Obtain weighted values of solar stills'environment factors by using supervised machine learning

Authors: Wang Yunpeng, Peng Guilong, Yang Nuo, Swellam W. Sharshir, AbdAllah W. Kandeal

Date: 2021-03-26T00:00:00+00:00

Abstract

Solar stills have attracted increasing attention in recent years due to their simple structure and environmentally friendly capabilities. In this study, a mature machine learning algorithm—random forest—is utilized to obtain the weighted values of environmental factors on evaporation efficiency. To examine the advancement of random forest over mathematical data analysis, we employed two traditional data science methods—pair plot method and Pearson correlation analysis—for comparison. The experimental data used in the analysis were derived from approximately 100 articles since 2014. The results indicate that thermal design is the most important factor for achieving high-efficiency solar evaporation. This will facilitate research on the evaporation efficiency of solar stills.

Full Text

Preamble

Weighted Values of Solar Evaporation' s Environmental Factors Obtained by Machine Learning

Yunpeng Wang a,b, #, Guilong Peng a,b, #, Swellam W. Sharshir b,c, AbdAllah W. Kandeal a,b, Nuo Yang a,b *

a State Key Laboratory of Coal Combustion, and b School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
c Mechanical Engineering Department, Faculty of Engineering, Kafrelsheikh University, Kafrelsheikh 33516, Egypt

# YW and GP contributed equally to this work.

*Corresponding authors: NY (nuo@hust.edu.cn)

Abstract

Enhancing the efficiency of solar evaporation is crucial for solar still performance. In this study, we obtain the weighted values of environmental factors (descriptors) affecting solar evaporation efficiency using a machine learning algorithm —random forest. To verify the advantages of random forest over traditional mathematical data analysis, we conduct two conventional methods for comparison: pairwise plots and Pearson correlation analysis. Experimental data were compiled from approximately 100 articles published since 2014. The results demonstrate that traditional methods fail to produce reasonable weighted values, whereas random forest proves competent for this task. We find that thermal design is the most significant descriptor for achieving high efficiency. The primary challenge for more in-depth and comprehensive analysis is the lack of complete datasets. This work may advance research in solar evaporation and solar stills.

Keywords: Solar still; Solar evaporation; Machine learning; Environmental factors

1. Introduction

With population growth and the development of industrial and agricultural activities, the shortage of freshwater resources is becoming one of the most catastrophic problems facing the world. Given that seawater accounts for 97% of the planet's water resources, developing technologies for seawater desalination is essential. Many effective desalination methods have been proposed in the past, such as multistage flashing, reverse osmosis, multi-effect distillation, and vapor compression. Compared to other methods, solar stills have attracted increasing interest due to their eco-friendly nature, simple construction and maintenance, low installation cost, and long operational life. Solar evaporation is one of the crucial processes in solar stills. Consequently, over the past decades, many methods have been proposed to achieve high solar evaporation efficiency, such as using nanofluids, cotton cloth, sponge, and charcoal.

However, evaporation efficiency is affected by numerous factors, including material type, thermal design, ambient temperature, solar intensity, and others. Therefore, it is important to determine the importance or weighting of these different factors. While several important factors can be identified empirically, quantifying the importance of each descriptor is difficult, and few studies in the solar evaporation field have addressed this issue. On the other hand, quantitative analysis of descriptor importance is a widespread scientific challenge. In chemistry, for example, machine learning technology has been used to measure descriptor importance for polymer chain angles—random forest, for instance. It

has been found that machine learning can accurately measure the relationship between different factors (descriptors) and their influence on target values, representing a meaningful first step toward high-throughput screening of polymer chemistry to identify compositions with desirable bulk properties.

In the current study, we use and compare three methods for obtaining the weighted values of each descriptor in solar evaporation: two traditional data science methods—pairwise plots (PWP) and Pearson correlation analysis (PCA) —and a machine learning algorithm, random forest (RF). The weighted values are also called descriptor importance. First, the traditional data science methods are employed. Pairwise plots and Pearson correlation analysis are used to measure correlations between descriptor pairs. Then, random forest is conducted to measure the importance between evaporation efficiency and descriptors. The resulting descriptor importance can eventually help scientists design high-efficiency solar evaporation systems.

2. Methodology

The main flowchart of the current study is shown in Fig. 1 [Figure 1: see original paper]. Experimental data used in the analysis were collected from approximately 100 articles published since 2014 (details in Supporting Materials).

For the collected dataset, M = {X, y}1:t, where y is the objective value (energy efficiency of evaporation) and X represents the input descriptors corresponding to y, such as solar intensity, thermal design, surface diameter, absorptivity, Tamb (ambient temperature), and Tinterface (interface temperature). Due to the lack of details in the original articles, some data for surface diameter, absorptivity, Tamb, and Tinterface are missing. The Mean Completer method is used to fill missing data. Each descriptor is divided into three labels. The detailed distribution of dataset M is listed in Table 1 .

Three methods are investigated in this work: pairwise plots (PWP), Pearson correlation analysis (PCA), and the machine learning algorithm random forest (RF). The application of these three methods can be summarized as follows:

2.1 Pearson Correlation Analysis

PCA is typically adopted to quantify the correlation between two different descriptors. For descriptors 1 and 2 in M, such as solar intensity and solar absorptivity, PCA can be calculated as Eq.1:

$$\mathrm{PCA} = \frac{\sum(X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{[\sum(X_1 - \bar{X}_1)^2 \sum(X_2 - \bar{X}_2)^2]^{1/2}}$$

The PCA values are dimensionless and range from -1 to 1. The closer the value is to 1 or -1, the stronger the correlation between the two descriptors. A value

of 0 suggests no correlation. If the value is positive, a positive correlation exists; otherwise, a negative correlation exists.

2.2 Pairwise Plots

Pairwise plots can intuitively draw scatterplots for descriptor correlations and histograms for univariate distributions. As presented in Fig. 3 [Figure 3: see original paper], the subfigures on the diagonal represent the data distribution trend of a particular descriptor. The other subfigures help us visibly and qualitatively observe the relationship between each pair of descriptors. If a rising or falling trend forms on the diagonal, the corresponding pair of descriptors has a strong correlation. Otherwise, the two descriptors are not correlated.

2.3 Random Forest

Random forest is a typical ensemble method that combines multiple decision trees into one model to improve performance. It is widely applied in many scientific and engineering fields, such as statistics, materials science, and biology. The main steps of RF, shown in Fig. 2 [Figure 2: see original paper], can be expressed as follows:

(1) Data Preprocessing
In the present study, data representation is performed to convert data into symbols that can be read by computers. For instance, three types of thermal design—3D interface, 2D/1D interface, and volumetric—are represented as 0, 1, and 2 (details in Table 1). Finally, the dataset is divided into training set TA and test set TE according to a certain ratio.

(2) Model Construction
Based on the processed dataset TA, the bootstrap resampling method is used to randomly generate K sets of data. Then, K decision trees are grown. For example, in calculating Fig. 5a [Figure 5: see original paper], each dataset includes three descriptors (thermal design, absorptivity, and random descriptor) and the label (efficiency). In each node of a decision tree, the node splits the dataset into two parts according to the value of a chosen descriptor. After traversing all descriptors, the final node will be the label (efficiency) of this data. The model' s prediction is voted by K decision trees.

(3) Model Validation
The test dataset TE, which was not used in model construction, is employed to judge the model' s accuracy. If the accuracy of TA is much higher than that of TE (e.g., 0.9 for TA and 0.5 for TE), the model is considered overfitting. If the accuracies of both TE and TA are too low (e.g., 0.5 for TA and 0.5 for TE), the model is considered underfitting. Both overfitting and underfitting are unacceptable, and the model needs to be retrained in such cases. If the accuracies of TA and TE are high enough and the accuracy of TE is similar to that of TA, the model is considered well-trained.

3. Results and Discussion

As a starting point, two traditional data science methods—pairwise plots and Pearson correlation analysis—are performed to find the weighted values (i.e., descriptor importance). The results are displayed in Fig. 3 and Fig. 4 [Figure 4: see original paper]. In addition to mathematical data analysis, well-established machine learning algorithms are used to extract the relationship between descriptors and the target property in materials informatics. Subsequently, Fig. 5 shows the results calculated by the machine learning algorithm, random forest.

3.1 Results of Traditional Data Analysis

Fig. 3 shows the pairwise plots between different descriptors (environmental descriptors) and solar evaporation efficiency. As can be seen from the top row of plots in Fig. 3, there is no obvious linear relation between efficiency and other descriptors. Moreover, it is clear that the dataset is discrete and not evenly distributed. Most solar intensity data is around 1 kW, and the thermal design is set as discrete numbers. Since PWP is a method that focuses on the pure mathematical mapping in the dataset, it is reasonable that PWP cannot measure descriptor importance based on a defective dataset.

The PCA values are displayed in Fig. 4 [Figure 4: see original paper]. As can be seen from all investigated descriptors, all absolute descriptor values are below 0.3, which means all descriptors have weak correlation with efficiency. Therefore, similar to PWP, it is unlikely to draw reasonable results of descriptor importance based on the PCA values.

3.2 Results of Machine Learning Algorithms

Figs. 5a-5d show the descriptor importance quantified by RF using 2, 3, 4, and 6 selected descriptors, respectively. The results indicate that thermal design is the most important descriptor in solar evaporation among all chosen descriptors. The importance of thermal design is at least two times higher than that of other descriptors. This result shows that optimizing the heat transfer process in solar evaporation systems is essential for enhancing solar evaporation efficiency. This finding is reasonable because when thermal design is poor, only a small portion of solar energy is used for evaporation. For example, in traditional solar evaporation, some heat is used for heating bulk water instead of promoting evaporation. Therefore, the efficiency of solar evaporation in volumetric systems is lower than that in interface systems.

Besides, Fig. 5 shows that solar intensity is an unimportant descriptor. This is because, with optimized thermal and material design, high efficiency can be obtained regardless of high or low solar intensity, as reported in many works. Therefore, solar intensity is not important and is similar to a random descriptor. Herein, the random descriptor is a set of random data that has no relationship to energy efficiency and is used as a benchmark.

Meanwhile, the descriptor importance of solar absorptivity is much lower than expected, as shown in Fig. 5. Higher absorptivity enables more available energy for evaporation and should significantly affect efficiency. The reason may be that almost all reported works selected materials with very high absorptivity (>90%), which prevents the dataset from achieving ergodicity. Hence, its importance is underestimated in the calculation. Besides, the ambient temperature (Tamb) and evaporation interface temperature (Tinterface) are insignificant, which might be due to the small differences in Tamb and Tinterface between most works. However, temperature is actually a very important descriptor in natural convection-based evaporation processes. Therefore, to capture the real importance of temperature, more work should be conducted at different ambient and interface temperatures.

It can be concluded that more accurate calculations by machine learning require more complete data. Compared to other fields such as materials science, which have complete databases of physical properties and theoretical calculation methods, the current study faces the hard problem of insufficient reported data because authors do not provide exact values for some descriptors. For example, values for ambient temperature, evaporation surface diameter, absorptivity, and evaporation interface temperature are missing in some papers. Therefore, to obtain more accurate machine learning results, authors should provide complete datasets of experimental descriptors in their future works. On the other hand, some other potentially important descriptors on material design, such as thermal conductivity, contact angle, specific area, porosity, characteristic size, functional groups, and so forth, are not included and calculated by RF at the current stage because detailed material properties are not provided in most papers. It is worth noting that a full dataset of descriptors in research reports will help push the field forward.

3.3 Effect of Dataset Size

As mentioned in other research, the quality of the dataset determines the reliability of machine learning algorithms. To avoid overfitting and further quantify the effect of dataset size in the current study, the initial dataset was separated into three combinations of train/test sets: 70%/30%, 80%/20%, and 90%/10%, respectively. The results are summarized in Fig. 6 [Figure 6: see original paper]. As can be seen, with the decrease of the test set, there is no obvious difference between different models. Thermal design is the most important descriptor in all cases. Absorptivity is the second most important descriptor. Other descriptors are not important and are similar to a random descriptor. This demonstrates that the result of descriptor importance depends on the physical mechanism rather than the dataset, indicating that the current dataset is able to achieve a convergent solution.

4. Conclusion

In conclusion, the importance of factors affecting solar evaporation efficiency is analyzed using pairwise plots, Pearson correlation analysis, and random forest. Experimental data used in the analysis were collected from approximately 100 articles. The results indicate that pairwise plots and Pearson correlation analysis cannot measure descriptor importance based on defective datasets. In contrast, random forest can obtain reasonable results. The random forest results show that thermal design is the most important descriptor determining solar evaporation efficiency. It can be concluded that machine learning is helpful for quantitatively understanding the importance of various descriptors, which will help advance the solar still field.

Although machine learning obtained meaningful results, it should be emphasized that due to limitations in the amount and quality of experimental data in published articles, the current analysis provides more qualitative than quantitative results. It is expected that authors can provide more detailed data and standardized descriptors in future publications. This will promote the application of machine learning in solar still research.

Conflicts of Interest
There are no conflicts of interest to declare.

Acknowledgment

This work was sponsored by the National Key Research and Development Project of China (2018YFE0127800), China Postdoctoral Science Foundation (2020M682411), National Natural Science Foundation of China (51950410592), Fundamental Research Funds for the Central Universities (2019kfyRCPY045), and Graduate Innovation Funds for Huazhong University of Science and Technology (2020yjsCXCY067). The authors thank the National Supercomputing Center in Tianjin (NSCC-TJ) and China Scientific Computing Grid (ScGrid) for providing computational assistance.

References

Al-Othman, A., Tawalbeh, M., El Haj Assad, M., Alkayyali, T. & Eisa, A. Novel multi-stage flash (MSF) desalination plant driven by parabolic trough collectors and a solar pond: A simulation study in UAE. Desalination 443, 237-244, doi:https://doi.org/10.1016/j.desal.2018.06.005 (2018).

Shaaban, S. Performance optimization of an integrated solar combined cycle power plant equipped with a brine circulation MSF desalination unit. Energy Conversion and Management 198, 111794, doi:https://doi.org/10.1016/j.enconman.2019.111794 (2019).

Farsi, A. & Dincer, I. Development and evaluation of an integrated MED/membrane desalination system. Desalination 463, 55-68, doi:https://doi.org/10.1016/j.desal.2019.02.015

(2019).

Sadri, S., Ameri, M. & Haghighi Khoshkhoo, R. Multi-objective optimization of MED-TVC-RO hybrid desalination system based on the irreversibility concept. Desalination 402, 97-108, doi:https://doi.org/10.1016/j.desal.2016.09.029 (2017).

Sharshir, S. W. et al. Improving the solar still performance by using thermal energy storage materials: A review of recent developments. Desalination and water treatment 165, 1-15, doi:https://doi.org/10.5004/dwt.2019.24362 (2019).

Sharshir, S. W. et al. Augmentation of a pyramid solar still performance using evacuated tubes and nanofluid: Experimental approach. Applied Thermal Engineering 160, 113997, doi:https://doi.org/10.1016/j.applthermaleng.2019.113997 (2019).

Trisaksri, V. & Wongwises, S. Critical review of heat transfer characteristics of nanofluids. Renewable Sustainable Energy Reviews 512-523, doi:https://doi.org/10.1016/j.rser.2005.01.010 (2007).

Elsheikh, A. H., Sharshir, S. W., Mostafa, M. E., Essa, F. A. & Ahmed Ali, M. K. Applications of nanofluids in solar energy: A review of recent advances. Renewable and Sustainable Energy Reviews 82, 3483-3502, doi:https://doi.org/10.1016/j.rser.2017.10.108 (2018).

Kalidasa Murugavel, K. & Srithar, K. Performance study on basin type double slope solar still with different wick materials and minimum mass of water. Renewable Energy 36, 612-620, doi:https://doi.org/10.1016/j.renene.2010.08.009 (2011).

Abu-Hijleh, B. A. K. & Rababa' h, H. M. Experimental study of a solar still with sponge cubes in basin. Energy Conversion Management 1411-1418, doi:https://doi.org/10.1016/S0196-8904(02)00162-0 (2003).

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. Nature 559, 547-555, doi:https://doi.org/10.1038/s41586-018-0337-2 (2018).

Ouyang, Y. et al. Accuracy of Machine Learning Potential for Predictions of Multiple-Target Physical Properties. Chinese Physics Letters 37, 126301, doi:https://doi.org/10.1088/0256-307x/37/12/126301 (2020).

Brochu, E., Cora, V. M. & De Freitas, N. J. a. p. a. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599 (2010).

Wan, X. et al. Materials Discovery and Properties Prediction in Thermal Transport via Materials Informatics: A Mini Review. Letters 3387-3395, https://doi.org/10.1021/acs.nanolett.8b05196 (2019).

Ma, R., Huang, D., Zhang, T. & Luo, T. Determining influential descriptors for polymer chain conformation based on empirical force-fields and

molecular dynamics simulations. Chemical Physics Letters 704, 49-54, doi:https://doi.org/10.1016/j.cplett.2018.05.035 (2018).

Wang, Y. et al. Prediction of tubular solar still performance by machine learning integrated with Bayesian optimization algorithm. Applied Thermal Engineering 116233, doi:https://doi.org/10.1016/j.applthermaleng.2020.116233 (2021).

Ma, R., Liu, Z., Zhang, Q., Liu, Z. & Luo, T. Evaluating Polymer Representations via Quantifying Structure–Property Relationships. Journal of Chemical Information and Modeling 59, 3110-3119, doi: https://doi.org/10.1021/acs.jcim.9b00358 (2019).

Chen, X., Wang, M. & Zhang, H. The use of classification trees for bioinformatics. WIREs Data Mining and Knowledge Discovery 1, 55-63, doi:https://doi.org/10.1002/widm.14 (2011).

Calzetta, L. et al. Pharmacological treatments in asthma-affected horses: A pair-wise and network meta-analysis. Equine Veterinary Journal 710-717, doi:https://doi.org/10.1111/evj.12680 (2017).

Månsson, R. et al. Pearson Correlation Analysis of Microarray Data Allows for the Identification of Genetic Targets for Early B-cell Factor*[boxs]. Journal of Biological Chemistry 279, 17905-17913, doi:https://doi.org/10.1074/jbc.M400589200 (2004).

Breiman, Random Forests. Machine Learning 5-32, https://doi.org/10.1023/A:1010933404324 (2001).

Maltecca, C. et al. Predicting Growth and Carcass Traits in Swine Using Microbiome Data and Machine Learning Algorithms. Scientific Reports 9, 6574, doi: https://doi.org/10.1038/s41598-019-43031-x (2019).

Palmer, D. S., O' Boyle, N. M., Glen, R. C. & Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. Journal of Chemical Information and Modeling 47, 150-158, doi: https://doi.org/10.1021/ci060164k (2007).

Tao, P. et al. Solar-driven interfacial evaporation. Nature Energy 3, 1031-1041, doi: https://doi.org/10.1038/s41560-018-0260-7 (2018).

Peng, G. et al. High efficient solar evaporation by airing multifunctional textile. International Journal of Heat and Mass Transfer 118866, doi:https://doi.org/10.1016/j.ijheatmasstransfer.2019.118866 (2020).

Chen, C., Kuang, Y. & Hu, L. Challenges and Opportunities for Solar Evaporation. Joule 3, 683-718, doi:https://doi.org/10.1016/j.joule.2018.12.023 (2019).

Sharshir, S. W. et al. Influence of basin metals and novel wick-metal chips pad on the thermal performance of solar desalination process. Journal of Cleaner Production 248, 119224, doi:https://doi.org/10.1016/j.jclepro.2019.119224 (2020).

Poós, T. & Varju, E. Review for prediction of evaporation rate at natural convection. Heat and Mass Transfer 55, 1651-1660, doi: https://doi.org/10.1007/s00231-

018-02535-4 (2019).

Wang, Y. et al. A New Machine Learning Algorithm to Optimize A Reduced Mechanism of 2-Butanone and the Comparison with Other Algorithms. ES Materials & Manufacturing 6, 28-37, doi: https://doi.org/10.30919/esmm5f615 (2019).

Weber, D. G. G. a. R. L. S. a. H. K. M. a. B. W. Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, Transport Processes. https://doi.org/10.5281/zenodo.4527812 (2021).

Note: Figure translations are in progress. See original paper for figures.