Adaramola (2012) reported that the A-P coefficients varied with geographical location and weather conditions. Liu et al. (2012) calibrated the benchmarks for the A-P coefficients and obtained similar distributions of the variable values as in this study. They also pointed out that a and the sum of a and b correlate with altitude positively. This study showed that the benchmarks for a were mainly less than 0.25 (mean value is 0.19), while those for b was notably greater than 0.50 (mean value is 0.56) (Table 1). Liu et al. (2014) calibrated the A-P coefficients at different sites across China and found that the mean values were 0.18 and 0.56 for coefficients a and b , respectively. Hence, there are apparent differences between the calibrations and the FAO recommendation, which can cause significant errors in the estimation of daily R_s (Figure 4). Additionally, the largest differences between the A-P coefficients recommended by the FAO and those calibrated in this study were in the SMZ. Xia et al. (2021) pointed out that the FAO recommendations are advised in northern China, including the northeast, North China Plain, and the Loess Plateau. Liu, Mei, Li, Wang, Zhang, and Porter (2009) found that the FAO

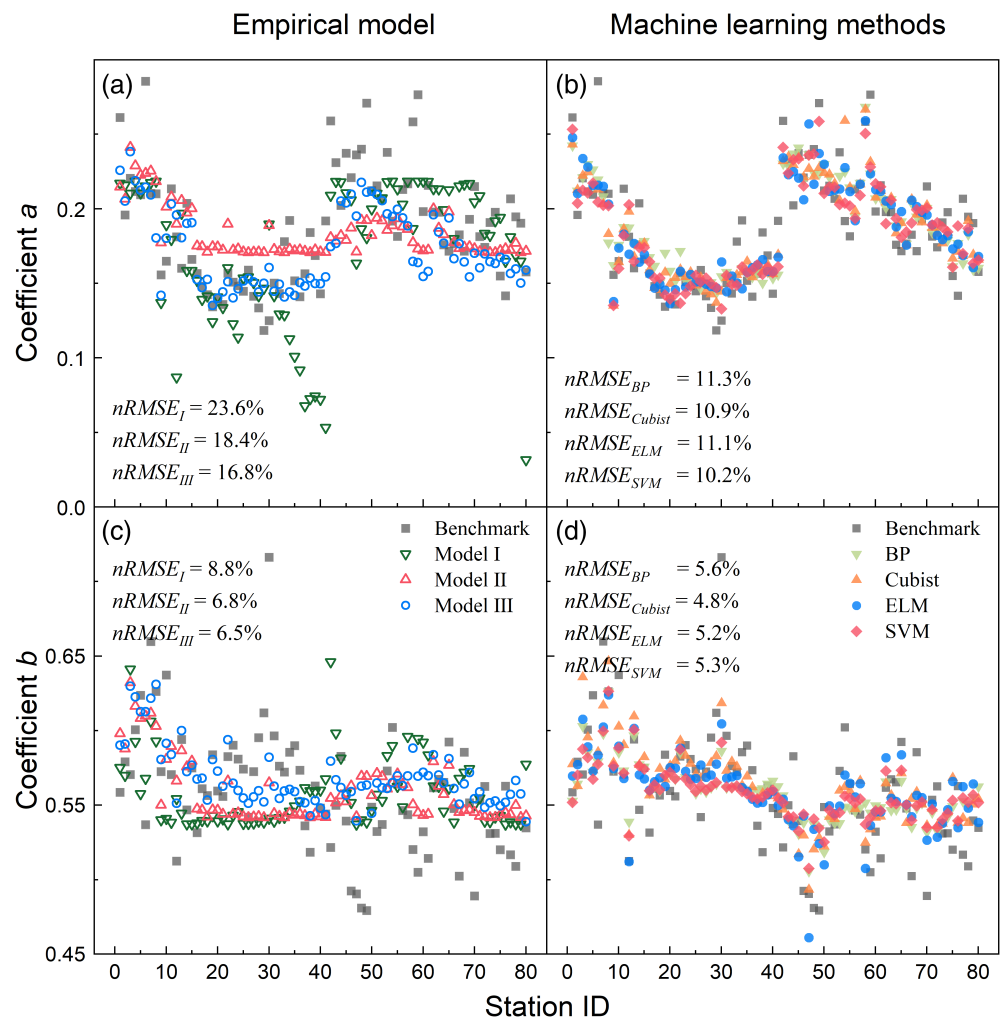


FIGURE 7 Comparisons between the a and b coefficients of the Ångström-Prescott (A-P) formula estimated with three empirical models (Model I, II, and III; a, c) and four machine learning methods (BP, Cubist, ELM, and SVM; b, d) at the 80 R_s measuring stations.

Region	MLM	Coefficient <i>a</i>			Coefficient <i>b</i>		
		R^2	RMSE	nRMSE (%)	R^2	RMSE	nRMSE (%)
MPZ	BP	0.916	0.008	3.8	0.791	0.021	3.6
	Cubist	0.877	0.011	5.1	0.793	0.021	3.6
	ELM	0.833	0.012	5.5	0.823	0.020	3.5
	SVM	0.692	0.017	7.6	0.721	0.024	4.2
TCZ	BP	0.845	0.020	8.7	0.832	0.026	5.0
	Cubist	0.773	0.023	10.4	0.721	0.034	6.5
	ELM	0.791	0.023	10.2	0.550	0.044	8.5
	SVM	0.853	0.019	8.6	0.702	0.035	6.8
TMZ	BP	0.755	0.024	11.9	0.871	0.021	4.0
	Cubist	0.925	0.013	6.5	0.827	0.025	4.6
	ELM	0.818	0.020	10.3	0.859	0.023	4.2
	SVM	0.836	0.019	9.7	0.792	0.027	5.1
SMZ	BP	0.854	0.011	7.2	0.877	0.018	3.1
	Cubist	0.889	0.009	6.3	0.880	0.018	4.1
	ELM	0.755	0.014	9.5	0.774	0.024	4.2
	SVM	0.881	0.010	6.4	0.829	0.021	3.6
National	BP	0.881	0.017	9.1	0.879	0.021	3.8
	Cubist	0.914	0.014	7.7	0.842	0.024	4.3
	ELM	0.868	0.018	9.7	0.778	0.029	5.2
	SVM	0.898	0.016	8.4	0.808	0.026	4.8

TABLE 3 Evaluation of the four machine learning methods (MLM) in the estimations of monthly *a* and *b* coefficients of the Ångström-Prešcott (A-P) formula in the four climatic zones and entire China. The grey-shaded cells show the approach obtains the highest estimation accuracy.

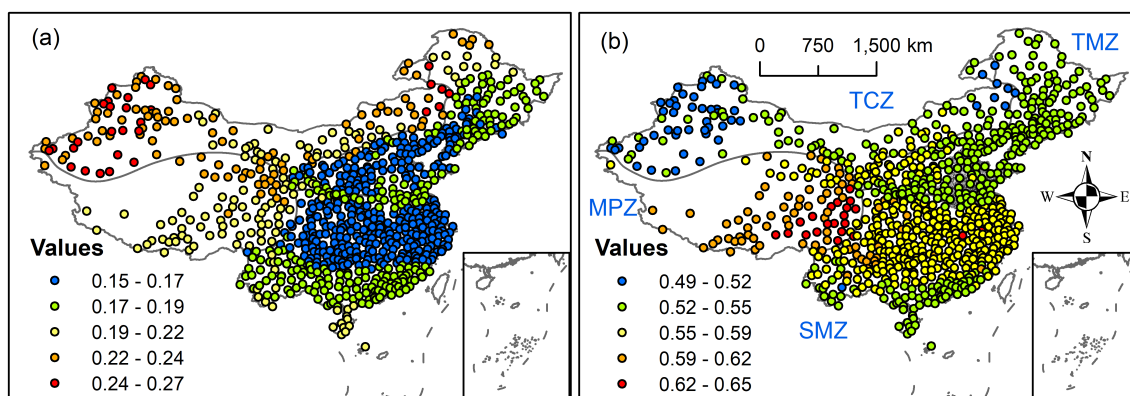


FIGURE 8 Estimation of yearly-scale coefficients *a* (a) and *b* (b) of the Ångström-Prešcott (A-P) formula with the optimal machine learning algorithm at the 839 national weather stations. The optimal machine learning algorithms selected for estimating the coefficients *a* and *b* are SVM and Cubist, respectively.

recommendations can provide acceptable R_s estimations where the altitude exceeds 1928 m. Thus, worse ET_0 estimates were found in the SMZ in south eastern China with lower altitudes when using the FAO-recommended values (Figure 5).

4.2 | Empirical and machine learning methods in A-P coefficient estimation

Machine learning is widely used for R_s and ET_0 estimations due to the flexible combination of predictors and satisfactory accuracy in dealing

with nonlinear problems (Fan et al., 2020; Gürel et al., 2020). However, machine learning works as a “black-box” model and usually becomes overfitted (Zhang et al., 2018). The performance of machine learning models is restricted by data training (Dong et al., 2022). Applying machine learning models would be inefficient when the scenario does not appear in the training dataset (Shu et al., 2022). Hence, the parameterization of the P-M model with machine learning can take advantage of the generalization ability of machine learning and the physical basis of the P-M models. This study used optimal machine learning algorithms to estimate the A-P coefficients of stations without R_s measurements. The coefficients were further used to

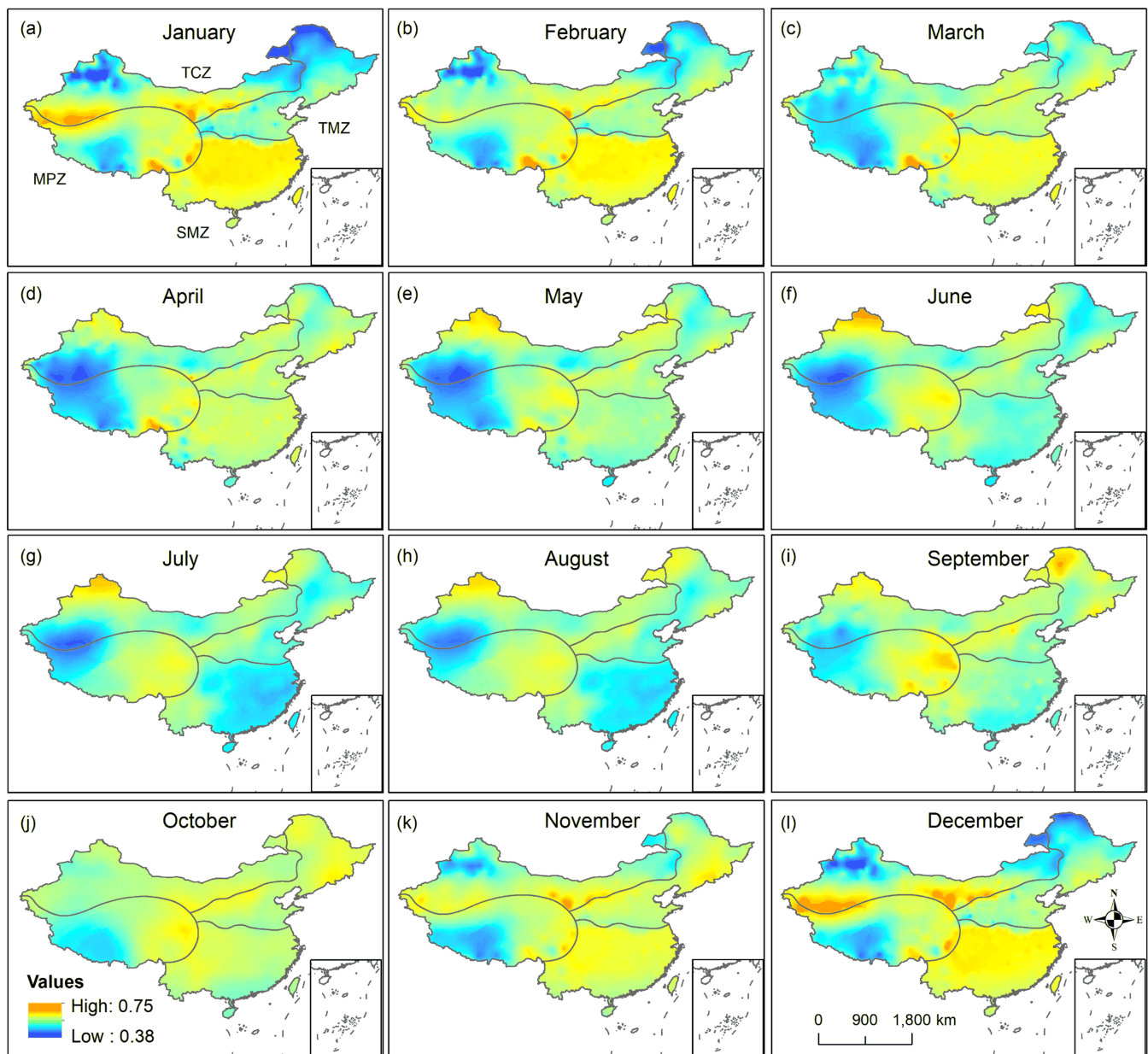


FIGURE 9 Spatial distributions of monthly coefficient a of the Ångström-Prescott (A-P) formula. The distributions are generated with the Kriging interpolation method based on the estimated coefficients a at 839 national weather stations through the Cubist.

calculate R_s to drive the P-M model to estimate the daily ET_0 . Satisfactory accuracy was obtained, indicating an exciting potential for combining machine learning with physical-process models for geoscientific predictions.

We compared the empirical and machine learning methods in estimating the A-P coefficients. In general, the machine learning methods outperformed the empirical models, especially for the stations with extreme a and b values (Figure 7). Large variations were found in the estimates of a and b among the empirical models, despite they were specially established for China. As in this study, the three empirical models published by Liu et al. (2009 and 2014) were established based on meteorological measurements over different periods. Hence, recalibrating the fundamental coefficients of the three

empirical models may better estimate the A-P coefficients. Generally, the machine learning methods could better estimate the a and b coefficients than the empirical models.

The machine learning methods performed exceptionally well at sites with extreme values of a and b . Fan et al. (2020) pointed out that machine learning methods take advantage of dealing with nonlinear problems with dispensable prior knowledge. Hence, machine learning methods have been widely used in direct estimation of solar radiation (Lu et al., 2023; Wang et al., 2017) and evapotranspiration (Niu et al., 2021). We evaluated the ET_0 estimated with machine learning methods and the PM-AP model with missing R_s measurements. The results indicated that using R_s estimations based on the daily and monthly A-P coefficients outperformed machine learning models in

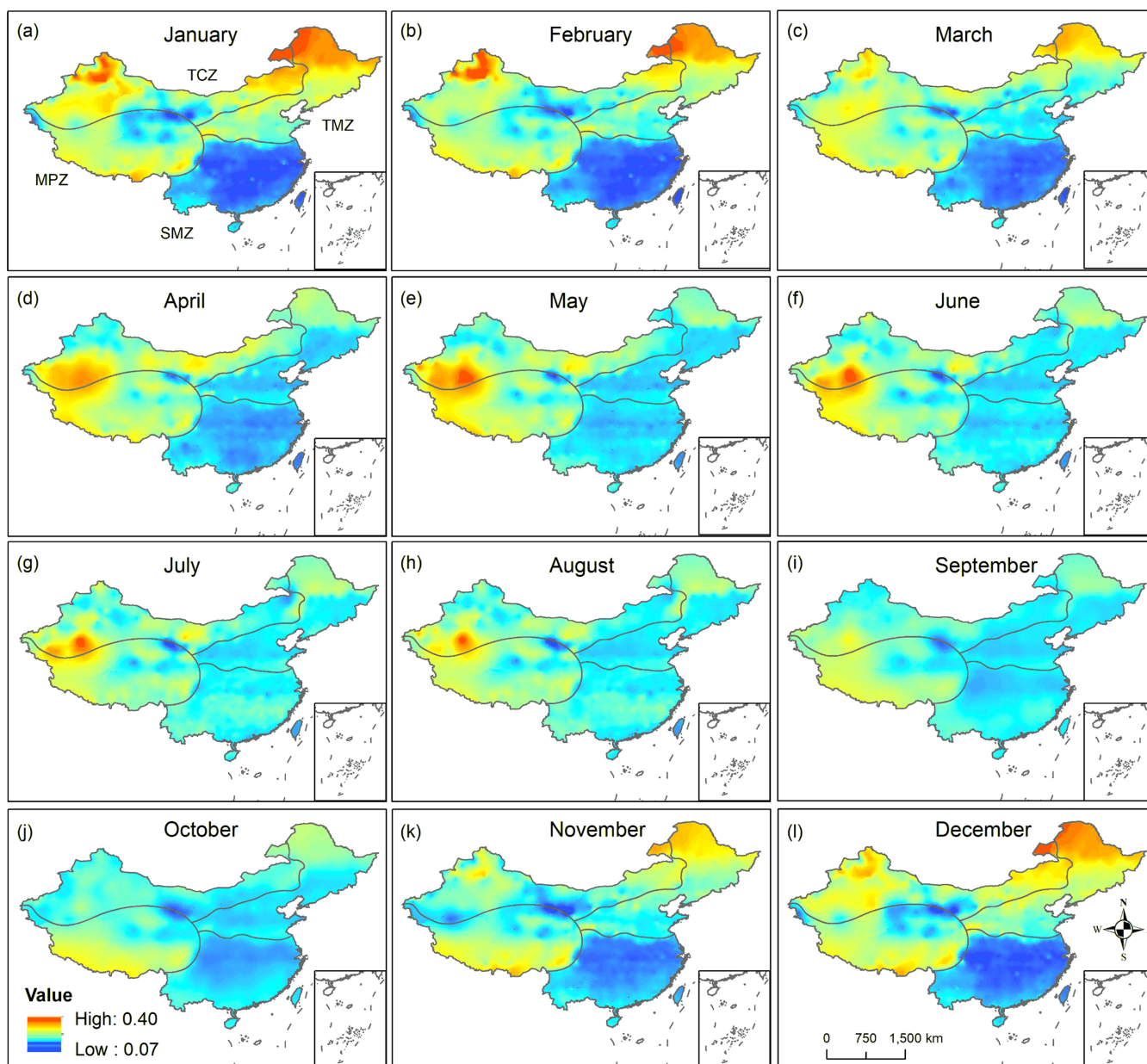


FIGURE 10 Spatial distributions of monthly coefficient b of the Ångström-Prescott (A-P) formula. The distributions are generated with the Kriging interpolation method based on the estimated values of coefficients b at 839 national weather stations through the BP.

ET_0 estimation (Table 2). Since the estimations of the A-P coefficients have been produced in this study, the PM- ET_0 model is recommended for ET_0 estimation in China for researchers who are not good at computer programming when only R_s data are missing.

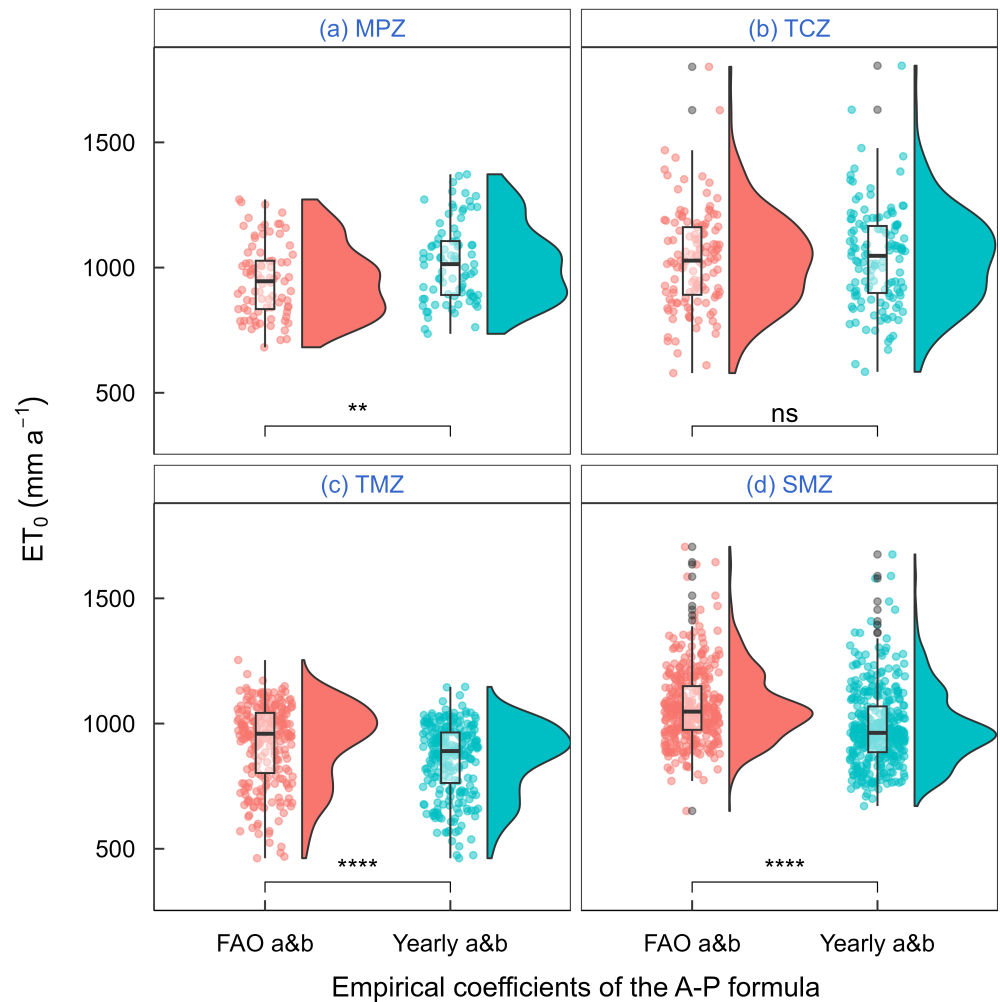
However, the A-P coefficients were estimated with the machine learning models that established based on observations from sparsely distributed stations. For example, there were only 7 and 16 stations in the MPZ and TCZ, respectively. Thus, the machine learning models and the A-P coefficient interpolation require further evaluation in these zones. Additionally, only four common machine learning algorithms were employed in this study. Searching for more efficient machine learning models for the parameterization of physical models is promising for geoscientific predictions (Nearing et al., 2021). In the

further, we will investigate new machine learning and deep learning methods in estimating R_s and ET_0 .

4.3 | Different time-scale A-P coefficients in R_s and ET_0 estimation

Great variation was found among the different time-scale A-P coefficient benchmark values (Table 1). The ranges of daily A-P coefficients were more extensive than their yearly ones, which may be caused by the data size used for linear regression (Liu, Mei, Li, Zhang, Wang, et al., 2009). The data size used for calibrating the values of A-P coefficients decreased substantially with a temporal gradient from yearly

FIGURE 11 Estimation of ET_0 for the 839 weather stations without direct R_s measurements in the four meteorological zones in China. Daily R_s used to drive the P-M model is calculated based on the Ångström-PreScott formula with the FAO-recommended coefficients (red) and yearly estimates based on the optimal machine-learning model (blue). The symbol * shows the significance level, and ns represents no significance.



to daily scales. Compared with R_s estimates with monthly and yearly A-P coefficients (both benchmarks and FAO recommendation), the R_s errors caused due to daily A-P coefficients were the smallest across China. Soler (1990) used monthly values in R_s estimation and obtained better results in Europe. Podestá et al. (2004) pointed out that seasonal or finer time-scale A-P coefficients should be used to eliminate the systematic residuals of R_s estimation. However, the daily A-P coefficients slightly outperformed the monthly and yearly coefficients. The errors of R_s were further transferred to the ET_0 estimation through the P-M model. The ET_0 estimation driven by the R_s estimated with the daily A-P coefficients provided the highest accuracy. Moreover, there was a nonsignificant difference among these three time-scale A-P coefficients in ET_0 estimation. Due to their convenience and accuracy, the yearly A-P coefficients are more suitable for estimating of daily R_s and ET_0 in China.

5 | CONCLUSION

This study proposes a new method for estimating ET_0 with missing R_s by combining machine learning with physical-based P-M models. Firstly, four machine learning algorithms were used to parameterize the

fundamental coefficients of the A-P formula at different time scales (daily, monthly, and yearly). Then, the estimated A-P coefficients were used to calculate R_s to drive the P-M model for ET_0 estimation (PM- ET_0). Additionally, ET_0 values were also directly estimated with machine learning methods (ML- ET_0). Values of the a descended from northwest to southeast China while the b shared an opposite distribution. The two coefficients were more scattered in the daily scale, followed by the monthly and yearly scales. Compared with the benchmarks, the FAO recommended a large a , but a small b for most stations in China, which resulted in the largest errors both in R_s and ET_0 , especially in south eastern China. Estimation of ET_0 with the FAO-recommended coefficients significantly improved by all of the site-calibrated A-P coefficients at different time scales, especially during the growing season from April to September. However, an insignificant decreasing gradient was found in ET_0 estimation with the A-P coefficients from the yearly to daily scales. Compared with the PM- ET_0 , the ML- ET_0 outperformed the ET_0 estimated with yearly A-P coefficients but underperformed those estimated with daily and monthly ones. Further machine learning methods were more reliable in estimating the A-P coefficients than empirical regression methods. Hence, in terms of robustness and convenience, using the A-P formula with yearly A-P coefficients to calculate R_s to drive the P-M model for ET_0 estimation is superior in China.

ACKNOWLEDGEMENTS

The authors thank the precious suggestions by anonymous reviewers and editors, which have greatly helped the improvement of the paper. This research was supported by the Natural Science Foundation of China (42021004), the Open Foundation of Jiangsu Key Laboratory of Agricultural Meteorology (JKLAM2305), the National Key R&D Program of China (No. 2018YFB1500901), the Natural Science Foundation of China (52079115, 41975143), the Key Research and Development Program of Shaanxi (2019ZDLNY07-03), and the Startup Foundation for Introducing Talent of NUIST (2023r101).

DATA AVAILABILITY STATEMENT

The data, materials, and codes of this article can be obtained by contacting the corresponding authors at jianqiang_he@nwsuaf.edu.cn.

ORCID

Shang Chen  <https://orcid.org/0000-0001-9307-5666>

Wenzhe Feng  <https://orcid.org/0009-0006-5011-5166>

REFERENCES

- Abdul-Aziz, J., A-Nagi, A., & Zumailan, A. A. R. (1993). Global solar radiation estimation using sunshine duration in Yemen. *Renewable Energy*, 3(6–7), 645–653. <https://doi.org/10.1016/j.enconman.2003.08.022>
- Adaramola, M. S. (2012). Estimating global solar radiation using common meteorological data in Akure, Nigeria. *Renewable Energy*, 47, 38–44. <https://doi.org/10.1016/j.renene.2012.04.005>
- Allen, R. G., Pereira, L. S., Raes, D., & Smith, M. (1998). *Crop evapotranspiration-guidelines for computing crop water requirements-FAO irrigation and drainage paper 56*. FAO.
- Almorox, J., & Hontoria, C. (2004). Global solar radiation estimation using sunshine duration in Spain. *Energy Conversion and Management*, 45(9–10), 1529–1535. <https://doi.org/10.1016/j.enconman.2003.08.022>
- Ångström, A. (1924). Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. *Quarterly Journal of the Royal Meteorological Society*, 50(210), 121–126.
- Badescu, V., & Dumitrescu, A. (2015). Simple solar radiation modelling for different cloud types and climatologies. *Theoretical and Applied Climatology*, 124, 141–160. <https://doi.org/10.1007/s00704-015-1400-7>
- Bristow, K. L., & Campbell, G. S. (1984). On the relationship between incoming solar radiation and daily maximum and minimum temperature. *Agricultural and Forest Meteorology*, 31, 159–166. [https://doi.org/10.1016/0168-1923\(84\)90017-0](https://doi.org/10.1016/0168-1923(84)90017-0)
- Chandler, W. S., Stackhouse, P. W., Hoell, J. M., Westberg, D., & Zhang, T. (2013). NASA prediction of worldwide energy resource high resolution meteorology data for sustainable building design. In *Proceedings of the Solar 2013 Conference of American Solar Energy Society* (p. 7). American Solar Energy Society.
- Chen, J.-L., Li, G.-S., & Wu, S.-J. (2013). Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. *Energy Conversion and Management*, 75, 311–318. <https://doi.org/10.1016/j.enconman.2013.06.034>
- Chen, R., Lu, S., Kang, E., Yang, J., & Ji, X. (2006). Estimating daily global radiation using two types of revised models in China. *Energy Conversion and Management*, 47(7–8), 865–878. <https://doi.org/10.1016/j.enconman.2005.06.015>
- Chen, S., He, C., Huang, Z., Xu, X., Jiang, T., He, Z., Liu, J., Su, B., Feng, H., Yu, Q., & He, J. (2022). Using support vector machine to deal with the missing of solar radiation data in daily reference evapotranspiration estimation in China. *Agricultural and Forest Meteorology*, 316, 108864. <https://doi.org/10.1016/j.agrformet.2022.108864>
- De, Souza, J. L., Lyra, G. B., Dos, Santos, C. M., Ferreira, Junior, R. A., Tiba, C., Lyra, G. B., et al. (2016). Empirical models of daily and monthly global solar irradiation using sunshine duration for Alagoas state, Northeastern Brazil. *Sustainable Energy Technologies and Assessments*, 14, 35–45. <https://doi.org/10.1016/j.se-ta.2016.01.002>
- Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656), 553–597. <https://doi.org/10.1002/qj.828>
- Dong, J., Zeng, W., Lei, G., Wu, L., Chen, H., Wu, J., Huang, J., Gaiser, T., & Srivastava, A. K. (2022). Simulation of dew point temperature in different time scales based on grasshopper algorithm optimized extreme gradient boosting. *Journal of Hydrology*, 604, 127452. <https://doi.org/10.1016/j.jhydrol.2022.127452>
- Drucker, H., Surges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems* (Vol. 9, pp. 155–161). The MIT Press.
- Ehnberg, J. S. G., & Bollen, M. H. J. (2005). Simulation of global solar radiation based on cloud observations. *Solar Energy*, 78, 157–162. <https://doi.org/10.1016/j.solener.2004.08.016>
- Ertekin, C., & Evrendilek, F. (2007). Spatio-temporal modeling of global solar radiation dynamics as a function of sunshine duration for Turkey. *Agricultural and Forest Meteorology*, 145, 36–47. <https://doi.org/10.1016/j.agrformet.2007.04.004>
- Fan, J., Wu, L., Ma, X., Zhou, H., & Zhang, F. (2020). Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. *Renewable Energy*, 145, 2034–2045. <https://doi.org/10.1016/j.renene.2019.07.104>
- Feng, Y., Gong, D., Zhang, Q., Jiang, S., Zhao, L., & Cui, N. (2019). Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. *Energy Conversion and Management*, 198, 111780. <https://doi.org/10.1016/j.enconman.2019.111780>
- Gürel, A. E., Ağbulut, Ü., & Biçen, Y. (2020). Assessment of machine learning, time series, response surface methodology and empirical models in prediction of global solar radiation. *Journal of Cleaner Production*, 277, 122353. <https://doi.org/10.1016/j.jclepro.2020.122353>
- 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. *Applied Energy*, 179, 437–450. <https://doi.org/10.1016/j.apenergy.2016.07.006>
- 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 Conversion and Management*, 220, 113111. <https://doi.org/10.1016/j.enconman.2020.113111>
- He, T., Liang, S., Wang, D., Shi, Q., & Goulden, M. L. (2015). Estimation of high-resolution land surface net shortwave radiation from AVIRIS data: Algorithm development and preliminary results. *Remote Sensing of Environment*, 167, 20–30. <https://doi.org/10.1016/j.rse.2015.03.021>
- Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory. And appli. *Neurocomputing*, 70(1–3), 489–501. <https://doi.org/10.1016/j.neucom.2005.12.126>
- Hussain, M., Rahman, L., & Rahman, M. (1999). Techniques to obtain improved predictions of global radiation from sunshine duration. *Renewable Energy*, 18, 263–275. [https://doi.org/10.1016/S0960-1481\(98\)00772-1](https://doi.org/10.1016/S0960-1481(98)00772-1)
- Jahani, B., Dinpashoh, Y., & Nafchi, A. R. (2017). Evaluation and development of empirical models for estimating daily solar radiation. *Renewable and Sustainable Energy Reviews*, 73, 878–891. <https://doi.org/10.1016/j.rser.2017.01.124>
- Jin, Z., Yezheng, W., & Gang, Y. (2005). General formula for estimation of monthly average daily global solar radiation in China. *Energy*

- Conversion and Management*, 46(2), 257–268. <https://doi.org/10.1016/j.enconma-n.2004.02.020>
- Karatzoglou, A., Smola, A., & Hornik, K. (2004). kernlab—an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9), 1–20. <https://doi.org/10.18637/jss.v011.i09>
- Kisi, O., & Parmar, K. S. (2016). Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *Journal of Hydrology*, 534, 104–112. <https://doi.org/10.1016/j.jhydrol.2015.12.014>
- Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., & Takahashi, K. (2015). The JRA-55 reanalysis: General specifications and basic characteristics. *Journal of the Meteorological Society of Japan Ser. II*, 93(1), 5–48. <https://doi.org/10.2151/jmsj.2015-001>
- Li, M.-F., Fan, L., Liu, H.-B., Wu, W., & Chen, J.-L. (2012). Impact of time interval on the Ångström–Prescott coefficients and their interchangeability in estimating radiation. *Renewable Energy*, 44, 431–438. <https://doi.org/10.1016/j.renene.2012.01.107>
- 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 Conversion and Management*, 87, 975–989. <https://doi.org/10.1016/j.enconman.2014.08.001>
- Liu, X., Mei, X., Li, Y., Wang, Q., Jensen, J. R., Zhang, Y., & Porter, J. R. (2009). Evaluation of temperature-based global solar radiation models in China. *Agricultural and Forest Meteorology*, 149(9), 1433–1446. <https://doi.org/10.1016/j.agrformet.2009.03.012>
- Liu, X., Mei, X., Li, Y., Wang, Q., Zhang, Y., & Porter, J. R. (2009). Variation in reference crop evapotranspiration caused by the Ångström–Prescott coefficient: Locally calibrated versus the FAO recommended. *Agricultural Water Management*, 96(7), 1137–1145. <https://doi.org/10.1016/j.agwat.2009.03.005>
- 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. *Agricultural and Forest Meteorology*, 149(3–4), 697–710. <https://doi.org/10.1016/j.agrfor-met.2008.10.027>
- 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. *Applied Energy*, 96, 327–338. <https://doi.org/10.1016/j.apenenergy.2011.12.083>
- Lu, Y., Zhang, R., Wang, L., Su, X., Zhang, M., Li, H., Li, S., & Zhou, J. (2023). Prediction of diffuse solar radiation by integrating radiative transfer model and machine-learning techniques. *Science of the Total Environment*, 859, 160269.
- Makadea, R., & Jamil, B. (2018). Statistical analysis of sunshine based global solar radiation (GSR) models for tropical wet and dry climatic region in Nagpur, India: A case study. *Renewable and Sustainable Energy Reviews*, 87, 22–43. <https://doi.org/10.1016/j.rser.2018.02.001>
- 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. *Renewable and Sustainable Energy Reviews*, 43, 1224–1237. <https://doi.org/10.1016/j.rser.2014.12.004>
- Naserpour, S., Zolfaghari, H., & Firouzabadi, P. Z. (2020). Calibration and evaluation of sunshine-based empirical models for estimating daily solar radiation in Iran. *Sustainable Energy Technologies and Assessments*, 42, 100855. <https://doi.org/10.1016/j.seta.2020.100855>
- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., & Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57(3), e2020WR028091. <https://doi.org/10.1029/2020WR028091>
- Niu, Z., Wang, L., Chen, X., Yang, L., & Feng, L. (2021). Spatiotemporal distributions of pan evaporation and the influencing factors in China from 1961 to 2017. *Environmental Science and Pollution Research*, 28(48), 68379–68397.
- Pawlak, D. T., Clothiaux, E. E., Modest, M. F., & Cole, J. N. S. (2004). Full-spectrum correlated-k distribution for shortwave atmospheric radiative transfer. *Journal of the Atmospheric Sciences*, 61, 2588–2601. <https://doi.org/10.1175/JAS3285.1>
- Peng, Z., Chen, H., Wei, Z., Zhang, B., Zhang, S., Gong, L., Yang, G., Cai, J., Li, W., & Zhang, Q. (2022). Coefficient correction of Ångström–Prescott equation for China and its influence on solar radiation and reference crop evapotranspiration at different temporal and spatial scales. *Journal of Cleaner Production*, 375, 134013. <https://doi.org/10.1016/j.jclepro.2020.122353>
- Persaud, N., Lesolle, D., & Ouattara, M. (1997). Coefficients of the Ångström–Prescott equation for estimating global irradiance from hours of bright sunshine in Botswana and Niger. *Agricultural and Forest Meteorology*, 88, 27–35. [https://doi.org/10.1016/S0168-1923\(97\)00054-3](https://doi.org/10.1016/S0168-1923(97)00054-3)
- Podestá, G. P., Núñez, L., Villanueva, C. A., & Skansi, M. A. (2004). Estimating daily solar radiation in the Argentine pampas. *Agricultural and Forest Meteorology*, 123(1–2), 41–53. <https://doi.org/10.1016/j.agrformet.2003.11.002>
- Prescott, J. A. (1940). Evaporation from a water surface in relation to solar radiation. *Transactions of the Royal Society of South Australia*, 64, 114–118.
- Quinlan, J. R. (1992). Learning with continuous classes. In S. Adams (Ed.), *Proceedings of Australian Joint Conference on Artificial Intelligence* (pp. 343–348). World Scientific.
- R, Core, Team. (2013). R: A language and environment for statistical computing. <https://www.r-project.org/>
- 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. *Agricultural Systems*, 88(2–3), 451–471. <https://doi.org/10.1016/j.agsy.2005.07.004>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by backpropagating errors. *Nature*, 323(6088), 533–536. <https://doi.org/10.1038/323533a0>
- Sabziparvar, A. A., Mousavi, R., Marofi, S., Ebrahimpak, N. A., & Heidari, M. (2013). An improved estimation of the Ångström–Prescott radiation coefficients for the FAO56 penman–Monteith evapotranspiration method. *Water Resources Management*, 27(8), 2839–2854. <https://doi.org/10.1007/s11269-013-0318-z>
- Shiri, J., Marti, P., & Singh, V. P. (2014). Evaluation of gene expression programming approaches for estimating daily evaporation through spatial and temporal data scanning. *Hydrological Processes*, 28(3), 1215–1225. <https://doi.org/10.1002/hyp.9669>
- Shu, Z., Zhou, Y., Zhang, J., Jin, J., Wang, L., Cui, N., Wang, G., Zhang, J., Wu, H., Wu, Z., & Chen, X. (2022). Parameter regionalization based on machine learning optimizes the estimation of reference evapotranspiration in data deficient area. *Science of the Total Environment*, 844, 157034. <https://doi.org/10.1016/j.scitotenv.2022.157034>
- Soler, A. (1990). Statistical comparison for 77 European stations of 7 sunshine-based models. *Solar Energy*, 45(6), 365–370. [https://doi.org/10.1016/0038-092X\(90\)90157-8](https://doi.org/10.1016/0038-092X(90)90157-8)
- Tian, H., Wang, P., Tansey, K., Zhang, S., Zhang, J., & Li, H. (2020). An IPSO-BP neural network for estimating wheat yield using two remotely sensed variables in the Guanzhong plain, PR China. *Computers and Electronics in Agriculture*, 169, 105180. <https://doi.org/10.1016/j.compag.2019.105180>
- Trnka, M., Žalud, Z., Eitzinger, J., & Dubrovský, M. (2005). Global solar radiation in central European lowlands estimated by various empirical formulae. *Agricultural and Forest Meteorology*, 131(1–2), 54–76. <https://doi.org/10.1016/j.agrformet.2005.05.002>
- Tymvios, F. S., Jacovides, C. P., Michaelides, S. C., & Scouteli, C. (2005). Comparative study of Ångström’s and artificial neural networks’

- methodologies in estimating global solar radiation. *Solar Energy*, 78, 752–762. <https://doi.org/10.1016/j.solener.2004.09.007>
- Vapnik, V. N. (1996). The nature of statistical learning theory. *Technometrics*, 38(4), 409. https://doi.org/10.1007/978-1-4757-3264-1_1
- Wang, L., Kisi, O., Zounemat-Kermani, M., Zhu, Z., Gong, W., Niu, Z., Liu, H., & Liu, Z. (2017). Prediction of solar radiation in China using different adaptive neuro-fuzzy methods and M5 model tree. *International Journal of Climatology*, 37(3), 1141–1155.
- Xia, X., Pan, Y., Zhu, X., & Zhang, J. (2021). Monthly calibration and optimization of Ångström-Prescott equation coefficients for comprehensive agricultural divisions in China. *Journal of Geographical Sciences*, 31(7), 997–1014. <https://doi.org/10.1007/s11442-021-1882-4>
- Xing, L., Cui, N., Guo, L., Du, T., Gong, D., Zhan, C., Zhao, L., & Wu, Z. (2022). Estimating daily reference evapotranspiration using a novel hybrid deep learning model. *Journal of Hydrology*, 614, 128567. <https://doi.org/10.1016/j.jhydrol.2022.128567>
- Xing, L., Feng, Y., Cui, N., Guo, L., Du, T., Wu, Z., Zhang, Y., Wen, S., Gong, D., & Zhao, L. (2023). Estimating reference evapotranspiration using penman-Monteith equation integrated with optimized solar radiation models. *Journal of Hydrology*, 620, 129407. <https://doi.org/10.1016/j.jhydrol.2023.129407>
- Yacef, R., Mellit, A., Belaid, S., & Şene, Z. (2014). New combined models for estimating daily global solar radiation from measured air

- temperature in semi-arid climates: Application in Ghardaïa, Algeria. *Energy Conversion and Management*, 79, 606–615. <https://doi.org/10.1016/j.enconman.2013.12.057>
- Zhang, J., Zhu, Y., Zhang, X., Ye, M., & Yang, J. (2018). Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of Hydrology*, 561, 918–929. <https://doi.org/10.1016/j.jhydrol.2018.04.065>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chen, S., Feng, W., He, L., Xiao, W., Feng, H., Yu, Q., Liu, J., & He, J. (2024). Parameterization of the Ångström–Prescott formula based on machine learning benefit estimation of reference crop evapotranspiration with missing solar radiation data. *Hydrological Processes*, 38(2), e15091. <https://doi.org/10.1002/hyp.15091>