SPECIAL TOPIC — Smart design of materials and design of smart materials

Forecasting solar still performance from conventional weather data variation by machine learning method

Wenjie Gao(高文杰)[1], Leshan Shen(沈乐山)[2,3], Senshan Sun(孙森山)[1], Guilong Peng(彭桂龙)[1], Zhen Shen(申震)[2,3], Yunpeng Wang(王云鹏)[1], AbdAllah Wagih Kandeal[6], Zhouyang Luo(骆周杨)[2,3], A.E Kabeel[7], Jianqun Zhang(张坚群)[4,†], Hua Bao(鲍华)[5,‡], and Nuo Yang(杨诺)[1,§]

[1] *State Key Laboratory of Coal Combustion, and School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan* 430074*, China*

[2] *Key Laboratory of Solar Energy Utilization & Energy Saving Technology of Zhejiang Province, Hangzhou* 311121*, China*

[3] *Zhejiang Energy Group R & D Institute Co., Ltd., Hangzhou* 311121*, China*

[4] *Zhejiang Zheneng Yueqing Electric Power Generation Co., Ltd., Yueqing* 325609*, China*

[5] *University of Michigan–Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai* 200240*, China*

[6] *Mechanical Engineering Department, Faculty of Engineering, Kafrelsheikh University, Kafrelsheikh* 33516*, Egypt*

[7] *Mechanical Power Engineering Department, Faculty of Engineering, Tanta University, Tanta, Egypt*

(Received 16 July 2022; revised manuscript received 24 September 2022; accepted manuscript online 10 October 2022)

Solar stills are considered an effective method to solve the scarcity of drinkable water. However, it is still missing a way to forecast its production. Herein, it is proposed that a convenient forecasting model which just needs to input the conventional weather forecasting data. The model is established by using machine learning methods of random forest and optimized by Bayesian algorithm. The required data to train the model is obtained from daily measurements lasting 9 months. To validate the accuracy model, the determination coefficients of two types of solar stills are calculated as 0.935 and 0.929, respectively, which are much higher than the value of both multiple linear regression (0.767) and the traditional models (0.829 and 0.847). Moreover, by applying the model, it is predicted that the freshwater production of four cities in China. The predicted production is approved to be reliable by a high value of correlation (0.868) between the predicted production and the solar insolation. With the help of the forecasting model, it would greatly promote the global application of solar stills.

Keywords: solar still, production forecasting, forecasting model, weather data, random forest

PACS: 88.40.–j, 92.60.Vb **DOI:** 10.1088/1674-1056/ac989f

1. Introduction

Seawater covers 70% of the earth, freshwater is mainly distributed in glaciers, ice caps, and underground.[1,2] With the increase in population and industrial activities, the shortage of drinkable water is a catastrophic issue the world facing.[3,4] As seawater accounts for 97% of water on the earth, desalination is an effective solution for the shortage of freshwater.[5]

Among the many desalination technologies, solar desalination[6] is one of the most environmentally friendly technologies. Fortunately, areas where freshwater is scarce happen to possess abundant solar energy.[7] Solar still is one of the solar desalination technologies, which is easy to install and maintain.[8] Solar still has broad application prospects in remote coastal areas and islands. Given this, solar desalination has received widespread attention in recent years.[9–18] However, the value of daily production fluctuates greatly and is much affected by climatic conditions, which are not easily forecasted.

Traditional models[19–21] show the function between production and a couple of important factors. Due to the complexity of heat and mass transfer, in reality, these models with simple functions are difficult to describe the heat and mass transfer process inside the solar still accurately, which limited to guide the design of solar stills.[22] Recently, it is an emerging and effective way to predict the performance of solar still by using the machining learning method.[23] Such as the multiple linear regression (MLR) method,[24] artificial neural network (ANN) method,[25,26] random forest (RF) method.[27,28] Among current algorithms, RF is an ensemble learning algorithm based on decision trees, with unexcelled accuracy,[29,30] and shows excellent performance in predicting.[28]

However, the previous studies just gave the functional relationship between the performance and a couple of professional parameters, such as basin plate temperature, glass cover temperature, feedwater temperature, *etc.*, which is not convenient to measure for customers. More importantly, the previous models cannot forecast production in advance, which is a big challenge.

†Corresponding author. E-mail: zhangjianq888@163.com
‡Corresponding author. E-mail: hua.bao@sjtu.edu.cn
§Corresponding author. E-mail: nuo@hust.edu.cn

The production is greatly affected by the weather. And, it is easy to obtain weather forecast data, such as air temperature, humidity, wind, atmospheric pressure, and air quality index. It will be a convenient and effective way to forecast the production if a model could be established between the production and the weather forecasting data.

Production forecasting is significant to promote the global application of solar still. Even in remote areas, it is not difficult to get conventional weather forecasting. Besides, forecasting can help to make a stable supply of water or a controllable desalination capacity. That is, with the help of forecasting, a proper substitute desalination strategy can be planned and chosen, such as using electrically powered desalination as compensation.

This work aims to make a model forecast the daily production of solar still based on convenient weather data. The required data to train the model was obtained by carrying out experimental measurements from July 2020 to March 2021. Based on the production and weather data, the forecasting model was conducted using the random forest method. To verify the practicability and accuracy of the model, the determination coefficients were calculated and compared. By applying the model, the freshwater production of four cities in China was forecasted from conventional weather data.

2. Experimental systems

The solar stills consist of a glass cover, basin, foam heat-insulation layer, water feeding tank, freshwater outlet, and required measuring instrument, as shown in Fig. 1(a). The bottom dimension is 50 cm × 50 cm. Singh and Tiwari[31] reported that the annual solar still yield reached a maximum value when the condensing glass cover inclination was equal to the latitude of the place. Thus the glass cover of the solar stills has an inclination angle of 30°, which is the preferred solar incidence angle at Hangzhou (120.2° E, 30.3° N). The equipment is installed on the roof of a building in Hangzhou, China. The solar still is placed horizontally and the front is south-facing.

The schematic diagram of solar still is shown in Fig. 1(b). The solar still has an interfacial evaporation structure at the bottom and insulation foams at the sidewall (BIF-SS). The BIF-SS adopts a three-layer composite structure: floating light absorption layer, water-conducting layer, and heat-insulating layer. The light-absorbing layer structure is made of black deerskin velvet fiber cloth, with 95% solar absorption. The water-conducting layer is made of cotton fiber cloth with a thickness of about 8 mm, and is in contact with seawater through the water-conducting channel. The sides and bottom are all wrapped with heat-insulating extruded foam XPS board, 2-cm thick. The thermal conductivity of the XPS board is 0.03 W/m·K. The freshwater is obtained from the freshwater collection tank, and recorded by cylinder manually. The solar still with interfacial evaporation structure is designed based on our previous work,[32] which has both high energy efficiency and salt rejection capacity. Meanwhile, a control group was set up on the solar still with an interfacial evaporation structure at both the bottom and the sidewall (BSI-SS). The schematic of BSI-SS is shown in Fig. 1(c).

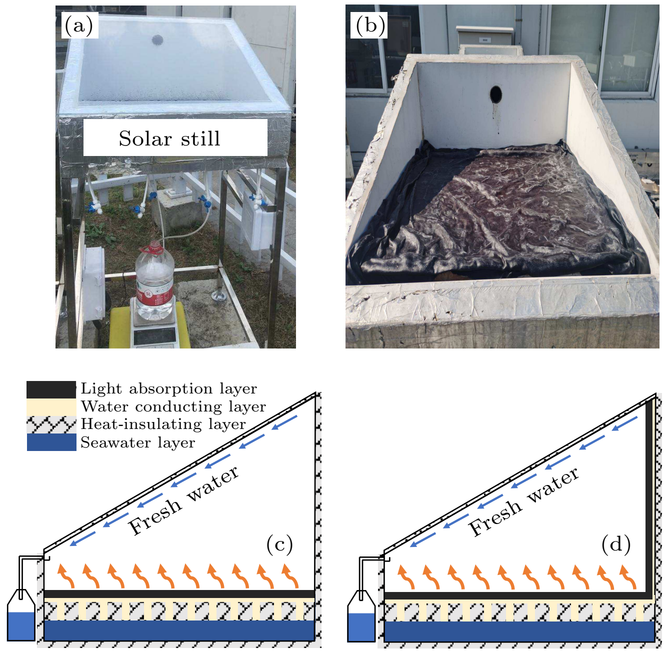

Fig. 1. (a) The photo of the solar still system for measurement in Hangzhou. (b) The internal structure of the solar still system. The diagrams of two solar still with an interfacial evaporation structure: (c) at the bottom and the insulation foams at the sidewall (BFI), and (d) on both the bottom and sidewall (BSI).

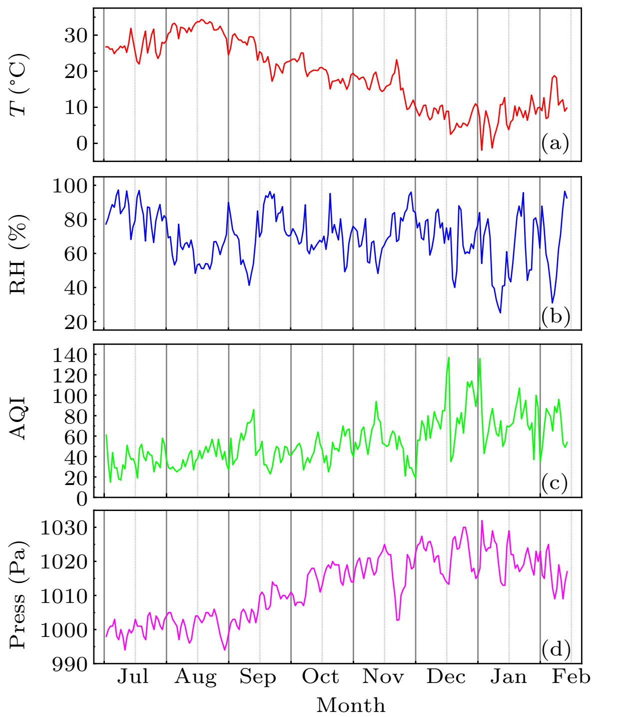

Fig. 2. The recorded weather data of Hangzhou was used as input in the model predicting the production of solar still. (a) Air temperature, (b) relative humidity, (c) air quality index, (d) atmospheric pressure.

Table 1. The test platform of meteorological data.

Name	Device model	Range	Accuracy	Resolution
Wind speed sensor	011E-MetOne	0–60 m/s	±0.1 m/s	0.04 m/s
Wind direction sensor	020C-MetOne	0–360°	±3°	< 0.1°
Environmental humidity sensor	HC2S3-Campbell	0–100% RH	±0.8% RH	0.1% RH
Atmospheric pressure sensor	CS106-Campbell	500–1100 kPa	±0.3 kPa	±0.1 kPa
Ambient temperature sensor	110PV-Campbell	−40–135 °C	±0.2 °C	–
Data logger	CR100-Campbell	0–4200 g	0.01 g	–

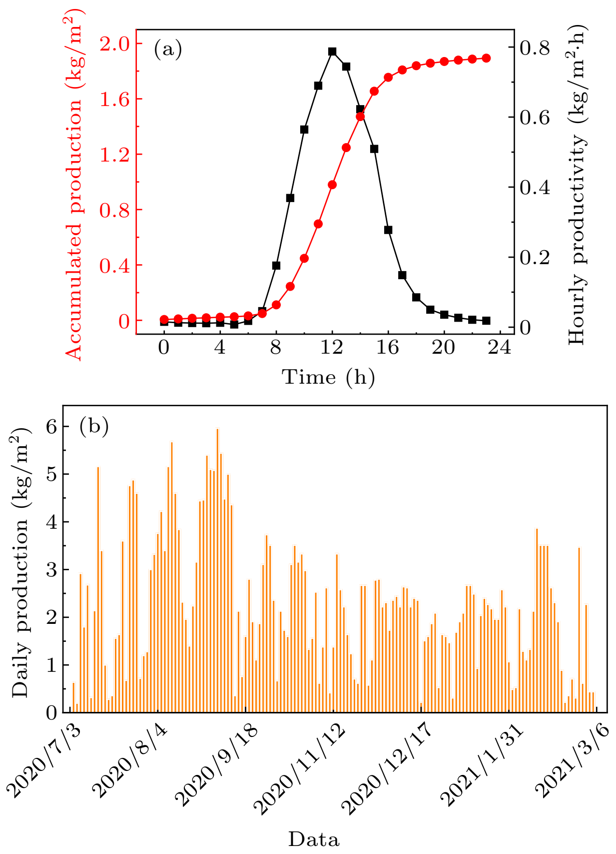

Fig. 3. (a) Both the accumulated production (red dots) and the hourly production (black squares) of the BIF solar still on March 9th, 2021. (b) The daily production of BIF-SS was measured from July 2020 to March 2021, which is a part of the dataset for building the forecasting model.

The measurements need a series of sensors which are shown in Table 1. The weather parameters were recorded every minute, including wind speed (W_{S}), wind direction (W_{D}), atmospheric pressure (Press.), air temperature (T), and relative humidity (RH). The air quality index (AQI) data is obtained from the website of www.tianqi.com. The recorded weather data of Hangzhou is shown in Fig. 2 which is expressed as daily average values. Affected by the El Niño, the average temperature in August is highest, which is significantly higher than that in July. Meanwhile, August is the driest month with the lowest average air humidity. It also can be seen from Fig. 2 that the AQI and atmospheric pressure are higher in the winter.

Figure 3(a) shows the hourly production of the BIF-SS on March 9th, the freshwater productivity gradually increases from 8:00 and reaches the highest at 12:00 about 0.8 kg/m^2·h. By 20:00, the productivity is close to 0. Figure 3(b) shows the recorded water production of the solar still from July 2020 to March 2021. Affected by the weather, the daily production varies. The freshwater production in August was the highest and significantly higher than in the other months. The highest daily production is 6.0 kg/m^2·day. The data of weather and production are listed in supporting materials (SM) I.

3. Machine learning methods

The forecasting model is established based on the dataset. The solar still dataset is given as $F = \{X, y\}_{1:i}$, where X is the input parameter, including Week, W_{S}, W_{D}, T, Press, RH, and AQI, and y is daily production, the target value corresponding to X.

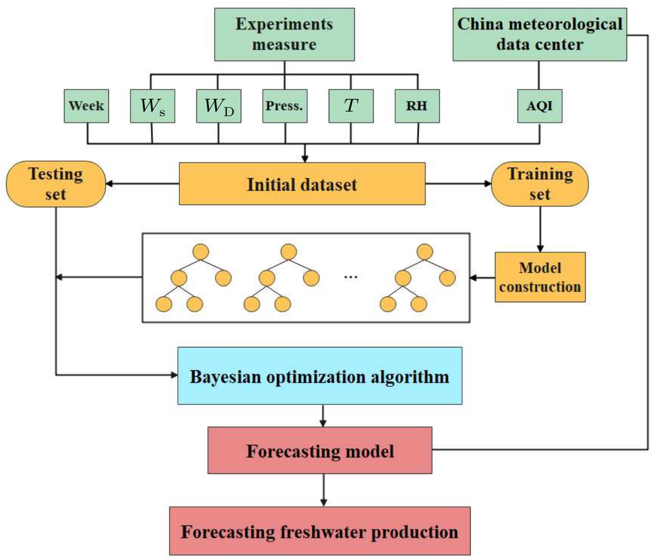

Fig. 4. The flowchart of the forecasting model includes data preprocessing, model construction, and algorithm optimization.

The basic steps include data preprocessing, model construction, and algorithm optimization. The process of data preprocessing refers to scaling the data attributes to a specific range. Because the data attributes with larger magnitudes will dominate, the accuracy of the model will be affected. The standardized method (Z-Scale) is used to scale the input data parameter. The Z-Scale method is based on the mean and standard deviation of the original data, the sample spacing can be maintained. After data standardization, the RF method is used to establish the forecasting model. First, select samples randomly, divided into training and test set. Then, build a decision tree for each piece of data, and get the predicting result. Last, vote on all the results and get the final result. The Bayesian optimization algorithm (BOA)[28] is used for searching the most appropriate hyper-parameters of the RF model. The Diagram of the forecasting model establishment is shown in Fig. 4 (Details in SM II).

4. Results and discussion

4.1. Forecasting results of RF model

Figure 5 shows the performance of the forecasting model for three different cases of testing datasets. The determination coefficient (R^2) and mean square error (MSE) are used to evaluate the performance of the forecasting model (details in SM II). With the increasing/decreasing of the size of the training/testing dataset, R^2 of the random forest models remains at a high level and improves gradually which indicates the model processes a good convergence. The value of R^2 and MSE are 0.935 and 0.209, respectively, when the test size is 10%.

The value of R^2 is much higher than that of multiple linear regression (0.767) and traditional models. For example, Kumar[20] developed a thermal model to predict the exact performance of solar stills for a different range of Grashof Numbers, the value of R^2 of Kumar's model was only 0.829. In Panchal's work,[21] the main parameters of the theoretical model were water temperature and inner glass cover temperature, and the R^2 of the model was 0.847. The results in Fig. 5 indicate that the RF method possesses a much higher predicting accuracy than traditional models (details of calculation in SM III).

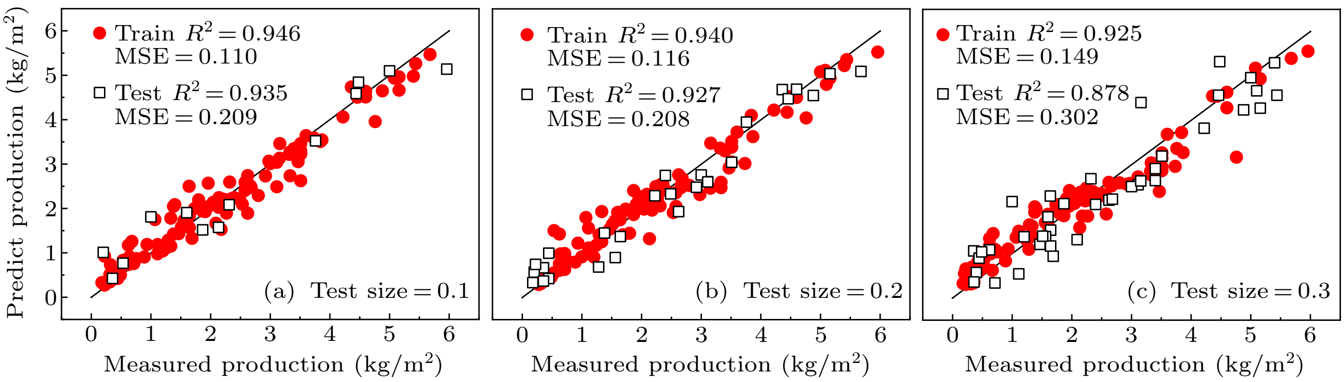

Fig. 5. For BIF-SS, the predicted values of production *versus* the measured values of production correspond to different sizes of the testing dataset, which are (a) 10%, (b) 20%, and (c) 30% of the dataset, respectively. The value of R^2 is much higher than that of multiple linear regression (0.767).

4.2. Correlation between productions and weather parameters

It was evaluated that the degree of correlation between the production of solar stills and the conventional weather forecasting parameters. The random forest method was preferred due to its superior forecasting performance. And the results are shown in Fig. 6. The three highest parameters are the daily highest temperature ($T_{\max}$), relative humidity (RH), and the daily lowest temperature ($T_{\min}$) whose values are 41%, 20%, and 18%, respectively. Moreover, Press., W_S, and W_D have similar importance values in the range of 2.3% to 3.6%, which is close to that of random orders (2.1%). The random orders were generated randomly, so it was a factor not correlated with the production and used as a normal value for comparison.

It indicates that $T_{\max}$, RH, and $T_{\min}$ are the three highest correlated factors correlating with the production. $T_{\max}$ has the highest correlation values. When the temperature rises due to increasing solar radiation, the evaporation rate will be increased. The relative humidity has a higher degree of correlation because the relative humidity directly reflects weather conditions and solar radiation. When the air humidity is high, it is usually cloudy or rainy and has low radiation intensity. Besides the three highest correlated factors, the air quality index has an importance value of 6%. AQI can also affect solar radiation energy. When the AQI is high, it means the air quality is poor and the particulate matters scatter the sunlight, which reduces the solar energy entering solar stills.

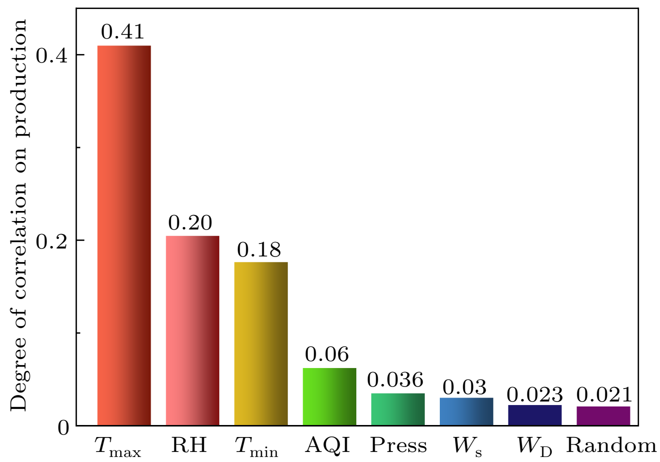

Fig. 6. The degree of correlation between the production of solar stills and the conventional weather forecasting parameters. The three parameters with the highest values are the daily highest temperature ($T_{\max}$), relative humidity (RH), and the daily lowest temperature ($T_{\min}$), whose values are 41%, 20%, and 18%, respectively.

4.3. Forecasting results between different types of solar still

A control group was set up to verify the accuracy and applicability of the predicting RF method. The solar evaporation experiments were done on the solar still with an interfacial evaporation structure at both the bottom and the sidewall (BSI-SS).

Figure 7 shows the results of the predicting performance based on the production data of BSI-SS. The predicting results are comparable to the BIF-SS. As shown in Fig. 8, 20% of production data is used as the test set. The forecasting models based on the two types of solar stills show high predicting accuracy, the R^2 of the BIF and BSI are 0.927 and 0.939. The results verify the high accuracy and applicability of the forecasting model.

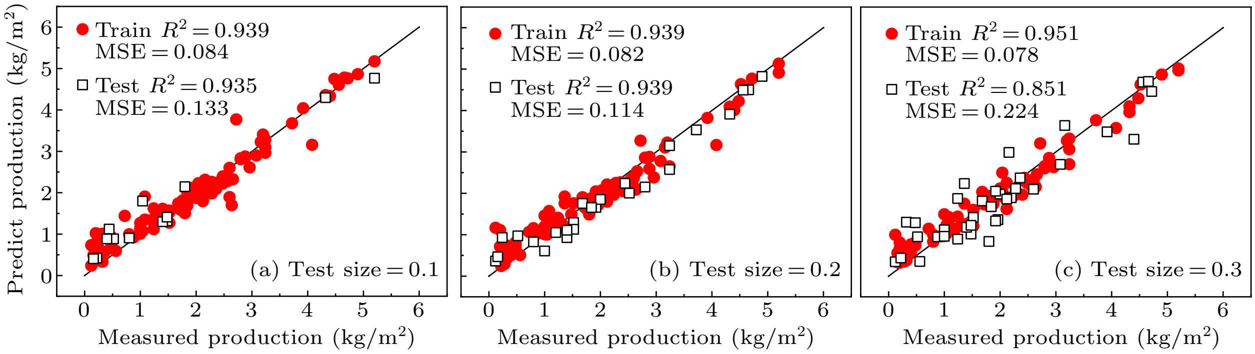

Fig. 7. For BSI-SS, the predicted values of production *versus* the measured values of production correspond to different sizes of the testing dataset, which are (a) 10%, (b) 20%, and (c) 30% of the dataset, respectively.

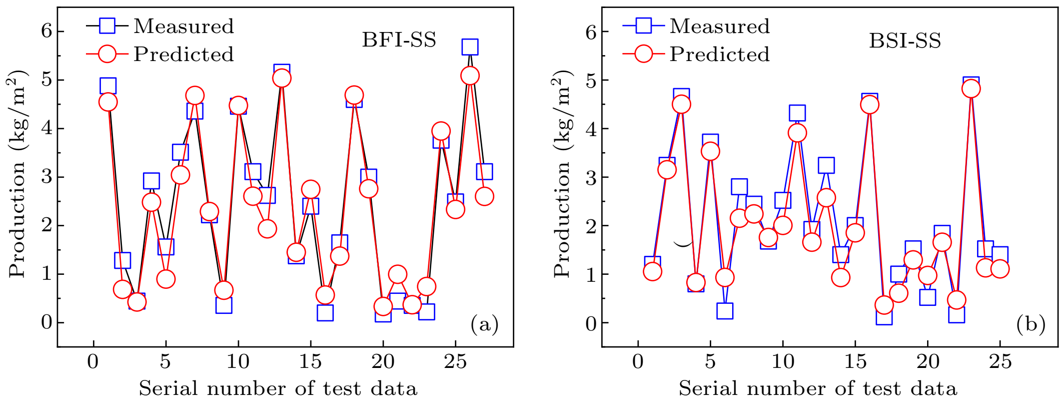

Fig. 8. Comparison of the values between the measured production data and the predicted production data, using 20% of production data as the test set. The x axis is the serial number of test data which is picked up randomly. (a) BIF-SS and (b) BSI-SS.

5. Appling forecasting model

By applying the forecasting model, freshwater production in four Chinese cities (Wuhan, Hefei, Chongqing, and Linzhi) was calculated and predicted from the weather data. It is obtained from the China meteorological data center (http://data.cma.cn) that the weather data from July 2020 to February 2021 includes air temperature, atmospheric pressure, wind speed and direction, relative humidity, and air quality index. The four cities are picked up because they have similar latitudes to Hangzhou ($\sim$ 30 N). Then, the daily productions from July 2020 to February 2021 were calculated and predicted based on the daily weather data.

The average daily production of the four cities is shown in Fig. 9. The average daily productions in Hefei and Wuhan are similar to that of Hangzhou, 2.18 kg/m^2 per day. Because the three cities have similar latitudes and are located close to the Yangtze River, that is, the climates of these three cities are similar. The production of Chongqing is the lowest among these cities, 2.1 kg/m^2 per day because Chongqing is foggy all year round and its intensity of solar radiation is lower than other cities. The production of Linzhi is the highest, 2.48 kg/m^2 per day. This is because Linzhi is located at the Qinghai–Tibet Plateau and has a high altitude (3.1 km) and insolation. The predicted daily production of the three cities was shown in SM IV.

Furthermore, the daily solar insolation data is obtained from the China meteorological data center to analyze the prediction accuracy. It needs a gauge to check the predicted values because there are no measured values of production. As shown above, solar insolation is not used in building the model. That is, the values of solar insolation are independent of the predicted production. Generally, the solar insolation is in direct proportion to the production, which can be used as a gauge to check the predicted values. Figure 10 shows the comparison of the predicted daily production and the solar insolation from July 2020 to February 2021 in Wuhan. Because of the higher/lower radiation intensity and temperature, the production should be higher in the summer/winter. The changing trend of the predicted production is similar to the daily solar insolation. And the correlation coefficient of the two data sets is 0.868, which indicates that the forecasting model possesses high accuracy.

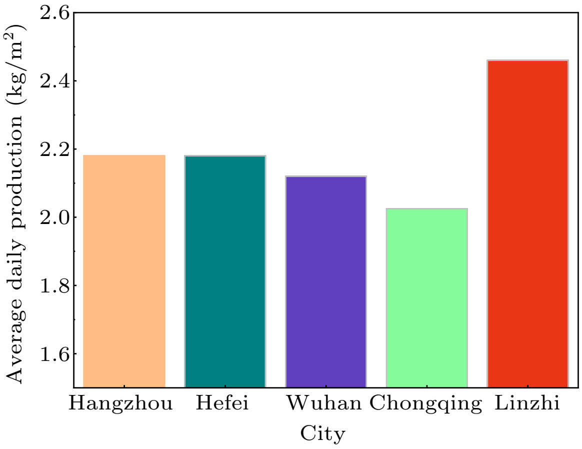

Fig. 9. The predicted average daily production of five cities in China by using the RF model. The production of Linzhi is the highest due to its high elevation and insolation. Chongqing is the lowest due to its dense mist and lowers radiation.

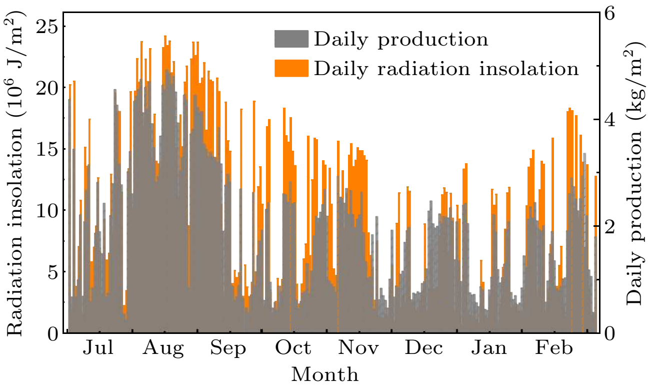

Fig. 10. A comparison of the predicted daily production and the solar insolation from July 2020 to February 2021 in Wuhan. The correlation coefficient of the predicted daily production and the solar insolation is 0.868, which indicates that the forecasting model possesses high accuracy.

6. Conclusion

In conclusion, it is proposed a method to forecast the production of solar still based on convenient weather data. The forecasting model is built by using marching learning and a measured dataset. To collect the dataset of production and weather data, a series of solar evaporation measurements on two types of solar stills (BIF-SS and BSI-SS) were performed from July 2020 to March 2021. The forecasting model was trained and established by using the random forest method, and then, optimized by the Bayesian algorithm.

Both the two forecasting models corresponding to the two solar stills have high accuracy whose determination coefficients (R^2) are much higher than the traditional model. The highest values of R^2 for BIF-SS (BSI-SS) on the training dataset and test dataset can reach 0.946 (0.951) and 0.935 (0.939), respectively.

To look for closely related weather parameters on the performance of the solar still, it was also calculated that the degree of correlation between the production and weather parameters. The three highest correlated parameters are maximum air temperature, Relative humidity, and minimum air temperature, whose degrees of correlation are 41%, 20%, and 18%, respectively.

By applying the model, the productions of BIF-SS and BSI-SS in four cities were predicted from their weather data from July 2020 to February 2021. To verify the reliability of the predicting results, the predicting results were compared with the daily solar insolation data. The correlation coefficient between predicted production and solar insolation is 0.864, indicating that the predictions have high accuracy.

There is universal applicability for our proposed idea to establish the predicting model. That is, the predicting method can be extended to any other type of stills. When forecasting the production of another type still, it just needs to follow our research processes and steps to establish another corresponding model. With the help of the forecasting model, it would greatly promote the global application of solar stills.

Data availability statement

Replication data and code can be found on the website for this project: http://nanoheat.energy.hust.edu.cn/Code_22_1.rar.

Acknowledgments

Project supported by the National Key Research and Development Program of China (Grant No. 2018YFE0127800), the Science, Technology & Innovation Funding Authority (STIFA), Egypt grant (Grant No. 40517), China Postdoctoral Science Foundation (Grant No. 2020M682411), and the Fundamental Research Funds for the Central Universities (Grant No. 2019kfyRCPY045).

References

[1] El-Samadony Y and Kabeel A E 2014 *Energy* **68** 744
[2] Abujazar M S S, Fatihah S, Lotfy E R, Kabeel A E and Sharil S 2018 *Desalination* **425** 94
[3] Kabeel A E, Arunkumar T, Denkenberger D C and Sathyamurthy R 2017 *Appl. Thermal Eng.* **114** 815
[4] Katekar V P and Deshmukh S S 2020 *Journal of Cleaner Production* **257** 120544
[5] Elimelech M and Phillip W A 2011 *Science* **333** 712
[6] Peng G L, Sharshir S W, Wang Y P, An M, Ma D K, Zang J F, Kabeel A E and Yang N 2021 *Journal of Cleaner Production* **311** 127432
[7] Mekonnen M M and Hoekstra A Y 2016 *Science Advances* **2** e1500323

[8] Sharshir S W, Peng G L, Elsheikh A H, *et al.* 2020 *Journal of Cleaner Production* **248** 119224

[9] Shalaby S M, Sharshir S W, Kabeel A E, Kandeal A W, Abosheiasha H F, Abdelgaied M, Hamed M H and Yang N 2022 *Energy Conversion and Management* **251** 114971

[10] Kandeal A W, El-Shafai N M, Abdo M R, Thakur A K and Sharshir S W 2021 *Solar Energy* **224** 1313

[11] Chen Q, Alrowais R, Burhan M, Ybyraiymkul D and Ng K C 2020 *Energy* **205** 118037

[12] Gao M, Zhu L, Peh C K and Ho G W 2019 *Energy & Environmental Science* **12** 841

[13] Peng G L, Deng S C, Sharshir S W, Ma D, Kabeel A E and Yang N 2019 *Int. J. Heat Mass Transfer* **147** 118866

[14] Chen S, Zhao P, Xie G, Wei Y and Zhang T 2021 *Desalination* **512** 115133

[15] Rahmani A, Kemmar F and Saadi Z 2021 *Desalination* **501** 114914

[16] Guo Y H, Dundas C M, Zhou X Y, Johnston K P and Yu G H 2021 *Advanced Materials* **33** 2102994

[17] SW Sharshir, Guilong Peng, L Wu, Nuo Yang, FA Essa, AH Elsheikh, SIT Mohamed and AE Kabeel 2017 *Appl. Thermal Eng.* **113** 684

[18] Cheng D, Gong W and Li N 2016 *Desalination* **394** 108

[19] Dunkle R 1961 *International Development in Heat Transfer* **5** 895

[20] Kumar S and Tiwari G N 1996 *Sol. Energy* **57** 459

[21] Panchal H 2016 *Technology and Economics of Smart Grids and Sustainable Energy* **1** 1

[22] Elango C, Gunasekaran N and Sampathkumar K 2015 *Renewable Sustainable Energy Rev.* **47** 856

[23] HAM A, MBB C, AAA D, MSZA E and AHF G 2020 *Renewable Energy* **162** 489

[24] Mashaly A F and Alazba A A 2016 *Computers and Electronics in Agriculture* **122** 146

[25] Ren Y S, Lei L, Deng X, Zheng Y, Li Y, Li J and Mei Z N 2019 *Sci. Rep.* **9** 15442

[26] Pei J, Deng L, Song S, Zhao M, Zhang Y, Wu S, Xie Y and Shi L P 2019 *Nature* **572** 106

[27] Belmokre A, Mihoubi M K and Santillán D 2019 *KSCE Journal of Civil Engineering* **23** 4800

[28] Wang Y P, Kandeal A W, Swidan A, Sharshir S W, Abdelaziz G B, Halim M A, Kabeel A E and Yang N 2020 *Appl. Thermal Eng.* **186** 116233

[29] Svetnik V 2003 *Journal of Chemical Information & Computer Sciences* **43** 1947

[30] Chan C W and Paelinckx D 2008 *Remote Sensing of Environment* **112** 2999

[31] Singh H N and Tiwari G N 2004 *Desalination* **168** 145

[32] Shi J, Luo X, Liu Z, Fan J and Bao H 2021 *Cell Reports Physical Science* **2** 100330