1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

**Assessing the potential of random forest method for estimating solar
radiation using air pollution index**

**Huaiwei Sun ^{a,b,*}, Dongwei Gui ^c, Baowei Yan ^a, Yi Liu ^a, Weihong Liao ^d, Yan
Zhu ^b, Chengwei Lu ^a, Na Zhao ^a**

Feb., 2016

For Publication in *Energy Conversion and Management*

^a School of hydropower and information engineering, Huazhong University of
Science and Technology, Wuhan 430074, China

^b State Key Laboratory of Water Resources and Hydropower Engineering Science,
Wuhan University, Wuhan 430072, China

^c Cele National Station of Observation & Research for Desert-Grassland Ecosystem in
Xinjiang, Chinese Academy of Sciences, Urumqi 830000, China

^d State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin,
China Institute of Water Resources and Hydropower Research, Beijing 100038, China

*** Corresponding author.**

Phone: 86 27 87543992 (Office) 86 34 1960 0578 (Cell)

E-mail address: huaiweisun@whu.edu.cn

26

27

ABSTRACT

28 Simulations of solar radiation have become increasingly common in recent years
29 because of the rapid global development and deployment of solar energy technologies.
30 The effect of air pollution on solar radiation is well known. However, few studies
31 have attempting to evaluate the potential of the air pollution index in estimating solar
32 radiation. In this study, meteorological data, solar radiation, and air pollution index
33 data from three sites having different air pollution index conditions are used to
34 develop random forest models. We propose different random forest models with and
35 without considering air pollution index data, and then compare their respective
36 performance with that of empirical methodologies. In addition, a variable importance
37 approach based on random forest is applied in order to assess input variables. The
38 results show that the performance of random forest models with air pollution index
39 data is better than that of the empirical methodologies, generating 9.1–17.0% lower
40 values of root-mean-square error in a fitted period and 2.0–17.4% lower values of
41 root-mean-square error in a predicted period. Both the comparative results of different
42 random forest models and variance importance indicate that applying air pollution
43 index data is improves estimation of solar radiation. Also, although the air pollution
44 index values varied largely from season to season, the random forest models appear
45 more robust performances in different seasons than different models. The findings can
46 act as a guide in selecting used variables to estimate daily solar radiation and improve
47 the accuracy of solar radiation estimation.

48

49 *Keywords:*

50 Random Forest; Solar radiation; Variable importance; Air pollution index

51

52 **1. INTRODUCTION**

53 Solar energy technology plays an essential role in providing a sustainable
54 energy future. Nearly half of the global solar photovoltaic (PV) is produced in China,
55 which is the largest solar PV manufacturer in the world. In addition, a Chinese
56 long-term plan estimates a target install capacity for solar PVs for the year 2020 to be
57 20 GW. For a given location, information on solar radiation is essential to the design
58 and evaluation of solar energy technologies. However, much of this information is not
59 readily available because of a limited number of meteorological stations in the world.
60 Several studies have devised methods to estimate solar radiation [1].

61 A major challenge in estimating solar radiation concerns its variability. Several
62 studies have observed alternating rates of rising and declining solar radiation at land
63 surfaces [2]. Spatiotemporal changes in the amount of solar irradiance have been
64 investigated by means of data collected over many years [3]. As numerous studies
65 reveal, urban development [4] and aerosols linked to air pollution affect variations in
66 solar radiation [5]. At least one study [6] has shown that human activities such as
67 burning of fuels and mining cause pollutants (tiny particles of elemental carbon, for
68 example) to be released into the atmosphere. Aerosols are known to act as both direct
69 and indirect radiative forcings in reducing radiation [7]. Direct radiative forcings
70 include methods of scattering and absorbing solar radiation. Indirect radiative forcings
71 include material means of attenuating surface solar radiation by acting as cloud
72 condensation nuclei to increasing cloud reflectivity and lifespan [8]. Confidential

73 evidence exists that increased air pollution leads to declines in solar radiation and thus
74 the number of hours of sunshine in China [9]. In addition, at least one study has
75 deduced that air pollution can affect solar radiation estimation that utilizes
76 meteorological variables. Specifically, Zhao et al. [10] have developed and validated
77 several expression models for predicting daily global radiation by introducing API
78 data as a new variable. A more thorough understanding of the API will enable us to
79 better estimate solar radiation.

80 Previous studies have revealed that sunshine duration is the most common
81 variable used to estimate solar radiation. In addition to the well-known
82 Angstrom-Prescott (A-P) model, different models have been introduced to improve
83 estimation performance [11]. Modifications to the A-P model are mainly for three
84 reasons [12]. First, the A-P coefficients are related to geographic elements or other
85 meteorological variables [13]. Second, the basic structure of models have been
86 changed from linear to nonlinear forms [14], such as raising relative sunshine to
87 second or third order [15]. Third, easily measured meteorological variables have been
88 added to the original A-P formula. Hargreaves et al. [16] proposed a method to
89 estimate solar radiation by maximum and minimum temperature. Relative humidity
90 data were also identified to be relevant with solar radiation and employed to estimate
91 solar radiation by Elagib et al. [17]. Also, De Jong et al. [18] related global solar
92 radiation to air temperatures, precipitation, and cloudiness in the wheat-growing area
93 of western Canada. As reported by Persaud et al. [19] in Botswana and Niger, the
94 coefficients of the A-P models are affected by air pollution, revealing that the API has

95 major potential in accurately estimating solar radiation. However, as more variables
96 have been introduced into the prediction model, the high-dimensional aspect of solar
97 radiation estimation has created an urgent need to measure variable importance and
98 avoid overfitting. An increasing number of machine learning techniques have been
99 studied in the area of energy. The random forest (RF) technique [20], which includes
100 an ensemble of decision trees and incorporates feature selection and interactions
101 naturally in the learning process, presents a good option for estimating solar radiation.
102 By using the RF method, one can obtain predictions and identify predictors that are
103 associated with the response via RF's inbuilt variable importance measures, and also
104 it provides the possibility of taking the ordering information into account and may
105 yield improved predictions. As far as we know, the potential of air pollution index has
106 been identified in estimating solar radiation [10], but the importance of air pollution
107 index has not been studied.

108 The aim of this study is to apply a RF approach for estimation of solar radiation
109 in three different sites with different API conditions in China, by using simple and
110 available meteorological variable and air pollution index. The main objectives of this
111 study are: (1) to quantify the effect of air pollution index for solar radiation estimation,
112 and (2) to investigate the use of the RF method in estimating solar radiation. This
113 article is structured as follows. We introduce the study sites, database, and brief theory
114 of RF models in Section 2. In Section 3, we described the obtained results and related
115 factors. Finally, some concluding remarks are given in Section 4.

116 2. DATA AND METHODS

117 2.1 Study sites and data

118 Based on the availability of meteorological, solar radiation, and air pollution
119 index (API) data, three sites having different API conditions are considered in this
120 study. Information about the general geography and climates of the three sites are
121 given in Table 1. Daily meteorological and solar radiation data for 2001–2012 are
122 provided by the China Meteorological Data Sharing Service System
123 (<http://old-cdc.cma.gov.cn/>), which is governed by the China Meteorological
124 Administration. As described by suppliers, global solar radiation is measured using
125 automatic telemetering radiometer (i.e., a thermoelectric type of radiometer in which
126 special optical black paint is used for induction surface) with relative error of $\pm 0.5\%$.

127 [Table 1]

128 API, which is used in this study to indicate the concentration of pollutants, is the
129 generalized model used to gauge air quality in China. Daily API data for 2001–2011
130 derive from the China National Environmental Monitoring Center
131 (<http://www.cnemc.cn/>). An API is a non-dimensional number in a range of 0–500
132 derived from the levels of three fundamental atmospheric pollutants of SO₂, NO₂, and
133 suspended particulates (PM₁₀). Among all pollutants considered in API, PM₁₀ is the
134 most dominant urban pollutant in China [21]. Fig. 1 shows the average API values in
135 Haikou, Changchun, and Urumqi as 36.4, 72.4, and 103.1, respectively. The API
136 values showed generally seasonal patterns in three sites. The highest values were

137 obtained during winter, and the lowest values were obtained during summer, as shown
138 in Table 1. Variations in seasonal patterns as revealed in/by API data may be because
139 of seasonal cycles derived from the normalized difference vegetation index (NDVI) as
140 well as anthropogenic factors such as winter heating [22].

141 [Figure 1]

142 2.2 Theory of RFs

143 RF is a fully non-parametric statistical method that was proposed by Breiman in
144 2001 [20]. It requires no distributional or functional assumptions on covariate
145 relations to the response. An RF algorithm is designed to produce accurate predictions
146 that do not overfit the data and construct multiple trees by drawing bootstrap samples
147 [23]. Each tree is built by a random subset of the original data from the training
148 procedure [20]. The use of bootstrap sampling in an RF model allows the remaining
149 unused subset, that is, the out-of-bag data (OOB), to be employed in the particular
150 evaluation of general problems. In general, RF model estimations are the averaged
151 result of most aggregations. Three user-defined parameters should be determined in
152 RF modeling, which are: the number of variables used to grow each tree (*mtry*), the
153 number of trees in the forest (*ntree*), and the minimum number of terminal nodes
154 (*nodesize*). The *mtry* parameter establishes the strength of each individual tree as well
155 as correlations between trees [24]. However, the predictive performance of the RF
156 model is improved by increasing the tree strength and decreasing the number of
157 correlations among trees [25]. In addition, the optimal parameter *nodesize* should be

158 set to fit an RF model. The relative importance of variables can be estimated from the
159 mean decrease in predictive accuracy when the variable is permuted. The operating
160 principle of RFs is illustrated in Fig. 2 and summarized as follows. First, j subsets of
161 the training sample $D_1, D_2, \dots, D_j$, are drawn from the total training sample set D
162 using the bootstrap sampling method. Second, j regression trees are constructed based
163 on the j subsets, and we thus obtain j regression results. Third, results of each
164 regression trees are aggregated to the optimal output.

165 [Figure 2]

166 2.3 Implementation of RF models

167 The input variables of RF models are selected based on previous studies and
168 empirical equations. As shown in Table 2, S (sunshine hours, h), S_0 (maximum
169 possible sunshine hours, h), T_{max} (maximum air temperature, °C), T_{min} (minimum
170 air temperature, °C), ΔT (air temperature difference, °C) and their derivative types
171 $(S/S_0, S/S_0^2, S/S_0^3, \exp(S/S_0), \Delta T^{0.5}, \exp(\Delta T))$ are used in previous studies. All of these
172 variables are considered in RF modeling.

173 In this study, the API is used to indicate the effect of air pollution on solar
174 radiation. Three models, linear, exponential, and logarithmic, were proposed by Zhao
175 et al. (2013) [10], and we consider these as input variables in this study.

176 [Table 2]

177 As previously described, RF and non-API RF modeling require three
178 user-defined parameters: $mtry$, $ntree$, and $nodesize$. The default value of $mtry$ is one

179 third the total number of predictors [25]. In this study, we applied the RF and non-API
 180 RF models with the default values of *mtry* and *nodesize* (5). However, the default
 181 value of *ntree* (500) has proven to be insufficient in generating stable results [32] and
 182 a higher value (1000) suggested by Yang et al. (2016) [33] was used in this study.
 183 Unlike in most artificial intelligence algorithms such as SVM, data scaling is not
 184 required in RFs, as no kernel values are generated when implementing the model [34].

185 For variable importance measures, we consider the unscaled permutation-based
 186 importance measure. Breiman [20] suggested that the difference in prediction
 187 accuracy before and after permuting $\mathbf{X}_j$, averaged over all regression trees, should
 188 be a measure for variable importance. Calculating permutation importance is achieved
 189 as follows. Let $\bar{D}^{(t)}$ be the OOB sample for a tree t (the total number of trees is
 190 *ntree*). The variable importance of variable $\mathbf{X}_j$ in tree t is then given as:

$$191 \quad VI^{(t)}(\mathbf{X}_j) = \frac{\sum_{i \in \bar{D}^{(t)}} I(y_i = \hat{y}_i^{(t)})}{|\bar{D}^{(t)}|} - \frac{\sum_{i \in \bar{D}^{(t)}} I(y_i = \hat{y}_{i, \pi_j}^{(t)})}{|\bar{D}^{(t)}|} \quad (1)$$

192 where $\hat{y}_i^{(t)}$ is the simulated result for observation I before and $\hat{y}_{i, \pi_j}^{(t)}$ is the
 193 predicted class for observation I after permuting its value of variable $\mathbf{X}_j$. The raw
 194 variable importance score for each variable is then computed as the mean importance
 195 over all trees.

$$196 \quad VI(\mathbf{X}_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(\mathbf{X}_j)}{ntree} \quad (2)$$

197 In this study, the contribution of the index of each variable is calculated from
 198 the variable importance score $VI(\mathbf{X}_j)$ as the ratio to the sum of all variables.

199 3. RESULTS AND DISCUSSION

200 3.1 RF performances

201 For a graphical illustration, the fitted and predicted solar radiation values
202 together with observation data are shown in Fig. 3. Collected data from 2001–2010
203 are used for fitting in RF models and the period 2011–2012 are used for prediction.
204 Because of the large variation of solar radiation seen in recent years, a tiny difference
205 in the observed average values of solar radiation appears during the fitted and
206 predicted periods. The average values of solar radiation during 2001–2010 are 14.198,
207 13.201, and 14.168 MJm⁻² in Haikou, Changchun, and Urumqi, respectively. Whereas
208 for the year 2011–12, the average observed values are 13.318, 13.612, and 13.512,
209 respectively. As shown in Fig. 3, most predictions of fitted solar radiation values agree
210 with the observations. The trend of actual radiation can be well captured from
211 predictions in the aforementioned three sites. In these three sites, even the extreme
212 points can be estimated by the RF algorithms during the fitted period. In addition, all
213 the RF models show good performance, as the predicted values approximate the
214 observed data. This means that RF models are able to estimate solar radiation.

215 [Figure 3]

216 [Table 3]

217 Values of statistical Nash-Sutcliffe efficiency (NSE) are used to compare the
218 performance of the models in estimating daily solar radiation. As Tables 3 and 4 show,
219 both RF and non-API RF models perform better than the empirical models. If we

220 consider the entire dataset (“All”) used in the Haikou site, RF models produced an
221 average NSE of 0.908, which is 0.058 higher than that produced by empirical models.
222 In addition, the RF models showed improvement of 0.085 and 0.121 in NSE values in
223 the Changchun and Urumqi sites, respectively. For the predicted period (2011–2012),
224 a considerable improvement in NSE was seen in Urumqi from the 0.546 of empirical
225 models to 0.862 by the RF models. Actually, the empirical models produced
226 acceptable performance, as shown in Table 3, which Zhao et al. (2013) in [10] also
227 indicated. However, better performances of RF models than empirical models
228 indicated that applying the RF models to daily solar radiation estimation was
229 necessary.

230 **3.2 Seasonal effects**

231 As the API data indicated significant seasonal patterns in the three sites,
232 seasonal effects on model performance for estimating solar radiation were
233 investigated by building different RF and empirical models. In Table 3, NSE values
234 obtained during the fitted period varied largely in different seasons. For example, RF
235 models applied in the Urumqi site produced an average NSE of 0.952 in autumn,
236 which is 0.138 higher than the 0.814 produced in winter. We observed that empirical
237 models produced greater differences in different seasons than did RF models during
238 the fitted period. This result is further confirmed for the predicted period, as shown in
239 Table 4. At the Urumqi site, the average NSE produced by the RF models is higher in
240 spring (0.901) than summer (0.461), whereas NSE indices produced by empirical

241 models were 0.854 in spring and 0.265 in summer.

242 **[Table 4]**

243 **[Figure 4]**

244 Seasonal effects are investigated using RMSE values, as shown in Fig. 4.
245 Significant differences were observed in different seasons, especially at the Urumqi
246 site. The RF model produced RMSE values of 2.229 MJm⁻² in summer and 1.189
247 MJm⁻² in winter during the fitted period, and 4.567 MJm⁻² in summer and 1.439
248 MJm⁻² in winter during the predicted period. For the empirical models, the largest
249 differences of RMSE values in the fitted and predicted periods were 1.034 MJm⁻² and
250 3.253 MJm⁻², respectively. These results indicate that the performances of both RF
251 and empirical models varied in different seasons. However, if we compare the
252 seasonal differences of model performances in the predicted period, the RF models
253 contributed to more robust results than did the empirical models, as shown in both
254 Table 3 and Fig. 4.

255

256 **[Figure 5]**

257 **[Figure 6]**

258 In order to investigate the differences of model performances between RF
259 model and empirical models, NSE values of different models during fitted period and
260 predicted period were plotted in Fig. 5 and Fig. 6. Among all sites, the proposed RF
261 model yield better NSE values than empirical models as plots over the 1:1 lines. The
262 differences between empirical models are also shown in Fig. 5 and Fig. 6.

263 Logarithmic model show advantages over other models (as shown in Fig. 5c, Fig 6a
264 and Fig 6c). Zhao et al. [10] had reported similar conclusion, although they ignored
265 the model performance of the predicted period.

266 The analysis of variance (ANOVA), which is an approach extensively used for
267 treatment assessment in climate and hydrology studies during the last decades, was
268 applied in this study to investigate the sources of variances of model performances in
269 estimating solar radiation. Following the three-way ANOVA approach applied by
270 Vetter et al. [35], NSE values of RF models and empirical models were used here to
271 decompose variance sources into three components. Table 5 shows the results of
272 three-way ANOVA in different models' performances. The p-values 0, 0.0144, and
273 0.0006 is small enough to conclude that the mean responses of sites, models, and
274 seasons are significantly different. It is indicated that RF models should prefer to use
275 in solar radiation, but the RF model performances varied largely in different sites and
276 seasons.

277 **[Table 5]**

278 Because meteorological observations, solar radiation, and API data yield
279 seasonal and random effects, the basic concern of solar forecasting methods is
280 seasonal effects on solar radiation estimation. Tymvios et al. determined that
281 data-driven methods depend considerably on the training data set and produce worse
282 prediction when the weather shows considerable variability [36]. In previous studies,
283 the monthly variations of coefficients in A-P models [37] or seasonal terms in
284 autoregressive models [38] are preferred in order to consider seasonal effects. A

285 detailed discussion of the seasonal dependency of model performance is presented by
286 Inman et al. [39]. However, the results from our study offer an alternative approach to
287 considering the seasonal dependency of solar radiation estimation. Although the API
288 data reveals the presence of major seasonal cycles, as shown in Fig. 1 and Table 1, the
289 RF models estimated in our study that use API show robust performances in different
290 seasons. Instead, we use the method of variable importance to determine the
291 contribution of API data to solar radiation estimation.

292 **3.3 Variable importance**

293 RF regression can act as a measure of variable importance, as described in
294 Section 2. Fig. 7 shows the contribution of each variable to RF models for the whole
295 data set and the different season subsets. Based on the contribution of the index at the
296 Haikou sites as shown in Fig. 7a, the values of “ S ”, “ $S \cdot \exp(API)$ ”, “ S/S_0 ”, “ $\exp(S/S_0)$ ”,
297 and “ S/S_0^2 ” are top five contributors of the 18 selected variables for the entire dataset
298 (indicated as “All”). This means that RF identifies these five variables as the most
299 important for estimating daily solar radiation. Among them, sunshine hours were most
300 important in estimating solar radiation. Chen et al. [34] and Sun et al. [40] reported
301 that estimation of solar radiation from sunshine hours offers an important alternative
302 in the absence of measured solar radiation. In addition to sunshine hours, the API data
303 related to “ $S \cdot \exp(API)$ ” is critical to estimating solar radiation. For the different
304 seasons of spring, summer, autumn, and winter (Fig. 7a), “ $S \cdot \exp(API)$ ” ranked as the
305 1st, 4th, 3rd, and 1st most important variables, respectively, , which further confirms the

[Figure 8]

However, we consider the fact that API data is unavailable in some cities. Therefore, the values of variable importance in non-API RF models are shown in Fig. 9. “ S ” is listed as the most important variable in nearly all simulations by non-API RF models. Compared to air temperature, the average contribution of sunshine duration (related to the variables “ S ”, “ S/S_0 ”, “ $\exp(S/S_0)$ ”, “ S/S_0^2 ” and “ S/S_0^3 ”) in non-API RF models is nearly 75.0%. This result confirms that sunshine duration can be used to achieve good performance in estimating solar radiation. However, combining API and sunshine duration data together in daily solar radiation estimation is recommended.

[Figure 9]

4. CONCLUSIONS

This study presented an application of an RF algorithm in estimating daily solar radiation. In addition, RF models were used to determine variable contribution. Solar radiation, meteorological, and API data from three sites with different API conditions were examined in this study. Both RF and non-API RF models were evaluated and simulations of these models showed good performance. The RF models that use API data outperformed both the non-API RF and empirical models. Nevertheless, the RF models that use API data consistently generated lower average RMSE values and better NSE values than did the non-API RF models. Specifically, at the three sites, RF models produced average values of RMSE that were lower than those of non-API RF models by 2.48, 1.18, and 0.41%, respectively.

An important feature of the RF algorithm application is that RF models can

350 identify variable importance. Eighteen variables were evaluated in this study as
351 possible variables to employ in RF models. Sunshine hours were the most important
352 contributing factor in estimating solar radiation, and API data was crucial. The results
353 of this study indicate that API data is a promising complement to traditional
354 approaches for estimating solar radiation.

355 **ACKNOWLEDGMENTS**

356 The authors would like to thank the National Sciences Foundation of China
357 (Grant 51309106, 51309105, and 51409109), the Science and Technology support
358 program of Wuhan City (2015020202010127) and the National Science and
359 Technology support program of China (2015BAB07B00) for financially supporting
360 this work. Special thanks go to the National Meteorological Information Center,
361 China Meteorological Administration for providing the long-term data records.

362 **REFERENCES**

- 363 [1]Khatib T, Mohamed A, Sopian K. A review of solar energy modeling techniques. *Renew Sustain*
364 *Energy Rev* 2012;16:2864–9.
- 365 [2]Wild, M., H. Gilgen, A. Roesch, A. Ohmura, C. N. Long, E. G. Dutton, B. Forgan, A. Kallis, V.
366 Russak, A. Tsvetkov. From dimming to brightening: Decadal changes in solar radiation at Earth's
367 surface, *Science* 2005; 308(5723), 847–50.
- 368 [3]Wang, Y., Y. Yang, N. Zhao, C. Liu, Q. Wang. The magnitude of the effect of air pollution on
369 sunshine hours in China, *J. Geophys. Res.* 2012; 117, D00V14, doi:10.1029/2011JD016753.
- 370 [4]Alpert, P., P. Kishcha, Y. J. Kaufman, R. Schwarzbard. Global dimming or local dimming?: Effect of
371 urbanization on sunlight availability, *Geophys. Res. Lett.* 2005; 32, L17802,
372 doi:10.1029/2005GL023320.
- 373 [5]Streets, D. G., Y. Wu, and M. Chin. Two-decadal aerosol trends as a likely explanation of the global
374 dimming/brightening transition, *Geophys. Res. Lett.* 2006; 33, L15806,
375 doi:10.1029/2006GL026471.

376 [6]Charlson, R. J., S. E. Schwartz, J. M. Hales, R. D. Cess, J. A. Coakley, J. E. Hansen, D. J. Hofmann.
377 Climate forcing by anthropogenic aerosols, *Science* 1992; 255(5043), 423–30,
378 doi:10.1126/science.255.5043.423.

379 [7]Ramanathan, V., P. J. Crutzen, J. T. Kiehl, D. Rosenfeld. Aerosols, climate, and the hydrological
380 cycle, *Science* 2001; 294(5549), 2119–24, doi:10.1126/science.1064034.

381 [8]Wild, M. Global dimming and brightening: A review, *J. Geophys. Res.* 2009; 114, D00D16,
382 doi:10.1029/2008JD011470.

383 [9]Che, H. Z., G. Y. Shi, X. Y. Zhang, R. Arimoto, J. Q. Zhao, L. Xu, B. Wang, Z. H. Chen. Analysis of
384 40 years of solar radiation data from China, 1961–2000, *Geophys. Res. Lett.* 2005; 32, L06803,
385 doi:10.1029/2004GL022322.

386 [10]Zhao N., X. Zeng, S. Han. Solar radiation estimation using sunshine hour and air pollution index in
387 China. *Energy Convers Manage* 2013;76: 846–51.

388 [11]Yorukoglu M, Celik AN. A critical review on the estimation of daily global solar radiation from
389 sunshine duration. *Energy Convers Manage* 2006;47: 2441–50.

390 [12]Liu XY, Mei XR, Li YZ. Calibration of the Angstrom–Prescott coefficients (a, b) under different
391 time scales and their impacts in estimating global solar radiation in the Yellow River basin. *Agric*
392 *For Meteorol* 2009;149:697–710.

393 [13]Iziomon MG, Mayer H. Assessment of some global solar radiation parameterizations. *J Atmos Sol*
394 *Terr Phys* 2002;64:1631–43.

395 [14]Almorox J, Hontaria C. Global solar radiation estimation using sunshine duration in Spain. *Energy*
396 *Convers Manage* 2004;45:1529–35.

397 [15]Besharat F, Dehghan AA, Faghih AR. Empirical models for estimating global solar radiation: a
398 review and cases study. *Renew Sustain Energy Rev* 2013;21:798–821.

399 [16]Hargreaves GH, Samani ZA. Estimating potential evapotranspiration. *J Irrig Drain Div*
400 1982;108:225–30.

401 [17]Elagib NA, Babiker SF, Alvi SH. New empirical models for global solar radiation over Bahrain.
402 *Energy Convers Manage* 1998;39(8):827–35.

403 [18]De Jong R, Stewart DW. Estimating global solar radiation from common meteorological variables
404 in western Canada. *Can J Plant Sci* 1993;73:509–18.

405 [19]Persaud N, Lesolle D, Ouattara M. Coefficients of the Angstrom–Prescott equation for estimating

406 global irradiance from hours of bright sunshine in Botswana and Niger. *Agric Forest Meteorol*
 407 1997;88:27–35.

408 [20]Breiman, L. Random forests. *Mach. Learn.* 2001; 45 (1), 5–32. DOI10.1023/A:1010933404324.

409 [21]Wang, L., Zhang P, Tan S, Zhao X, Cheng D, Wei W, Su J, Pan X. Assessment of urban air quality
 410 in China using air pollution indices (APIs) *J. Air Waste Manage. Assoc.* 2013; 63 170–8.

411 [22]Li, L., Y. Wang. What drives the aerosol distribution in Guangdong - the most developed province
 412 in Southern China? *Scientific Reports*, 2014. <http://dx.doi.org/10.1038/srep05972>.

413 [23]Lahouar, A., J. B. Hadj Slama. Day-ahead load forecast using random forest and expert input
 414 selection. *Energy Conversion and Management* 2015; 103: 1040–51

415 [24]Peters, J., Verhoest, N., Samson, R., Boeckx, P., De Baets, B. Wetland vegetation distribution
 416 modelling for the identification of constraining environmental variables. *Landsc. Ecol.* 2008; 23,
 417 1049–65.

418 [25]Ließ, M., Glaser, B., Huwe, B. Uncertainty in the spatial prediction of soil texture: comparison of
 419 regression tree and Random Forest models. *Geoderma* 2012; 170, 70–9.

420 [26]Ångström A. Solar and terrestrial radiation. *QJR Meteorolog Soc* 1924;50:121–6.

421 [27]Prescott JA. Evaporation from a water surface in relation to solar radiation. *Trans R Soc S Aust*
 422 1940;64:114–8.

423 [28]Bahel V, Bakhsh H, Srinivasan R. A correlation for estimation of global solar radiation. *Energy*
 424 1987;12:131–5.

425 [29]Bakirci K. Correlations for estimation of daily global solar radiation with hours of bright sunshine
 426 in Turkey. *Energy* 2009;34:485–501.

427 [30]Allen RG, Pereira LS, Raes D, Smith M. Crop evapotranspiration-guidelines for computing crop
 428 water requirements. In: *FAO Irrigation and drainage paper 56*. Rome: Food and Agriculture
 429 Organization of the United Nations; 1998.

430 [31]Bristow KL, Campbell GS. On the relationship between incoming solar radiation and daily
 431 maximum and minimum temperature. *Agric For Meteorol* 1984;31:159–66.

432 [32]Grimm, R., Behrens, T., Märker, M., Elsenbeer, H. Soil organic carbon concentrations and stocks
 433 on Barro Colorado Island-digital soil mapping using Random Forests analysis. *Geoderma* 2008;
 434 146, 102–13.

435 [33]Yang, R., G. Zhang, F. Liu, Y. Lu, F. Yang, F. Yang, M. Yang, Y. Zhao, D. Li. Comparison of

436 boosted regression tree and random forest models for mapping topsoil organic carbon
437 concentration in an alpine ecosystem. *Ecological Indicators* 2016; 60: 870–78..

438 [34]Chen J, Li G, Wu S. Assessing the potential of support vector machine for estimating daily solar
439 radiation using sunshine duration. *Energy Convers Manage* 2013;75:311–8.

440 [35]Vetter T., Huang S., Aich V., Yang T., Wang X., Krysanova V., Hattermann F. Multi-model climate
441 impact assessment and intercomparison for three large-scale river basins on three continents. *Earth*
442 *Syst. Dynam.*, 2015, 6,17-43.

443 [36]Tymvios FS, Silas CM, Chara SS. Estimation of Surface Solar Radiation with Artificial Neural
444 Networks. In: Badescu, V, editors. *Modeling Solar Radiation at the Earth's Surface*, Berlin
445 Heidelberg: Springer Inc; 2008, p. 221–56.

446 [37]Sahin AD, Sen Z. Statistical analysis of the Angstrom formula coefficients and application for
447 Turkey. *Solar Energy* 1998; 62: 29–36.

448 [38]Sun, H., D. Yan, N. Zhao, J. Zhou. Empirical investigation on modeling solar radiation series with
449 ARMA–GARCH models. *Energy Conv. Mana.* 2015; 92, 385-95.

450 [39]Inman R., T. Hugo. C. Pedro, F. Coimbra. Solar forecasting methods for renewable energy
451 integration. *Progress in Energy and Combustion Science* 2013; 39: 535-76.

452 [40]Sun, H., N. Zhao, X. Zeng, D. Yan. Study of solar radiation prediction and modeling of
453 relationships between solar radiation and meteorological variables. *Energy Conv. Mana.* 2015;105,
454 880-90.

455

456

LIST OF TABLES

457 Table 1. Information about the general geography, climates, and API data records of the three sites.

Stations	Latitude	Longitude	Daily solar radiation (MJ/m ²)			Average daily API in the different seasons				
			Mean	Max.	Min.	All	Spring	Summer	Autumn	Winter
Haikou	20°01'48"	110°21'00"	14.09	30.52	0.02	36.4	35.8	27.1	39.6	43.1
Changchun	43°54'00"	125°13'12"	13.28	31.50	0.01	72.4	75.9	58.6	68.6	87.0
Urumqi	43°46'48"	87°39'00"	14.04	34.62	0	103.1	86.6	60.4	89.6	177.4

Table 2. Input variables for estimating solar radiation proposed in literatures.

Models	Types	Input variables	Equations	Source
Sunshine duration based	Linear	S, S ₀ , S/S ₀	$R_s = Ra(aS/S_0 + b)$	[26] and [27]
	Cubic	S, S ₀ , S/S ₀ , S/S ₀ ² , S/S ₀ ³	$R_s = Ra(aS/S_0 + bS/S_0^2 + cS/S_0^3 + d)$	[28]
	Exponential	S, S ₀ , S/S ₀ , exp(S/S ₀)	$R_s = Ra(aS/S_0 + b \exp(S/S_0) + c)$	[11] and [29]
Temperature based	Power	Tmax, Tmin, ΔT , $\Delta T^{0.5}$	$R_s = Ra(a\Delta T^{0.5} + b)$	[30] and [16]
	Exponential	Tmax, Tmin, ΔT , exp(ΔT)	$R_s = Ra(a - a \exp(b\Delta T))$	[31]
	Linear	API/100	$R_s = Ra(a + bS/S_0 + cAPI/100 + dS/S_0(API/100))$	[10]
API based	Exponential	exp(API/100)	$R_s = Ra(a + bS/S_0 + c \exp(API/100) + dS/S_0 \exp(API/100))$	[10]
	Logarithmi	log(API/100)	$R_s = Ra(a + bS/S_0 + c \log(API/100) + dS/S_0 \log(API/100))$	[10]

459

460 Table 3. Statistical NSE results of different models for daily solar radiation estimation during fitted
461 period (2001-2010)

Sites	Method	Seasons				
		All	Spring	Summer	Autumn	Winter
Haikou	RF model	0.910	0.908	0.870	0.863	0.901
	Non-API RF model	0.905	0.902	0.867	0.850	0.894
	Linear model	0.851	0.859	0.860	0.809	0.857
	Exponential model	0.850	0.859	0.859	0.809	0.856
	Logarithmic model	0.850	0.859	0.860	0.807	0.857
Changchun	RF model	0.938	0.905	0.903	0.937	0.831
	Non-API RF model	0.936	0.904	0.898	0.934	0.826
	Linear model	0.852	0.899	0.891	0.880	0.716
	Exponential model	0.851	0.897	0.890	0.879	0.715
	Logarithmic model	0.853	0.899	0.893	0.876	0.718
Urumqi	RF model	0.958	0.914	0.894	0.952	0.819
	Non-API RF model	0.956	0.911	0.894	0.951	0.809
	Linear model	0.836	0.850	0.890	0.880	0.651
	Exponential model	0.828	0.850	0.890	0.878	0.637
	Logarithmic model	0.844	0.855	0.891	0.882	0.659

462

463 Table 4. Statistical Nash-Sutcliffe efficiency (NSE) results of different models for daily solar radiation
464 estimation during predicted period (2011-2012)

Sites	Method	Seasons				
		All	Spring	Summer	Autumn	Winter
Haikou	RF model	0.913	0.874	0.886	0.892	0.843
	Non-API RF model	0.909	0.866	0.876	0.885	0.834
	Linear model	0.846	0.801	0.856	0.853	0.780
	Exponential model	0.846	0.802	0.855	0.853	0.778
	Logarithmic model	0.848	0.801	0.856	0.853	0.787
Changchun	RF model	0.950	0.932	0.899	0.956	0.821
	Non-API RF model	0.949	0.932	0.897	0.953	0.819
	Linear model	0.855	0.923	0.896	0.882	0.614
	Exponential model	0.853	0.923	0.898	0.880	0.627
	Logarithmic model	0.859	0.925	0.892	0.877	0.619
Urumqi	RF model	0.863	0.901	0.462	0.658	0.731
	Non-API RF model	0.862	0.900	0.461	0.652	0.702
	Linear model	0.538	0.854	0.258	0.335	0.456
	Exponential model	0.531	0.849	0.261	0.333	0.448
	Logarithmic model	0.570	0.859	0.275	0.350	0.463

465

466 Table 5. Results of three-way ANOVA in different models' performances.

Source	Sites	Models	Seasons
F Values	41.98	3.87	5.83
p-values (Prob > F)	0	0.0144	0.0006

467

468

469

LIST OF FIGURES

470

471

472

Fig. 1 The daily API in three different sites.

473

474

Fig. 2 Random Forest regression principle.

Fig. 3 Time series plots of the observed vs. estimated solar radiation by Random Forest regression models in three sites.

Fig. 4 Performance of the RF and empirical models for the whole dataset and different season's sub-sets.

Fig. 5 Comparison of NSE values of RF models and empirical models in three sites during fitted period.

Fig. 6 Comparison of NSE values of RF models and empirical models in three sites during predicted period.

Fig. 7 Measure for variable importance given by RF regression model.

Fig. 8 Comparison of model performance of RF models and non-API RF models in three sites.

499

500

Fig. 9 Measure for variable importance given by non-API RF regression model.