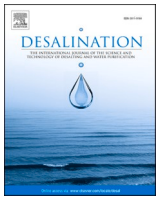
Desalination 587 (2024) 117946

Contents lists available at ScienceDirect

Desalination

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



A data-driven method to construct prediction model of solar stills

Senshan Sun [a,b], Juxin Du [a], Guilong Peng [c,*], Nuo Yang [d,*]

[a] School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[b] School of Integrated Circuits, Wuhan National Laboratory for Optoelectronics (WNLO), Key Laboratory of Material Chemistry for Energy Conversion and Storage, Huazhong University of Science and Technology, Wuhan 430074, China
[c] School of Mechanical and Energy Engineering, Shaoyang University, Shaoyang 422000, China
[d] Department of Physics, College of Science, National University of Defense Technology, Changsha 410073, China

HIGHLIGHTS

- A new data-driven method is proposed which is superior to the expert-driven method.
- Data-driven method integrates data acquisition and model construction in real-time.
- The data-driven method is more effective in 70 % of the comparisons.
- A 14.7 % reduction in required data size can be achieved.

ARTICLE INFO

Keywords:
Solar stills
Machine learning
Data acquisition
Production predicting
Process optimization

ABSTRACT

The interdisciplinary field between solar desalination and machine learning is the subject of a cutting-edge study. Generally, the studies treat data acquisition and model construction as independent processes, leading to problems such as insufficient dataset size or resource wastage. This study proposes a data-driven method that integrates data acquisition with model construction processes. By using the Bayesian optimization algorithm, the method accelerates the convergence of model accuracy. By comparing the results of 100 pairs of simulations, it is found that the models using the data-driven method are more accurate than traditional expert-driven methods in 70 % of compared results. Additionally, when it makes a model with the mean absolute percentage error as 5 %, the proposed data-driven method requires 220 additional data on average, compared to 258 with the traditional expert-driven method, representing a 14.7 % reduction. This work offers new ways and a broad application of the interdiscipline between solar desalination and machine learning.

1. Introduction

Freshwater scarcity remains one of the pressing global challenges today, with approximately 2 billion people worldwide lacking access to safely managed drinking water services [1]. In contrast to freshwater resources, seawater resources on Earth are abundant. Therefore, desalination of seawater can be employed to address freshwater shortages in coastal cities [2]. Solar stills are commonly used solar-powered desalination devices known for their compact structure, small footprint, and ease of operation and maintenance [3]. Moreover, compared to other commercially scaled desalination technologies, solar stills are considered more environmentally friendly [4], thus garnering broader attention from researchers.

In recent years, researchers have explored various methods to enhance the productivity of solar stills, including the addition of nanoparticles [5], phase change materials [6], structural optimization [7], and external auxiliary devices [8]. Although the productivity of solar stills has been improved, constructing an accurate prediction model for solar still production remains a major challenge. Many researchers have obtained different models of solar stills through thermodynamic calculations. Peng et al. [9], grounded in thermodynamics, assumed the most ideal model for heat and mass transfer in the solar stills, and proceeded to calculate the theoretical upper limit of the solar still's performance. A. Mohamed et al. [10] conducted theoretical research on the performance of a newly designed rhombic solar still, considering radiation exchange between the surfaces of the still, and proposed a novel theoretical model for solar stills. However, like others, this model has certain limitations and limited application scenarios.

* Corresponding authors.
E-mail addresses: 4195@hnsyu.edu.cn (G. Peng), nuo@nudt.edu.cn (N. Yang).

https://doi.org/10.1016/j.desal.2024.117946
Received 21 May 2024; Received in revised form 20 July 2024; Accepted 20 July 2024
Available online 22 July 2024

On the other hand, machine learning (ML) methods offer a new avenue for predicting and optimizing solar stills. Various ML methods, which have been widely studied in other fields, could be used in the solar still field, such as Generative Artificial Intelligence (GAI) [11] and meta-learning [12]. Wang et al. [13] used a random forest model to predict the output of tubular solar stills, and calculated the optimal hyper-parameters through Bayesian optimization. They achieved an R-squire (R^2) value of 0.976 for the test set fit. Gao et al. [14] proposed a model for predicting solar still output using conventional weather data and applied the model to four other cities, achieving a correlation coefficient of 0.868 between output and irradiance, demonstrating the reliability of the prediction model. Ammar et al. [15] used long short-term memory neural networks (LSTM) to predict the output of stepped solar stills, achieving an R^2 value of 0.99 for the test set fit. Santos et al. [16] studied the use of artificial neural networks (ANN) to predict solar still output and explored the impact of different variable combinations on prediction results.

Furthermore, in the application of ML methods to the solar still field, it is essential to understand the relationship between model accuracy and dataset characteristics, for example, data quality [17,18] and dataset size [19,20]. Larger datasets do not invariably equate to superior performance [21]. Especially in the field of solar stills, Peng et al. [22] through extensive experimentation and modeling research, found that too small datasets greatly reduce the generalization ability of machine learning prediction models. Conversely, when the dataset reaches a certain size, the accuracy of the prediction model does not necessarily increase with the increase in data volume. This means that there is a delicate relationship between dataset size and model accuracy.

However, previous studies (expert-driven method) typically involved the establishment of a dataset through experimental data to train machine learning models, with data acquisition and model construction treated as independent processes. These approaches often overlooked the relationship between dataset size and model accuracy during the data acquisition process. Moreover, as Fig. 1 shows, dataset sizes were often determined based on researchers' experience. Typically, researchers would adjust parameters intermittently to generate different datasets over several days of experimentation, halting data collection once they deemed it sufficient. This practice introduces significant uncertainty into the data acquisition process and can result in either insufficient or excessive datasets. Thus, it is imperative to devise a method that bridges the gap between the data acquisition and model construction processes of the solar still field. Active learning [23], and reinforcement learning [24] could be the possible ways, which have been widely used in other fields, such as the advanced materials field for optimizing data acquisition procedures.

The primary objective of this study is to leverage machine learning optimization algorithms to establish a connection between the data acquisition and model construction processes (data-driven method). By utilizing existing data, the optimization algorithm generates a set of recommended experimental parameters to guide subsequent experiments. All parameters are determined exclusively through machine learning algorithms. Furthermore, the algorithm continuously monitors model accuracy in real-time, providing researchers with guidance on when to terminate experiments. This approach ensures that the dataset size is sufficiently large to achieve optimal accuracy while avoiding the wastage of experimental data.

In this work, a substantial experimental dataset is first collected. Using this dataset, a highly accurate machine learning prediction model is trained, tested, and compared with experimental values. Subsequently, this high-accuracy model is employed to validate the proposed data-driven method. Finally, the effectiveness of the data-driven data acquisition method is compared with the expert-driven method.

2. The experiment and methods

2.1. Experiment system

A solar still experimental setup was constructed to acquire data (Fig. 2). This system primarily consists of three components: the solar still system, the control system, and the measurement system.

In the solar still system, three types of solar stills were utilized for experimentation and research: single-slope, double-slope, and pyramid. Each solar still is covered with an insulated layer using 4 cm XPS (extruded polystyrene) foam. The water tank has dimensions of 25 cm × 25 cm, with an electrical heating panel attached to the bottom to simulate solar heating. Within the inner chamber, a fan enhances air circulation. A thermostat cover placed above the glass cover simulates ambient temperature, with a distance of about 1 cm between the glass cover and the thermostat cover. The inside surface of the glass cover is treated with an anti-fog coating, rendering it ultra-hydrophilic. Additionally, a fibrous water channel and a water-leading wire are incorporated to guide the condensed freshwater stream from the glass cover to the collection bottle. These optimized designs have been proven to significantly reduce production fluctuations and experiment time [22].

In the control system, the fan and electrical heating panel are controlled by different DC power sources to adjust input power. The thermostat cover temperature is controlled by a thermostat water bath to simulate various ambient temperatures. This setup enables the convenient achievement of different working conditions.

In the measurement system, it encompasses all measuring devices,

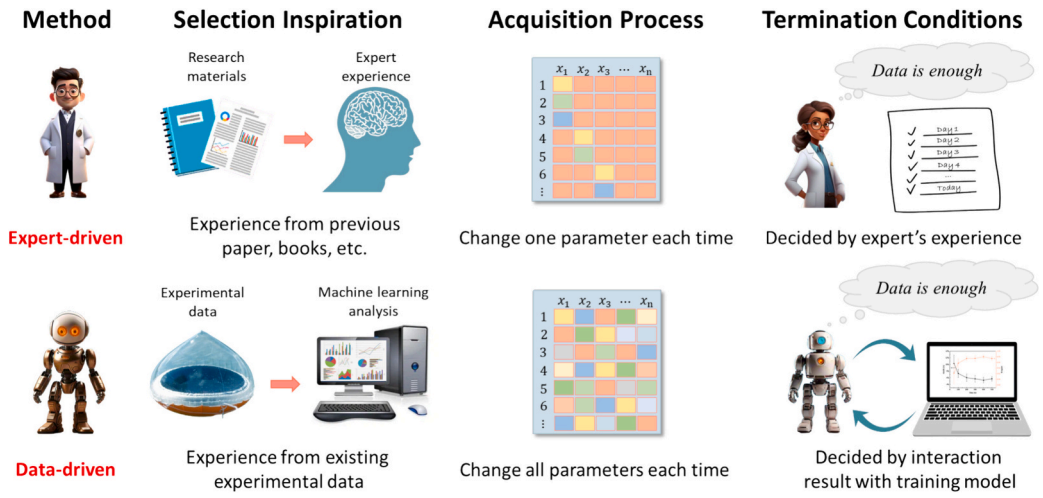


Fig. 1. The difference between traditional expert-driven and novel data-driven methods (Figure source refers to Table S1 in the Supplementary Information).

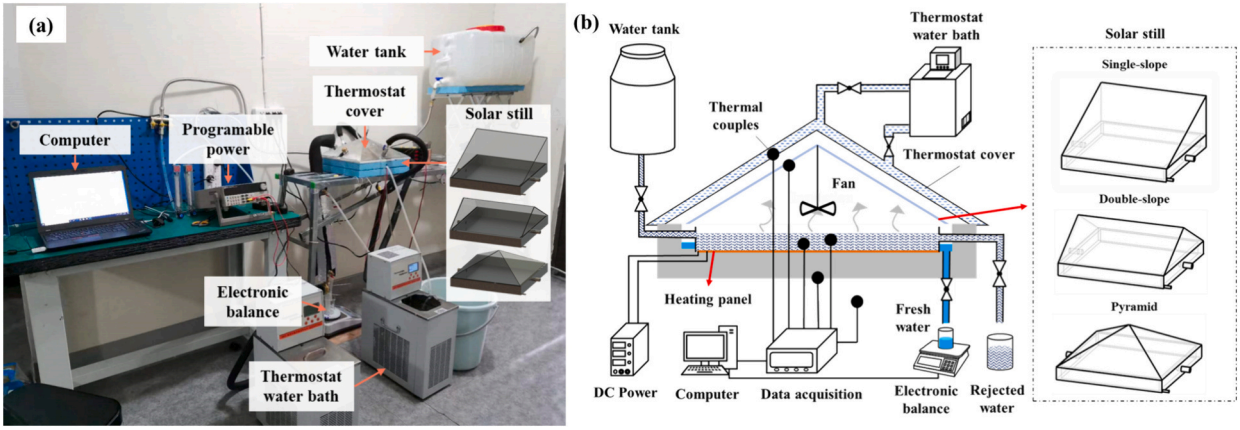


Fig. 2. The experimental setup (a) Photo (b) Schematic diagram.

such as the data acquisition unit, electronic balance, and thermal couple. The specific temperature measurement locations are shown in Fig. 2(b). All devices and sensors used in the experiment are listed in Table S2 in the Supplementary Information. Based on the experimental setup, four input parameters and one productivity value were collected, including water temperature, fan voltage, ambient temperature, solar still types, and freshwater productivity. These input parameters have been shown to significantly influence productivity [22]. Therefore, a large dataset comprising 813 experimental data was established. The pair plot and correlation heatmap of the collected data can be found in Fig. S1 and S2 in the Supplementary Information.

2.2. Machine learning assistance model

To rapidly and accurately establish a machine learning prediction model, a data-driven data acquisition method using machine learning is proposed (Fig. 3).

Initially, data are randomly acquired through experimentation to create the initial dataset. To achieve optimal effectiveness, the initial data should be as diverse as possible. For example, this study constructs an initial dataset with 20 data, with productivity ranging from 0.1 kg/(m^2·h) to 1.1 kg/(m^2·h).

Using this initial dataset, an initial prediction model is established. The backpropagation neural network (BPNN) is a common machine learning regression algorithm widely used in property prediction and engineering [25,26], particularly in solar still machine learning prediction models [27]. Therefore, this study selects the BPNN model to construct the prediction model. With four input parameters, to prevent overfitting and minimize model complexity, two hidden layers are considered sufficient to adequately fit the dataset in this study.

In the forward pass process, the output value can be computed sequentially from the left side to the right side. For simplicity, bias terms

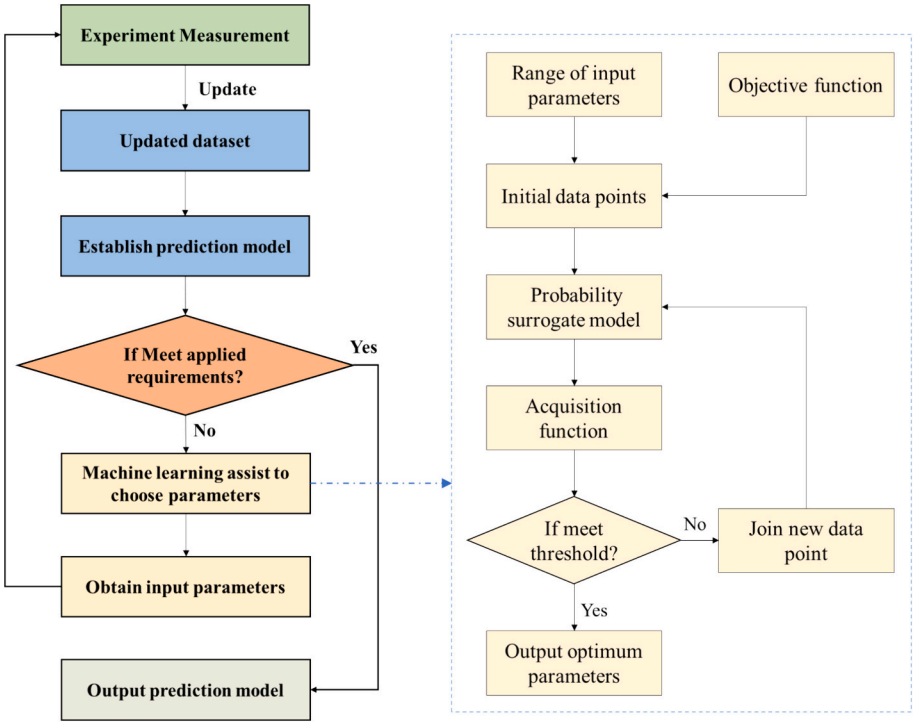


Fig. 3. The flowchart of the data-drive method based on machine learning. In this work, the applied requirement is the R-squire of the prediction model. The optimization model is Bayesian optimization, and the threshold is defined as the R-squire. (the detailed objective function can refer to Note S3 in the Supplementary Information).

are omitted. Each unit value can be represented as

$$z_j = \sum w_{ij} x_i \tag{1}$$

$$y_j = f(z_j) \tag{2}$$

Where i, j are respectively the labels of the current and next layer, z_j is the input value of the unit in the next layer, w_{ij} is the weight between two units, x_i is the value of the unit in the current layer, y_i is the value of the unit in the next layer. $f(.)$ is the activation function.

In the backward pass process, the weights between two units can be adjusted, with the computation proceeding from the right side to the left side. Each updated weight can be represented as

$$w' = w + \Delta w = w - \frac{\partial E}{\partial w} \tag{3}$$

Where $\frac{\partial E}{\partial w}$ is error derivatives. In the output layer, the cost function E can be chosen

$$E = 0.5(o - t)^2 \tag{4}$$

Where o is the output value of the BPNN model, t is the true value. The error E in the other layers can be represented

$$\frac{\partial E_i}{\partial z_i} = \frac{\partial E_i}{\partial y_i} \frac{\partial y_i}{\partial z_i} \tag{5}$$

$$\frac{\partial E_i}{\partial y_i} = \sum w_{ij} \frac{\partial E_j}{\partial z_j} \tag{6}$$

Δw can be represented

$$\Delta w_{ij} = -y_i \frac{\partial E_j}{\partial z_j} \tag{7}$$

Next, the evaluation of the prediction model should be compared against the applied requirements, typically represented by a threshold value. If focusing on absolute error, this threshold can be the mean absolute error; if relative error is significant, the mean absolute percentage error may be used. In this study, the R-squared value is set as the threshold value, considering the overall quality and universality of the prediction model. However, due to the limited amount of initial data, the initial prediction model typically lacks high accuracy and seldom meets this threshold value.

In the subsequent step, an optimization algorithm is employed based on the obtained prediction model to assist in selecting the next experimental parameters. This step represents a self-corrective process, utilizing the existing prediction model and an optimization algorithm to provide potential next experimental parameters aimed at improving model accuracy. This process forms a virtuous cycle between the prediction model and improvement effectiveness. As the accuracy of the prediction model increases, so does the improvement in effectiveness, leading to a more accurate prediction model. Conversely, when the dataset is too small, the improvement effectiveness may be weak. However, even a slight increase in data can modify the prediction model, enhancing the effectiveness of the improvement process.

Bayesian optimization (BO) is regarded as one of the most outstanding optimization algorithms. Ghahramani, the former chief scientist at Uber, considers Bayesian optimization as one of the cutting-edge technologies in probabilistic machine learning [28]. Therefore, this study utilizes the BO algorithm as an example to validate the proposed method of machine learning assistance. The core concept of Bayesian optimization is to construct a model that can be iteratively updated and queried to guide optimization decisions [29]. The detailed optimization process is illustrated in Fig. 3. In the phase of selecting the next experimental parameters using machine learning, the objective function needs to be initially formulated. It can be represented as

$$Index = Fun(Ps^*, Ds) \tag{8}$$

Where *Index* is the evaluating index of the prediction model, and it is R-square (R^2) in this instance. $Fun(.)$ is the objective function. *Ds* is the data set, which is updated in the former part but is invariable in this optimization part to provide a data-driven experience. Ps^* is target parameter set including $T_w{}^*, T_{amb}{}^*, U_f{}^*, Type^*$ which are new parameters in the next experiment. This function indicates that when new parameters Ps^* are selected as the experimental parameters, a probable R-square is generated. This R-square is computed by training the dataset with the addition of new data Ps^* and its corresponding production. The production under the Ps^* condition can be acquired using the internal prediction model within the objective function. For detailed construction, please refer to Note S3 in the Supplementary Information.

Based on the objective function and the range of *Ps*, a few initial data points can be randomly generated. Then, the probabilistic surrogate model, a key component in BO, can be fitted. In this section, Gaussian process regression serves as the probabilistic surrogate model, with the Matern 2.5 kernel function utilized. This can be represented as follows:

$$R^2 = GP(T_w, T_{amb}, U_f, Type) \tag{9}$$

The acquisition function is employed to select the optimum *Ps* using the probabilistic surrogate model. The confidence boundary strategy, widely applied in the field of K-arm gambling machines [30], has been selected for the acquisition function in this study, specifically the Upper Confidence Bound (UCB) function, represented as:

$$\alpha_t(Ps^*; Ps_{1:t}) = \mu(Ps_{1:t}) + \sqrt{\beta_{1:t}}\sigma(Ps_{1:t}) \tag{10}$$

Where μ is expectation; σ is variance; β is a constant value to balance expectation and variance, and $\sqrt{\beta}$ is 2.576 quoted the setting of Python package "bayes_opt" [31]; t is the number of data in *Ds*.

When new Ps^* and R^{2*} are selected by acquisition function, R^{2*} should be compared with the set threshold. If it meets the threshold, output this Ps^* as the next experiment parameters. If not, put Ps^* into *Ds* and fit the new probabilistic surrogate model, then repeat this process until R^{2*} meets the threshold.

Consequently, a set of experimental input parameters can be obtained, and subsequent experiments can reference these parameters. The new experimental data are then incorporated into the dataset, and this process is repeated until the prediction model's accuracy meets the applied requirements.

Table 1 illustrates the difference in data acquisition methods between expert-driven and data-driven methods. The expert-driven method relies on the experience of experts, which is often based on previous research materials. However, this method provides limited assistance, as research objectives and models can vary widely. Experts typically adjust one parameter at a time to collect different data. While this method is efficient for data acquisition, it does not consider the relationship between dataset size and model accuracy. Experts often terminate experiments based on their judgment after several days of experimentation, deeming the data collection sufficient.

In contrast, the data-driven method relies on existing experimental

Table 1
The difference in data acquisition methods between traditional expert-driven and novel data-driven methods.

Specifications	Expert-driven	Data-driven
Selection inspiration	Most from previous researches	Most from acquired data
Acquisition process	Change one of the parameters each time	Change all parameters each time
Termination conditions	Expert's experience	Feedback result with training model

data from the current research system, providing more targeted guidance. In this method, all parameters are determined by machine learning algorithms. Consequently, the model accuracy generally improves as more data is collected. This method allows for real-time monitoring of the experiment, enabling researchers to decide whether to continue or terminate the experiment based on the model's performance. The detailed procedures for the two data acquisition methods are shown in Note S4 in the Supplementary Information.

In this work, a highly accurate prediction model is employed to simulate the processes of solar still experiments, aiming to comprehensively demonstrate the superiority of the data-driven method. The conventional data acquisition method is subject to occasional randomness, necessitating numerous experiments to substantiate the proposed approach. Conducting numerous comparisons to draw general conclusions would be labor-intensive and time-consuming. To alleviate this burden and enhance feasibility, the study utilized a highly accurate model trained on 813 experimental data to simulate experiments. While the accurate model still requires a substantial amount of data, this approach remains practical and valid. Table S3 in the Supplementary Information Note S5 provides a comparison between experimental and simulation data acquisition methods.

3. Results and discussions

To demonstrate the effectiveness of the highly accurate prediction model in simulating experiments, Fig. 4 illustrates the independent validation of dataset size and the scatter points for the BPNN prediction model. In Fig. 4(a), two evaluation indices, R-square (R^2) and mean absolute percentage error (MAPE), are computed. As the dataset size increases, MAPE gradually decreases. When the dataset size exceeds 200, MAPE stabilizes at approximately 6 %. Similarly, R^2 stabilizes around 0.98 once the dataset size exceeds 200. The mean absolute error is illustrated in Fig. S4 in the Supplementary Information Note S6, showing a similar trend to R^2. Although additional data may not significantly enhance the accuracy of the prediction model, it does not have a detrimental effect. Therefore, a dataset comprising 813 experimental data is used to train the BPNN prediction model. Fig. 4(b) presents the distribution of scatter points around the optimized BPNN prediction model programmed. Consistently meeting expectations, the scatter points closely align around the diagonal (black dotted line), representing equal values. For the testing set, the R^2 and MAPE are 0.99 and 3.4 %, respectively. Furthermore, a comparison between two Python packages for the BPNN model is detailed in the Supplementary Information Note S7.

To further demonstrate the accuracy of the prediction model, the visual performance of the BPNN model is shown in Fig. 5. The orange line represents the predicted production, while the green stars indicate the experimental data points. In Fig. 5(a), as the water temperature increases, the production also increases, with the experimental points aligning closely with the prediction line. Fig. 5(b-d) illustrates the range of predicted production (yellow area) between the upper and lower water temperatures under identical conditions. During the measurement, the water temperature varies slightly under the same input power of the electrical heating panel. For instance, in Fig. 5(b), where a constant water temperature of 50 °C is targeted, slight fluctuations are observed, with temperatures ranging from 47.5 °C to 52.5 °C. Consequently, the upper and lower lines represent the predicted production at water temperatures of 52.5 °C and 47.5 °C, respectively. Despite the complexity of the relationship between parameters and production, the BPNN prediction model demonstrates excellent predictive capability, as evidenced by the agreement between experimental and predicted values. These results affirm that the prediction model is sufficiently accurate to simulate the experimental process, making this approach applicable across numerous cases to highlight the advantages of using machine learning for data acquisition.

To ensure consistent and robust findings, a comparison is conducted across 100 pairs of cases. In each pair, the initial datasets from both the expert-driven and data-driven methods are identical, while the subsequent addition of data differs. This approach ensures fairness in the comparison. The iterative model accuracy of the expert-driven and data-driven methods is shown in Note S8 in the Supplementary Information. Fig. 6(a) shows the number of superior cases for varying sizes of additional data. When the added data size is 40, the data-driven method demonstrates superiority in 50 of the 100 pairs of cases, indicating a modest impact. At this stage, the prediction models from both methods exhibit similar accuracy. However, as the amount of added data increases, the accuracy of the prediction model using the data-driven method improves. Specifically, when the added data size is 100, the data-driven method shows superiority in 70 of the 100 pairs, which is 2.3 times higher than the traditional expert-driven method. Nevertheless, with continued data addition, the superiority of the data-driven method diminishes. By the time 300 data are added, the data-driven method outperforms the expert-driven method in only 64 of the 100 pairs. This trend is visually represented in Fig. 6(b), where the red circular arcs denote the superiority percentage of the data-driven method across all cases. A larger circular arc angle signifies a higher percentage of the data-driven method's superiority. With the data-driven method, as the amount of added data increases, the percentage initially rises before declining.

Initially, when the dataset is small, the internal prediction model in

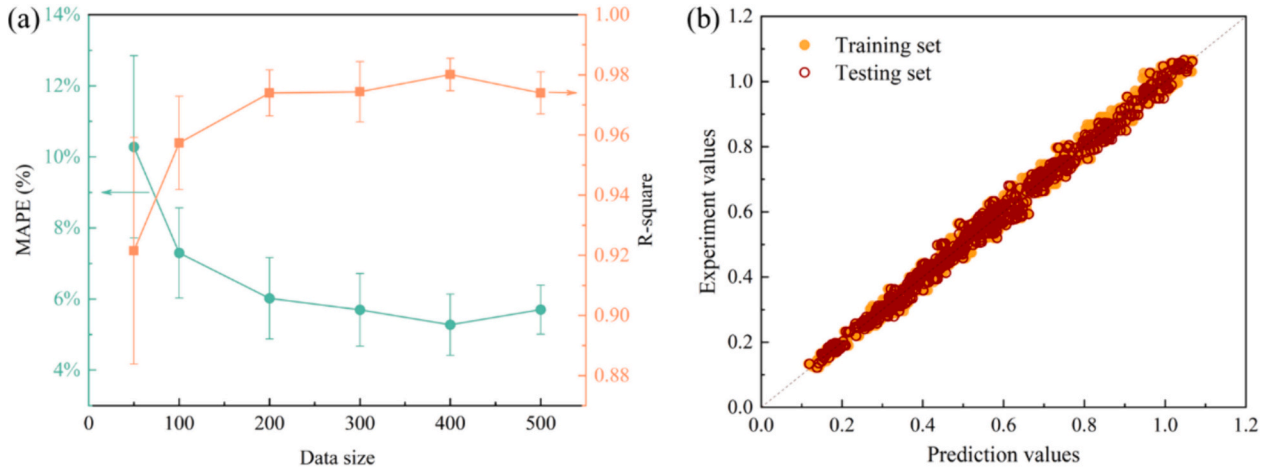


Fig. 4. (a) The accuracy of the BPNN model using a big dataset in different data sizes (b) The scattered diagram of the optimum BPNN prediction model trained by all data.

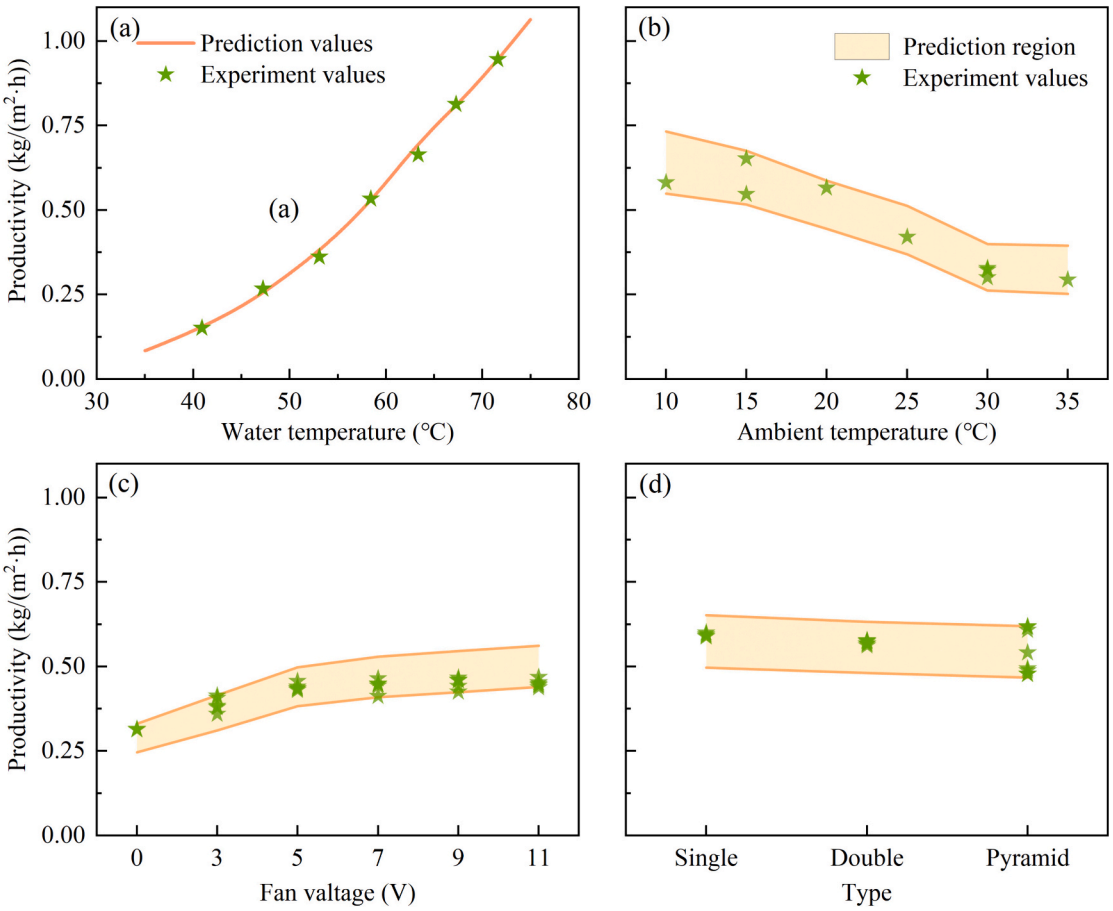


Fig. 5. The visual comparison between BPNN prediction values and experiment values. (a) different water temperatures, (b) different voltage of fan, (c) different ambient temperature, (d) different solar still types.

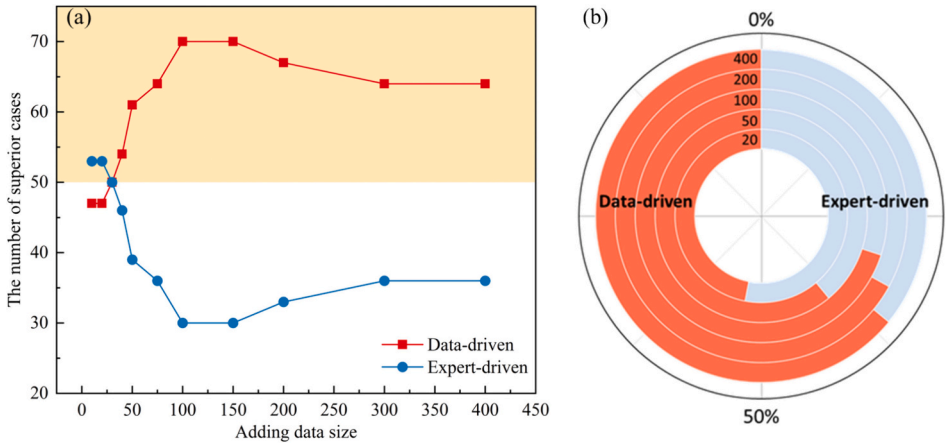


Fig. 6. The performance comparison between expert-driven and data-driven methods. In all 100 pairs of cases, each pair includes one expert-driven method and one data-driven method and is individually compared. In (a), the number of superior cases inferred by MAPE is respectively counted. (b) shows the tendency of data-driven superiority percentage in different adding data sizes. A larger circular arc angle signifies a higher percentage of the data-driven method's superiority.

the data-driven method may not be accurate enough to predict production reliably. This suggests that the selected experiment parameters may not significantly improve the accuracy of the next prediction model. As the dataset becomes larger, the internal prediction model becomes more accurate, leading to more reliable optimization performance. Consequently, the data-driven method becomes crucial in providing the next experiment parameters. With the continued addition of data, the superiority of the data-driven method gradually diminishes. This result can be explained by the fact that when the dataset is sufficiently large, the prediction model achieves convergent accuracy. At this stage, regardless of the method used to increase the data, the accuracy of the prediction model remains largely unchanged.

To further demonstrate the superiority of the data-driven method, Fig. 7(a) compares the average number of additional data required to achieve targeted MAPE. For instance, when the targeted MAPE is 6 %, the data-driven method requires an average of 69 additional data in 100

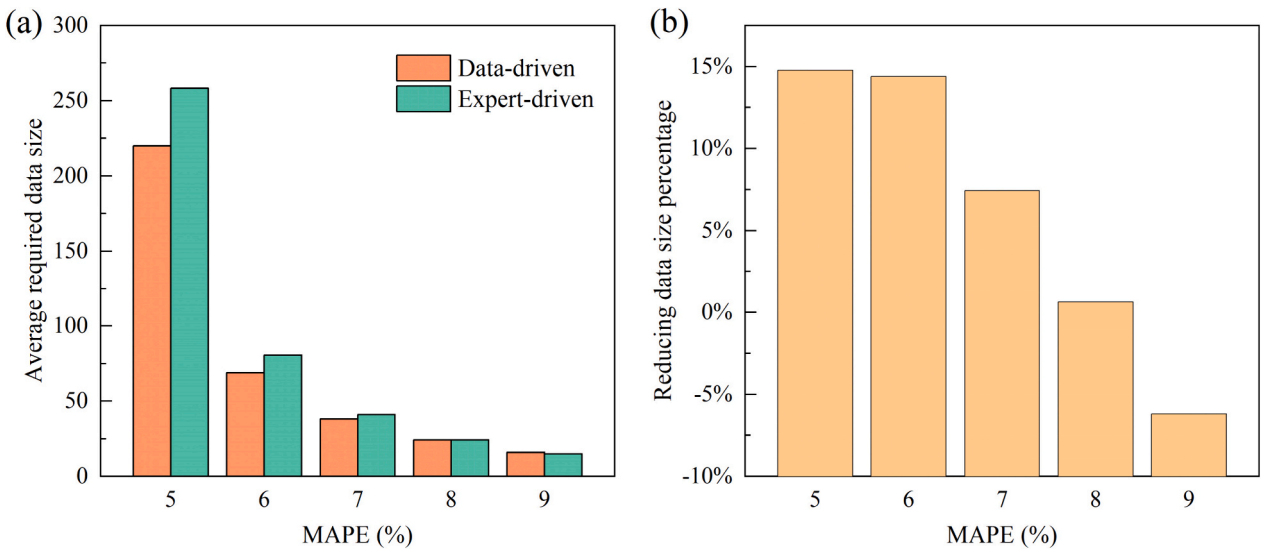


Fig. 7. The superiority of the data-driven method to the expert-driven method. (a) Compare the average number of additional data required to achieve targeted MAPE and (b) percentage reduction in data size.

pairs of cases, while the expert-driven method requires an average of 81 additional data. However, the superiority of the data-driven method becomes less apparent when the targeted MAPE is relatively high, such as 9 % or 8 %. Nonetheless, for more stringent accuracy requirements, such as a MAPE of 5 %, the data-driven method only requires an average of 220 additional data, compared to the expert-driven method, which requires an average of 258 additional data. Fig. 7(b) illustrates the percentage reduction in data size achieved by the data-driven method compared to the expert-driven method, showing a maximum reduction of 14.7 %. Specifically, when aiming for a highly accurate prediction model with a MAPE of 5 %, the data-driven method consistently outperforms the expert-driven method across various evaluation metrics. This highlights the effectiveness of the proposed data-driven method for constructing high-accuracy models.

4. Conclusion

This work proposes a data-driven method for data acquisition in the interdisciplinary study between solar stills and machine learning. The method overcomes the limitation of the traditional expert-driven method which relies on expert experiences. By employing the Bayesian optimization algorithm, the data-driven method effectively integrates data acquisition with model construction, guiding the selection of optimal experimental parameters. This accelerates the convergence of the BPNN prediction model, achieving higher accuracy.

The steps to obtain a prediction model are as follows. A solar still experimental setup is used to gather 813 data, comprising four input parameters and one productivity value. These data are then utilized to train and analyze a BPNN prediction model. After gathering approximately 300 data, the prediction model achieves convergent accuracy. Subsequently, the highly accurate BPNN prediction model, trained using the full dataset, is employed as a simulation experiment system to compare different data acquisition methods.

Then, it is demonstrated that the superiority of the data-driven method over the expert-driven method by a comparison of 100 pairs of results. When a small amount of data (~40 data) is added to the training set, both data acquisition methods show similar effects. However, with the addition of more data (150 or 200 data), models utilizing the data-driven method exhibit greater accuracy in up to 70 % of all comparisons. At this stage, the models have already achieved convergent accuracy. Further data addition beyond this point has minimal effect on improving accuracy, resulting in a decreasing superiority rate for the data-driven method.

Furthermore, it is recorded and compared that the dataset size required to meet the targeted mean absolute percentage error. When targeting a relatively large MAPE (8 %), the advantage of the proposed method is less pronounced than the expert-driven, which reduces the data size requirement by 0.7 %. However, aiming for a stricter MAPE of 5 %, models utilizing the data-driven method require an average of 220 data, compared to 258 data for models using the expert-driven method. It demonstrates a maximum reduction of 14.7 % in data requirements, highlighting its effectiveness in constructing highly accurate models.

In summary, the proposed data-driven method for data acquisition represents a significant integration of machine learning and data acquisition processes. It introduces novel ways that promise to advance research in interdisciplinary between solar desalination and machine learning.

CRediT authorship contribution statement

Senshan Sun: Investigation, Methodology, Formal analysis, Software, Writing – original draft. **Juxin Du:** Data curation, Software. **Guilong Peng:** Methodology, Formal analysis, Writing – review & editing. **Nuo Yang:** Conceptualization, Supervision, Writing – review & editing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used the ERNIE Bot to polish the schematic drawing Fig. 1. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

There are no conflicts of interest to declare.

Data availability

Data will be made available on request.

Acknowledgment

The work was sponsored by the National Key Research and Development Program of China (2018YFE0127800). The work was carried out at the National Supercomputing Center in Tianjin (NSCC-TJ), and the calculations were performed on TianHe-HPC.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.desal.2024.117946.

References

[1] U. Nations, The United Nations World Water Development Report 2023: Partnerships and Cooperation for Water, UNESCO, Paris, 2023.
[2] S. Lattemann, T. Höpner, Environmental impact and impact assessment of seawater desalination, DESALINATION 220 (2008) 1–15.
[3] H.A. Maddah, M. Bassyouni, M.H. Abdel-Aziz, M.S. Zoromba, A.F. Al-Hossainy, Performance estimation of a mini-passive solar still via machine learning, RENEW. ENERG. 162 (2020) 489–503.
[4] A.G.M. Ibrahim, I. Dincer, A solar desalination system: Exergetic performance assessment, ENERG. CONVERS. MANAGE. 101 (2015) 379–392.
[5] S.W. Sharshir, A.W. Kandeal, M. Ismail, G.B. Abdelaziz, A.E. Kabeel, N. Yang, Augmentation of a pyramid solar still performance using evacuated tubes and nanofluid: experimental approach, Appl. Therm. Eng. 160 (2019) 113997.
[6] F.A. Essa, Z.M. Omara, A.S. Abdullah, S. Shanmugan, H. Panchal, A.E. Kabeel, R. Sathyamurthy, W.H. Alawee, A.M. Manokar, A.H. Elsheikh, Wall-suspended trays inside stepped distiller with Al2O3/paraffin wax mixture and vapor suction: experimental implementation, J. ENERGY STORAGE 32 (2020) 102008.
[7] G. Peng, S.W. Sharshir, Progress and performance of multi-stage solar still – a review, DESALINATION 565 (2023) 116829.
[8] M. Al-harahsheh, M. Abu-Arabi, H. Mousa, Z. Alzghoul, Solar desalination using solar still enhanced by external solar collector and PCM, Appl. Therm. Eng. 128 (2018) 1030–1040.
[9] G. Peng, Z. Xu, J. Ji, S. Sun, N. Yang, A study on the upper limit efficiency of solar still by optimizing the mass transfer, Appl. Therm. Eng. 213 (2022) 118664.
[10] S.A. Mohamed, H. Hassan, Investigation the performance of new designed solar still of rhombus shaped based on new model, Sol. Energy 231 (2022) 88–103.
[11] Y. Liu, Z. Yang, Z. Yu, Z. Liu, D. Liu, H. Lin, M. Li, S. Ma, M. Avdeev, S. Shi, Generative artificial intelligence and its applications in materials science: current situation and future perspectives, J. MATERIOMICS 9 (2023) 798–816.
[12] Y. Liu, S. Wang, Z. Yang, M. Avdeev, S. Shi, Auto-MatRegressor: liberating machine learning alchemists, Sci. Bull. 68 (2023) 1259–1270.
[13] Y. Wang, A.W. Kandeal, A. Swidan, S.W. Sharshir, G.B. Abdelaziz, M.A. Halim, A. E. Kabeel, N. Yang, Prediction of tubular solar still performance by machine learning integrated with Bayesian optimization algorithm, Appl. Therm. Eng. 184 (2021) 116233.
[14] W. Gao, L. Shen, S. Sun, G. Peng, Z. Shen, Y. Wang, A.W. Kandeal, Z. Luo, A. E. Kabeel, J. Zhang, H. Bao, N. Yang, Forecasting solar still performance from conventional weather data variation by machine learning method, CHINESE PHYS B 32 (2023) 35–41.
[15] A.H. Elsheikh, V.P. Katekar, O.L. Muskens, S.S. Deshmukh, M.A. Elaziz, S. M. Dabour, Utilization of LSTM neural network for water production forecasting of a stepped solar still with a corrugated absorber plate, PROCESS SAF. ENVIRON. 148 (2021) 273–282.
[16] Q. He, H. Zheng, X. Ma, L. Wang, H. Kong, Z. Zhu, Artificial intelligence application in a renewable energy-driven desalination system: a critical review, Energy and AI 7 (2022) 100123.
[17] Y. Liu, S. Ma, Z. Yang, X. Zou, S. Shi, A data quality and quantity governance for machine learning in materials science, J. Chin. Ceram. Soc. 51 (2023) 427–437.
[18] H. C., J. C., J. D., Data evaluation and enhancement for quality improvement of machine learning, IEEE T. RELIAB. 70 (2021) 831–847.
[19] Y. Liu, X. Zou, Z. Yang, S. Shi, Machine learning embedded with materials domain knowledge, Journal of the Chinese Ceramic Society 50 (2022) 863–876.
[20] Y. Liu, Z. Yang, X. Zou, S. Ma, D. Liu, M. Avdeev, S. Shi, Data quantity governance for machine learning in materials science, NATL SCI REV, 10 (2023) nwad125.
[21] K. Li, D. Persaud, K. Choudhary, B. DeCost, M. Greenwood, J. Hattrick-Simpers, Exploiting redundancy in large materials datasets for efficient machine learning with less data, Nat. Commun. 14 (2023) 7283.
[22] G. Peng, S. Sun, Y. Qin, Z. Xu, J. Du, S.W. Sharshir, A.W. Kandel, A.E. Kabeel, N. Yang, Optimized data collection and analysis process for studying solar-thermal desalination by machine learning, ArXiv abs/2307.12594 (2023).
[23] Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang, W. E, DP-GEN: a concurrent learning platform for the generation of reliable deep learning based potential energy models, COMPUT, PHYS. COMMUN. 253 (2020) 107206.
[24] P. Rajak, A. Krishnamoorthy, A. Mishra, R. Kalia, A. Nakano, P. Vashishta, A.I.U.S. Argonne National Lab. ANL, Autonomous reinforcement learning agent for chemical vapor deposition synthesis of quantum materials, npj Comput. Mater. 7 (2021) 1–9.
[25] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J. MATERIOMICS 3 (2017) 159–177.
[26] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, NATURE 559 (2018) 547–555.
[27] A.S. Abdullah, A. Joseph, A.W. Kandeal, W.H. Alawee, G. Peng, A.K. Thakur, S. W. Sharshir, Application of machine learning modeling in prediction of solar still performance: a comprehensive survey, RESULTS ENG 21 (2024) 101800.
[28] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, NATURE 521 (2015) 452–459.
[29] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. de Freitas, Taking the human out of the loop: a review of Bayesian optimization, P. IEEE 104 (2016) 148–175.
[30] T.L. Lai, H. Robbins, Asymptotically efficient adaptive allocation rules, Adv. Appl. Math. 6 (1985) 4–22.
[31] F. Nogueira, Bayesian Optimization: Open Source Constrained Global Optimization Tool for Python, 2014.

Supplementary Information

A data-driven method to construct prediction model of solar stills

Senshan Sun[1, 2], Juxin Du[1], Guilong Peng[3, †], Nuo Yang[4, †]

[1]School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[2]School of Integrated Circuits, Wuhan National Laboratory for Optoelectronics (WNLO), Key Laboratory of Material Chemistry for Energy Conversion and Storage, Huazhong University of Science and Technology, Wuhan 430074, China

[3]School of Mechanical and Energy Engineering, Shaoyang University, Shaoyang 422000, China

[4]Department of Physics, College of Science, National University of Defense Technology, Changsha 410073, China

†Corresponding E-mail: G.P. (4195@hnsyu.edu.cn) and N.Y. (nuo@nudt.edu.cn)

Note S1 Figure source in Fig. 1

Table S1 Figure source in Fig. 1

Section	Drawing method	Reference
Method	AI draw	ERNIE Bot
Selection Inspiration	Website	• Noisechannelrocker. Brain - Human Brain Human Head Clip Art PNG, https://favpng.com/png_view/brain-human-brain-human-head-clip-art-png/qfagg9AU • RESEARCH PAPER, https://flectone.ru/research-paper.html • Solar Still, https://www.force4.co.uk/item/Aquamate/Solar-Still/DXG • StudenT Computer & Educational Services, https://www.facebook.com/profile.php?id=100048610826198
Termination Conditions	AI draw	ERNIE Bot

Note S2 Experiment system and data

Table S2 Specifics of devices and sensors in the experiments.

Name	Brand	Type	Range	Error
Fan	LFFAN	LFS0512SL	0 ~ 4800 RPM	-
Electronic balance	ANHENG	AH-A503	0 ~ 500 g	±0.01 g
Power #1&#3	WANPTEK	NPS3010W	0 ~ 30 V	±0.1 %
Power #2	ITECH	IT6932A	0 ~ 60 V	± 0.03 %
Data acquisition unit	CAMPBELL SCIENTIFIC	CR1000X& AM25T	25 Channels	-
Thermostat water bath	QIWEI	DHC-2005-A	-20 ~ 99.9 °C	± 0.2 °C
Heating panel	BEISITE	Custom-made	0 ~ 2000 W/m^2	-
Thermal couple	ETA	T-K-36-SLE	-200 ~ 260 °C	± 1.1 °C

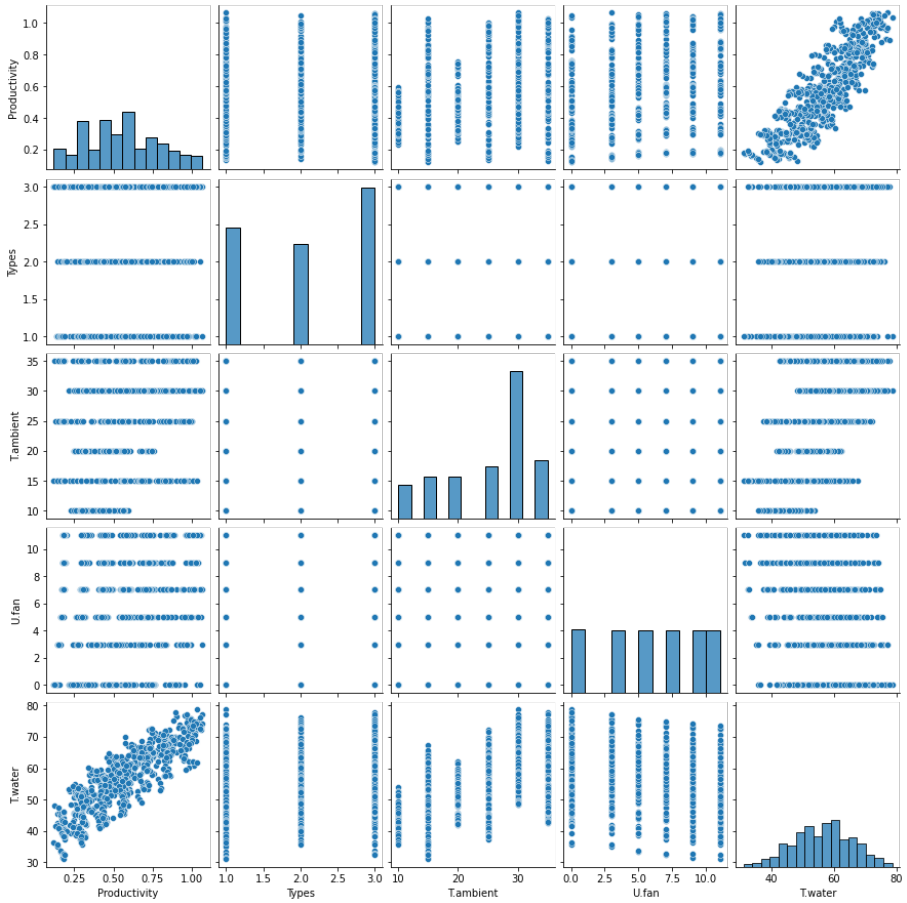


Fig. S1 The pairs plot of collected data

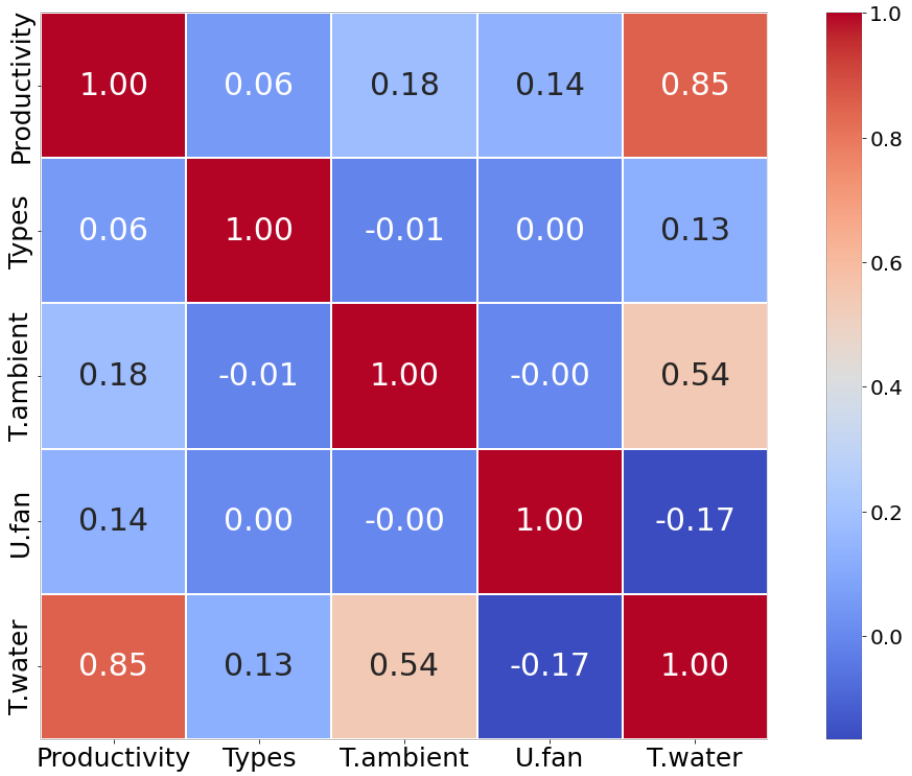


Fig. S2 The correlation heat map of collected data

Note S3 The optimization process

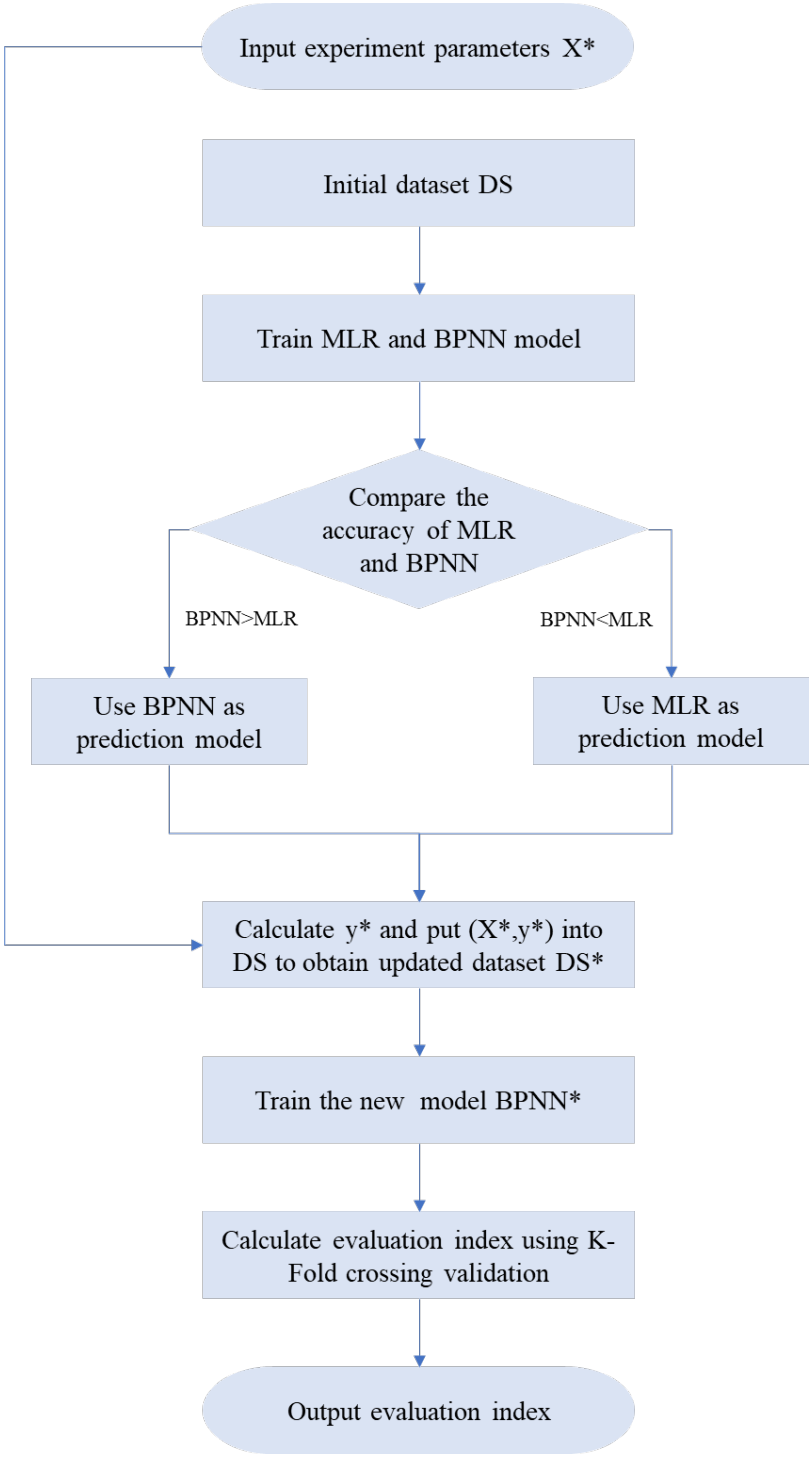


Fig. S3 The objective function construction

In the objective function, the input factors consist of the next experiment parameters Ps^*, and the output value is the R-square calculated by the BPNN model trained with the updated dataset. The updated dataset includes Ps^* and its corresponding production. Initially, a prediction model is acquired by training the

original dataset obtained from the preceding step. In the study by Peng and Sun[1], it was observed that when the dataset is small, the Multiple Linear Regression (MLR) model remains stable. Therefore, two prediction models, BPNN and MLR, are established and compared. The more accurate prediction model is selected as the Inner Prediction Model (IPM) in the objective function. With the IPM, the production corresponding to Ps^* can be computed. Subsequently, with the addition of new data, an updated dataset DS^* is formulated. A new BPNN model can be trained using DS^*, and the evaluation index R-square can be determined. Finally, the R-square value is outputted.

Note S4 Expert-driven and data-driven data acquisition procedures

In this work, four input parameters (x_1, x_2, x_3, x_4) are used to predict freshwater production.

1. Expert-driven data acquisition method

In the first experiment with the expert-driven method, one parameter x_{vp} is randomly selected as the variable parameter, while the other three parameters are constant parameters $(x_{cp1}, x_{cp2}, x_{cp3})$. The variable and constant parameters are randomly generated by a computer as x_{vp}^1 and $(x_{cp1}^0, x_{cp2}^0, x_{cp3}^0)$, respectively. The first set of input parameters $(x_{vp}^1, x_{cp1}^0, x_{cp2}^0, x_{cp3}^0)$ is used to obtain the freshwater production P^1 through experiment. In the second experiment, the computer only generates the variable parameter as x_{vp}^2 which is different from the existing x_{vp}, keeping the constant parameters unchanged, and obtaining freshwater production P^2 through experiment. In the i_{th} experiment, the data $(P^i, x_{vp}^i, x_{cp1}^0, x_{cp2}^0, x_{cp3}^0)$ is recorded. When experts believe that the variable parameter needs alteration, this cycle completes, and the next cycle starts. This cycle repeats until the expert deems the dataset sufficient, completing the data acquisition process.

2. Data-driven data acquisition method

After obtaining an initial dataset with i sets of data, it is used to construct the prediction model with an initial model estimating index EI^0. In the $(i+1)_{th}$ experiment, four input parameters $(x_1^{i+1}, x_2^{i+1}, x_3^{i+1}, x_4^{i+1})$ are recommended based on the proposed data-driven method, and the freshwater production P^{i+1} is obtained. In the $(i+2)_{th}$ experiment, the dataset is updated with the new data $(P^{i+1}, x_1^{i+1}, x_2^{i+1}, x_3^{i+1}, x_4^{i+1})$, the prediction model is rebuilt, and a new estimating index EI^1 is calculated. The next set of input parameters $(x_1^{i+2}, x_2^{i+2}, x_3^{i+2}, x_4^{i+2})$ is then recommended, and the corresponding freshwater production P^{i+2} is obtained. This process continues iteratively. When the experiment reaches the $(i+j)_{th}$ iteration and the estimating index EI^j exceeds the target threshold ($EI^j > EI_{threshold}$), the data acquisition process is automatically stopped.

Note S5 Comparison between experimental and simulation data acquisition methods

Table S3 The difference in verification method between the prediction model and experiment

Specifications	Simulation	Experiment
The necessary amount of experimental data	0	Unexpected
Acquisition time per data	A few seconds (sec)	A few minutes (min)
Production Accuracy	±3.4%	0
Parameter Accuracy	0	Unknown
Data-driven Operation	Automatic	Manual

Note S6 The independence validation of the BPNN model

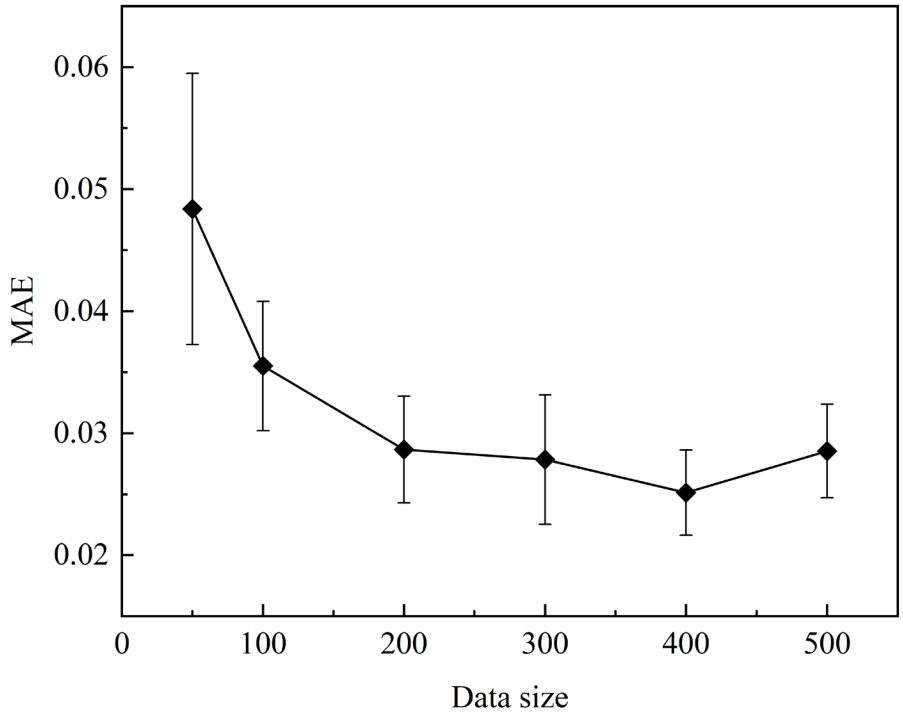


Figure S4 The evaluating index MAE of the BPNN model in different data size

Note S7 Comparing different program packages in Python

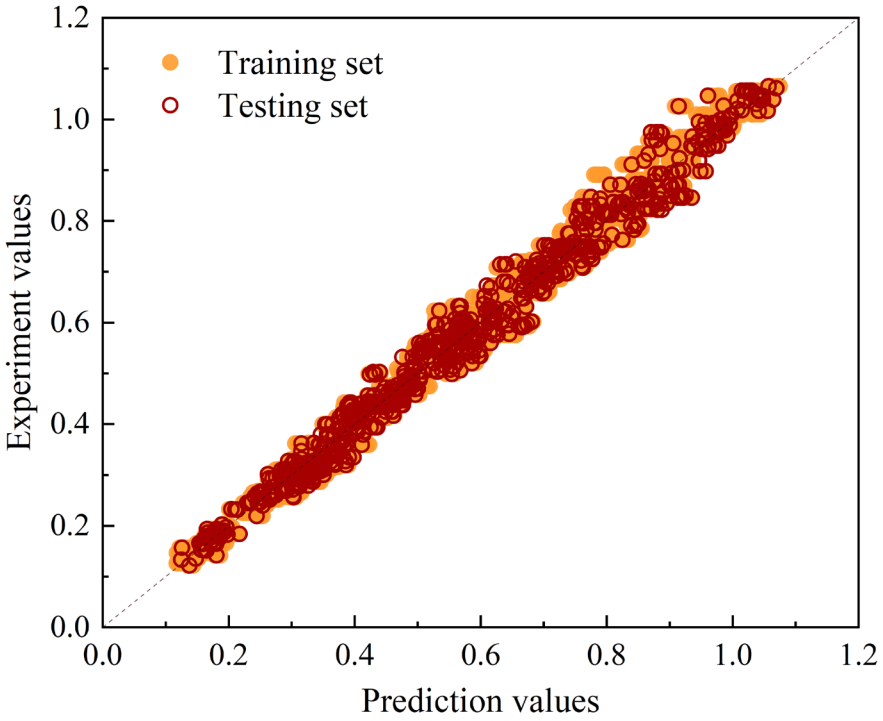


Fig. S5 The scattered diagram of the BPNN model using all data with the “Sklearn” package.

Figure S5 illustrates the scattered point distribution for the BPNN prediction model programmed using the "Sklearn" package, with the test set R^2 and MAPE being 0.98 and 4.8%, respectively. Although both Python packages for machine learning yield accurate results in the scatter diagram, the "Pytorch" package notably provides a better fit for the prediction model. This can be attributed to the "Pytorch" package's proficiency in the field of machine learning model fitting and its inclusion of more adjustable hyperparameters. On the other hand, the "Sklearn" package serves as a comprehensive tool for data fitting and processing, making it easier to program for model fitting compared to the "Pytorch" package. Therefore, the BPNN model programmed using the "Sklearn" package is selected for this study.

Note S8 Iterative model accuracy

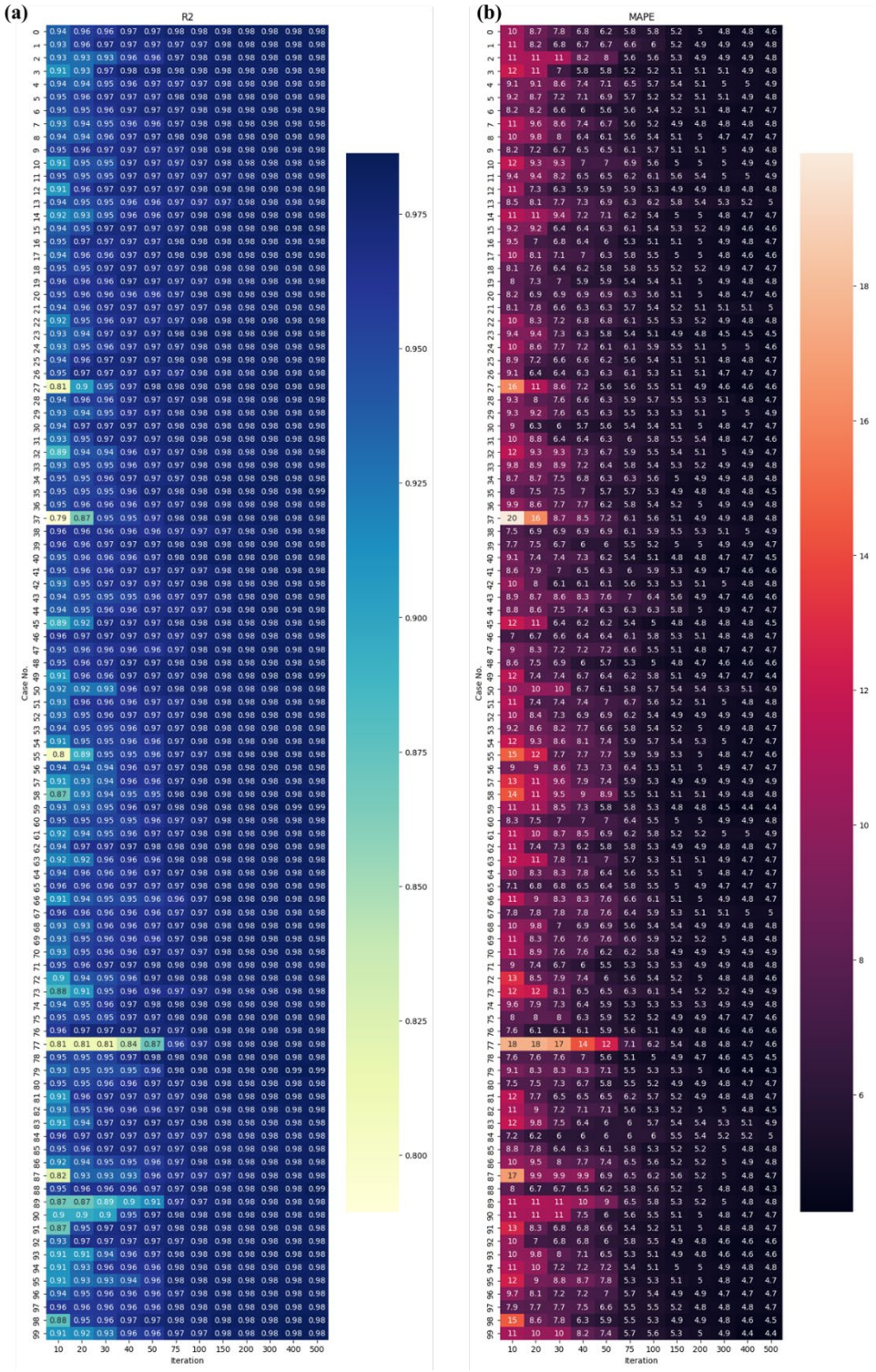


Fig. S6 The heatmap of iterative model accuracy with proposed data-driven data acquisition method: (a)R-squire, (b) Mean absolute percentage error

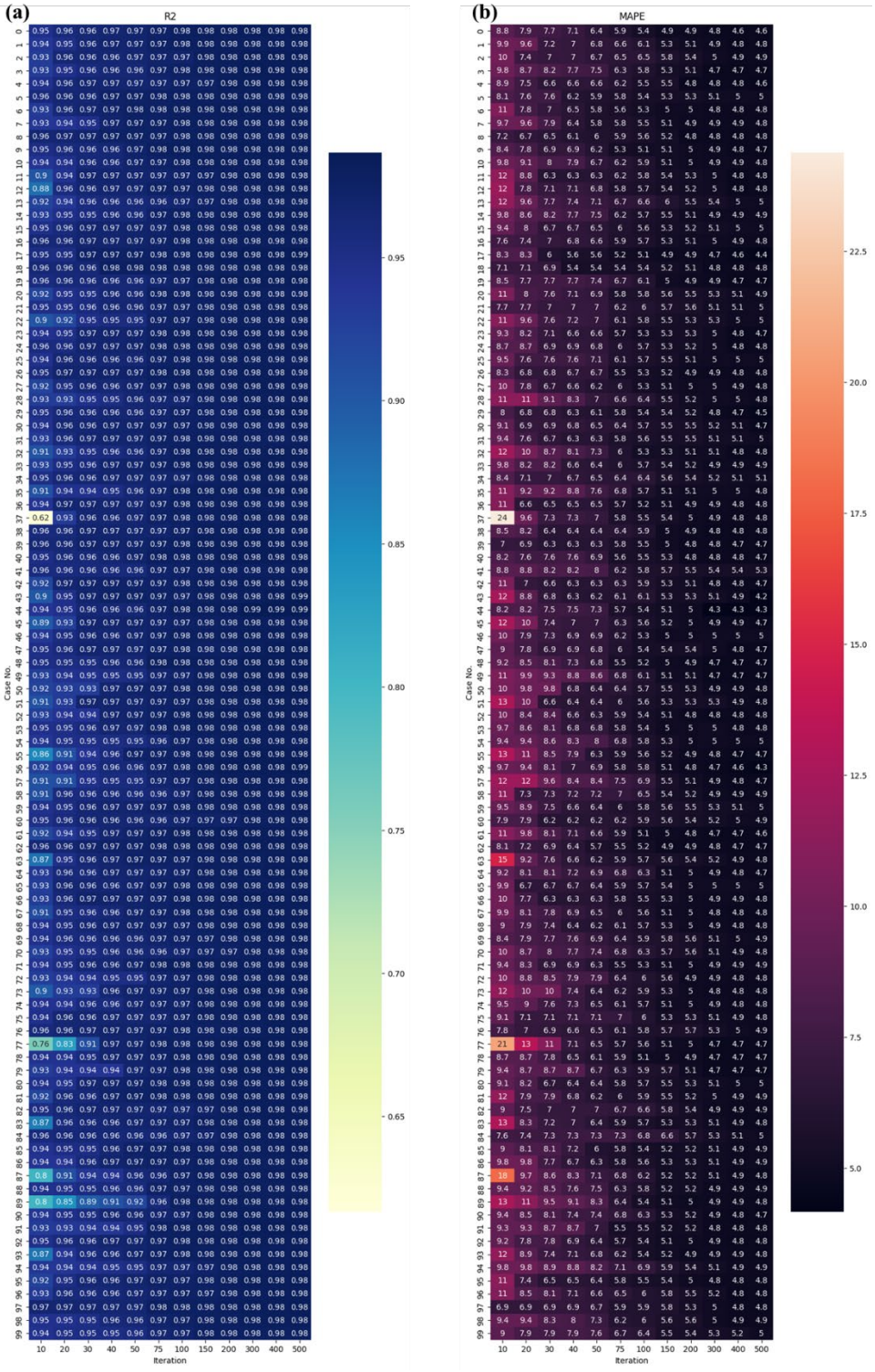


Fig. S7 The heatmap of iterative model accuracy with traditional expert-driven data acquisition method: (a)R-squire, (b) Mean absolute percentage error

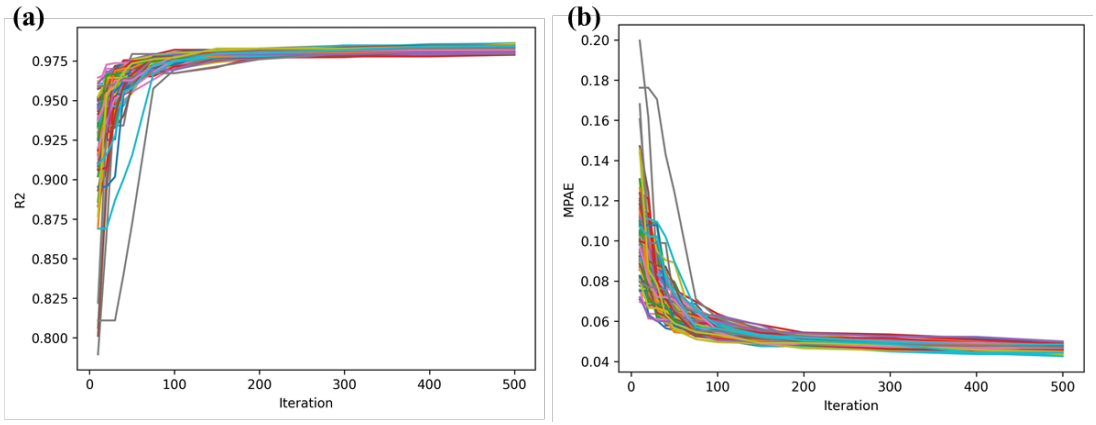


Fig. S8 The change line of iterative model accuracy with the proposed data-driven data acquisition method: (a) R-squire, (b) Mean absolute percentage error

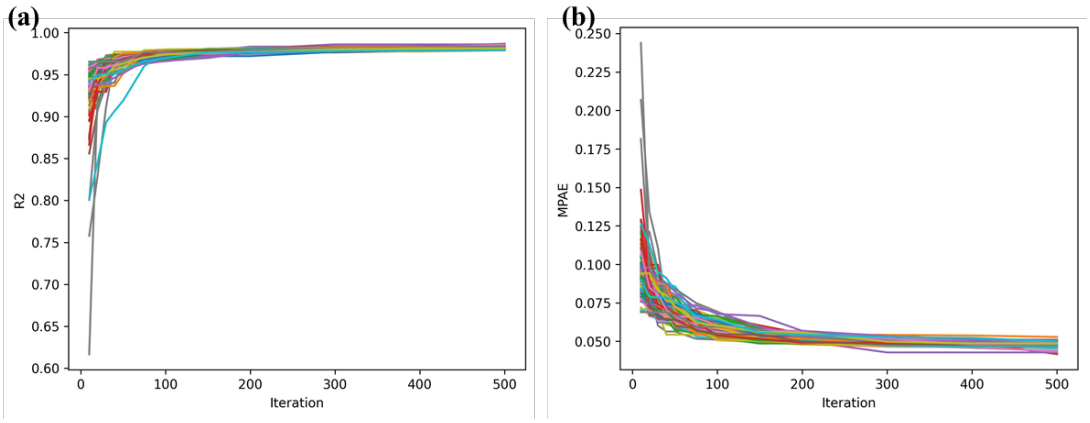


Fig. S9 The change line of iterative model accuracy with traditional expert-driven data acquisition method: (a)R-squire, (b) Mean absolute percentage error

Reference

[1] G. Peng, S. Sun, Y. Qin, Z. Xu, J. Du, S. W. Sharshir, et al. Optimized data collection and analysis process for studying solar-thermal desalination by machine learning. Arxiv, 2023, abs/2307.12594