9-1-2023

Harnessing the power of neural networks for the investigation of solar-driven membrane distillation systems under the dynamic operation mode

Pooria Behnam
Edith Cowan University

Masoumeh Zargar
Edith Cowan University

Abdellah Shafieian
Edith Cowan University

Amir Razmjou
Edith Cowan University

Mehdi Khiadani
Edith Cowan University

10.1016/j.solener.2023.06.007
Behnam, P., Zargar, M., Shafieian, A., Razmjou, A., & Khiadani, M. (2023). Harnessing the power of neural networks for the investigation of solar-driven membrane distillation systems under the dynamic operation mode. Solar Energy, 261, 63-82. https://doi.org/10.1016/j.solener.2023.06.007

Solar Energy 261 (2023) 63–82

Contents lists available at ScienceDirect

Solar Energy

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

Harnessing the power of neural networks for the investigation of solar-driven membrane distillation systems under the dynamic operation mode

Pooria Behnam [a], Masoumeh Zargar [a], Abdellah Shafieian [a], Amir Razmjou [a,b,c], Mehdi Khiadani [a,*]

[a] School of Engineering, Edith Cowan University, 270 Joondalup Drive, Joondalup, Perth, WA 6027, Australia
[b] UNESCO Centre for Membrane Science and Technology, School of Chemical Engineering, University of New South Wales, Sydney, NSW 2052, Australia
[c] Mineral Recovery Research Center (MRRC), School of Engineering, Edith Cowan University, Joondalup, Perth, WA, 6027, Australia

ARTICLE INFO

Keywords:
Direct contact membrane distillation
Solar desalination
Dynamic modeling
Machine learning
Neural networks

ABSTRACT

Accurate modeling of solar-driven direct contact membrane distillation systems (DCMD) can enhance the commercialization of these promising systems. However, the existing dynamic mathematical models for predicting the performance of these systems are complex and computationally expensive. This is due to the intermittent nature of solar energy and complex heat/mass transfer of different components of solar-driven DCMD systems (solar collectors, MD modules and storage tanks). This study applies a machine learning-based approach to model the dynamic nature of a solar-driven DCMD system for the first time. A small-scale rig was designed and fabricated to experimentally assess the performance of the system over 20 days. The predictive capabilities of two neural network models: multilayer perceptron (MLP) and long short-term memory (LSTM) were then comprehensively examined to predict the permeate flux, efficiency and gain-output-ratio (GOR). The results showed that both models can efficiently predict the dynamic performance of solar-driven DCMD systems, where MLP outperformed the LSTM model overall, especially in the prediction of efficiency. Additionally, it was indicated that the accuracy of the models for the prediction of GOR can be significantly improved by increasing the size of the dataset.

1. Introduction

Access to safe clean water has emerged as a serious challenge during the last decade as a result of industrialization, climate change and rapid population growth [1]. On the one hand, solar desalination technologies can play a key role in addressing water scarcity, as 97% of water sources are saline or brackish in nature. On the other hand, water-stressed regions often benefit from great solar irradiance, which can be used to power desalination systems [2]. Among differing small-scale solar-powered desalination technologies, the membrane distillation (MD) process has received significant attention over the past few years [3]. This mainly stems from the great advantages of the MD process such as, its low operational temperature, compact structure and high efficiency for highly saline water desalination [4–7]. The MD approach involves a combined thermal/membrane desalination process in which the vapor pressure difference between the feed and permeate sides results in evaporation within the feed flow. Following this, the generated vapor transfers through a thin hydrophobic membrane, and condensation occurs at the permeate side. In accordance with the condensation method on the permeate side, the MD process can be mainly classified into four general categories. These categories are namely: direct contact membrane distillation (DCMD), air–gap membrane distillation (AGMD), vacuum membrane distillation (VMD), and sweep gas membrane distillation (SGMD) [8,9]. Heretofore, the performance of solar-driven MD systems comprising varied solar systems and MD modules has been widely investigated using experimental and theoretical methods.

Many researchers have sought to experimentally examine the performance of solar-driven MD systems mainly due to the complexity involved in modelling these systems. This has mostly been due to the intermittent nature of solar energy and complex heat/mass transfer in different components of solar-driven MD systems, such as solar collectors, MD modules and storage tanks. The integration of a heat pipe evacuated tube solar collector with the DCMD module has previously

---

[*] Corresponding author.
E-mail address: m.khiadani@ecu.edu.au (M. Khiadani).

https://doi.org/10.1016/j.solener.2023.06.007
Received 5 March 2023; Received in revised form 1 June 2023; Accepted 5 June 2023

Nomenclature

a	Activation function
A	Area (m^2)
AGMD	Air-gap membrane distillation
b	Bias term
C	Specific heat of feed water (J/kg.C), Memory cell
$\acute{C}$	Potential cell sate
DCMD	Direct contact membrane distillation
FCCD	Face-centered composite design
GOR	Gain-output-ratio
h	Evaporation enthalpy (J/kg)
I	Solar intensity (W/m^2)
IQR	Interquartile range
K	Number of folds
LSTM	Long short-term memory
$\dot{m}$	Mass flow rate (kg/s)
M	Mass flow rate (kg/h)
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MD	Membrane distillation
MLP	Multilayer perceptron
N	Number of samples
N_f	Number of features
Q	Mass flow rate (L/min), Quartile
R^2	Coefficient of determination
RMSE	Root mean square error
SGMD	Sweep gas membrane distillation
t	Timestep
t_w	Number of timesteps
T	Temperature (°C), T'th input
T0	Feed inlet temperature (°C)
T1	Feed outlet temperature (°C)
T2	Collector inlet temperature (°C)
T3	Collant inlet temperature (°C)
T4	Collector outlet temperature (°C)
T5	Collant outlet temperature (°C)
tanh	Hyperbolic tangent activation function
U	Weights
VMD	Vacuum membrane distillation
w	Weights
W	Power (W)
X	Input data
y_i	Experimental target value
$\hat{y}_i$	Predicted target value
$\bar{y}$	Mean value of the experimental data
1D	One-dimensional
2D	Two-dimensional

Greek symbols

η	Efficiency
Γ_f	Forget gate
Γ_i	Input gate
Γ_o	Output gate
σ	Logistic activation function

Subscripts

f	Forget
fg	Latent heat of evaporation of water
i	Input
max	Maximum data
min	Minimum data
o	Output
p	Pressure

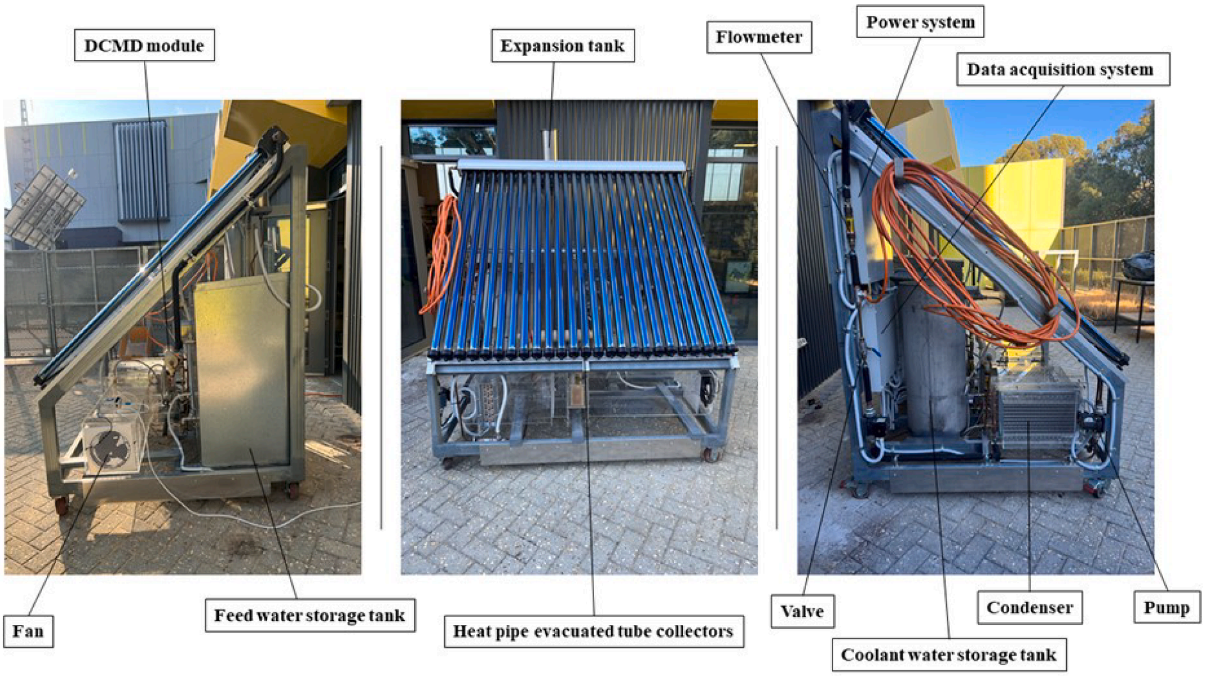


Fig. 1. The fabricated solar-driven DCMD system from different views.

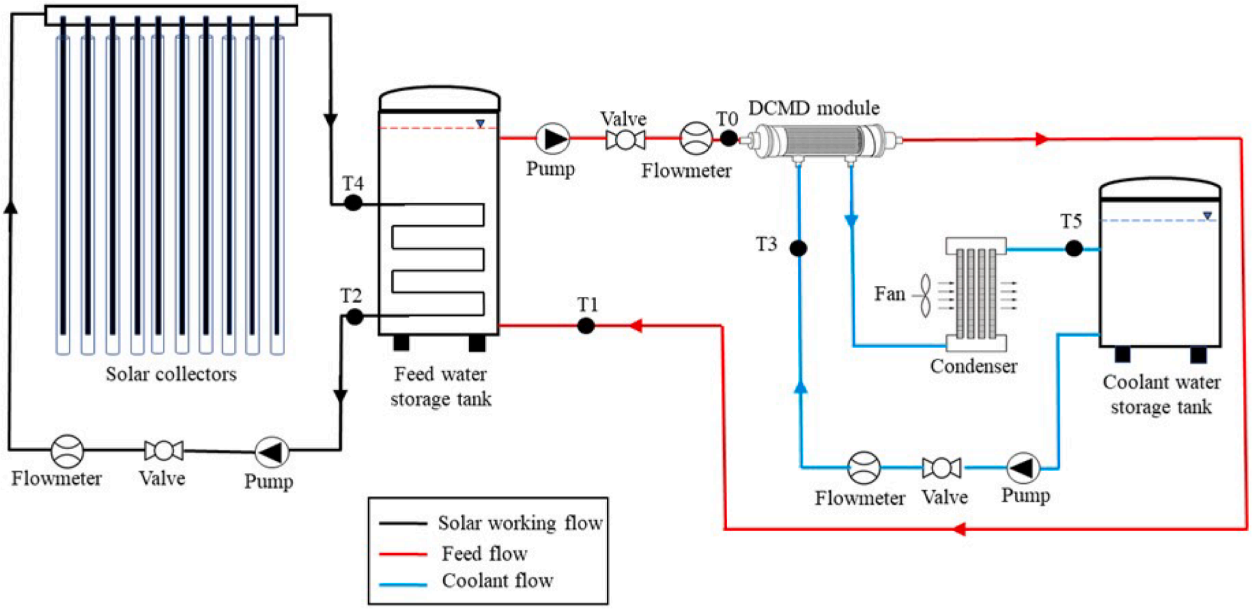


Fig. 2. Schematic illustration of the experimental setup.

Table 1
Characteristics of solar loop's main components.

Parameter		Value/Type
Solar collector	Number of evacuated tubes	25
	Tube length	1.8 m
	Inner diameter of glass tube	0.047 m
	Outer diameter of glass tube	0.058 m
	Absorptivity	0.97
	Transmittance	0.88
	Emissivity	0.07
	Heat pipe diameter	0.008 m
	Heat pipe working fluid	Ethanol
	Heat pipe material	Copper
Coil heat exchanger	Heat exchanger material	Copper
	Length	34 m
	Heat transfer area	1.45 m^2
Feed water storage tank	Feed water storage tank volume	210 L
	Insulation thickness	0.05 m

Table 2
The properties of the DCMD module.

Parameter	Value/Type
Material	PVDF
Module length	0.47 m
Effective area	0.118 m^2
Inner diameter of hollow fibers	0.8 mm
Outer diameter of hollow fibers	1.6 mm
Pore size	0.2 µm
Porosity	45%
Hydrostatic pressure	0.25 MPa

been examined by Shafieian and Khiadani [10]. They recommended modelling and optimization of the system by considering flow rate changes in solar and feed loops for future investigations. Bamasag et al. [11] have proposed a novel compact solar-driven DCMD desalination system by inserting hollow fiber membranes inside evacuated tube collectors. The performance of the proposed system was experimentally analyzed and the permeate flux generation shown to vary in the range of

Table 3
Specifications of the condenser.

Parameter	Value/Type
Length	310 mm
Height	210 mm
Width	900 mm
Fin material	Aluminum
Tube material	Copper
Tube coil diameter	9.5 mm
Header	25.4 mm

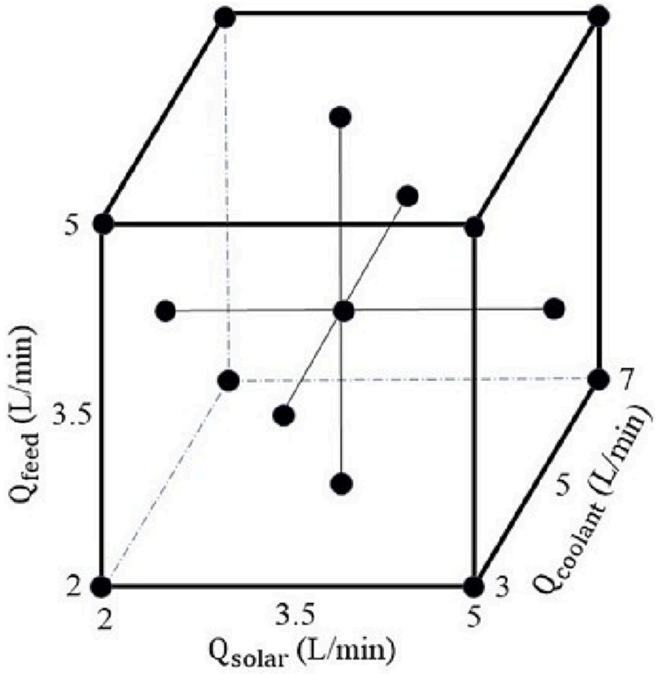


Fig. 3. The distribution of mass flow rates using face-centered composite design approach.

Table 4
Values of mass flow rates during 20 days of experiments designed by the FCCD.

Day	Point type	Q_{solar}(L/min)	Q_{feed}(L/min)	$Q_{coolant}$(L/min)
1	Center	3.5	3.5	5
2	Axial	3.5	3.5	7
3	Axial	3.5	5	5
4	Axial	3.5	3.5	3
5	Axial	2	3.5	5
6	Factorial	5	5	3
7	Factorial	5	2	7
8	Center	3.5	3.5	5
9	Factorial	2	5	3
10	Factorial	5	2	3
11	Center	3.5	3.5	5
12	Factorial	2	2	7
13	Axial	5	3.5	5
14	Center	3.5	3.5	5
15	Center	3.5	3.5	5
16	Axial	3.5	2	5
17	Factorial	2	5	7
18	Factorial	5	5	7
19	Factorial	2	2	3
20	Center	3.5	3.5	5

2.2 to 6.5 L/m^2.h. Kabeel et al. [12] have experimentally showed that using an evaporative cooler in the coolant side of a solar-driven DCMD desalination system increased permeate flux by 1.25 times. Moreover, Abdelgaied et al. [13] experimentally investigated the application of both phase change materials and evaporative cooling in a solar-driven DCMD desalination system. The results indicated an approximately 45% enhancement in the gain-output-ratio (GOR) of a solar-driven DCMD system. Hejazi et al. [14] evaluated the performance of the solar-driven DCMD system under intermittent working conditions with different washing and liquid drainage scenarios. Their results showed that daily permeate flux and energy consumption remained insignificantly affected by the different scenarios.

Accurate modelling of solar-driven MD systems can result in better understanding of the performance of these systems, and facilitate improvements in performance optimization. Despite the vital importance of modelling integrated solar-driven MD systems, limited research has thusfar addressed the dynamic characteristics of these promising systems [15,16]. In this regard, dynamic models that consider variations of operational parameters over time should be applied due to the intermittent nature of solar energy [16]. A few investigators have sought to apply one-dimensional (1D) or two-dimensional (2D) physical-based modelling approaches. These approaches consider unsteady variations of operational parameters either along or perpendicular to membrane length. Chang et al. [17] have developed a 1D approach on the Aspen platform for the dynamic modelling of a solar AGMD system. The model's accuracy in essential parameters such as permeate flux or efficiency were not fully compared to experimental results, despite comparing the effect of feed flow rate and temperature to experimental results. Duong et al. [16] have applied the 1D model for the DCMD module, using TRNSYS software to dynamically simulate the performance of an integrated solar DCMD system. These authors found that, in contrast to the stand-alone DCMD module, increasing the feed flow rate reduced the performance of direct solar-driven DCMD desalination systems. Eleiwi et al. [18] have proposed a 2D mathematical model to analyze the unsteady performance of a DCMD module. While the model accurately predicted permeate flux, it was developed solely for the DCMD module and did not consider the integration of solar energy. The performance of a solar-driven DCMD system was examined by developing a 2D dynamic mathematical model [19]. The accuracy of the model was compared with experimental permeate fluxes for 150 days. The results showed that the model failed to fully capture variations in permeate flux over the 150 days of experiments. To simplify dynamic modelling approaches, some researchers have applied a lumped-parameter method, treating the entire MD module as a single control volume [20–22]. In these types of studies the focus to date has been on dynamic modelling of the MD module, whilst modelling of the integrated solar-driven MD systems has not been considered. Further, the dynamic characteristics of essential performance indicators, such as permeate flux or efficiency of the MD modules were not fully investigated.

Data-driven methods have received noticeable attention in recent years in relation to performance analysis of different solar-driven energy systems [23–28]. In particular, artificial neural networks have exhibited great capability among data-driven methods, largely due to their great ability to capture the non-linear characteristics of these systems [29]. The application of data-driven methods for performance analysis of solar desalination systems has been the focus of several studies. The results reported by Zarei and Behyad [30] have revealed that the multi-layer perceptron neural network accurately predicted the optimum design parameters of a solar greenhouse desalination system. Further, some researchers have applied machine learning models for performance prediction of solar still desalination systems [31,32]. Their results showed that the long short-term memory (LSTM) method properly modelled the dynamic performance of solar still desalination systems. This was mainly due to the model's great capability in learning patterns. In another study, Salem et al. [33] have applied different machine learning techniques to analyze the performance of a complex solar desalination system. The system was comprised of solar still and humidification-dehumidification desalination systems. The researchers

Table 5
Summary of the experimental dataset for data collected in 20 days.

	Inputs											Outputs		
	T0 (°C)	T1 (°C)	T2 (°C)	T3 (°C)	T4 (°C)	T5 (°C)	$T_{ambient}$ (°C)	Solar intensity (W/m^2)	Q_{solar} (L/min)	Q_{feed} (L/min)	$Q_{coolant}$ (L/min)	Permeate flux (L/m^2.h)	Efficiency (%)	GOR
Number of data samples	160	160	160	160	160	160	160	160	160	160	160	160	160	160
Mean	44.1	42.1	43.4	32.8	46.5	32.1	22.4	350.7	3.5	3.5	5.0	3.5	33.2	0.58
Standard deviation	3.7	3.7	3.9	3.8	4.0	3.8	2.1	174.6	1.1	1.1	1.4	1.6	27.3	0.21
Minimum	31.4	30.0	29.1	23.9	34.7	22.2	16.9	43.9	2.0	2.0	3.0	1.1	5.4	0.13
First quartile (25%)	41.5	39.7	40.9	30.6	43.9	30.0	20.9	171.0	3.1	3.1	4.5	2.0	9.9	0.39
Second quartile (50%)	44.5	42.7	44.2	33.5	47.0	32.6	22.1	431.4	3.5	3.5	5.0	3.7	22.7	0.61
Third quartile (75%)	47.0	44.8	46.2	35.3	49.5	35.0	23.7	500.9	3.9	3.9	5.5	4.6	57.2	0.76
Maximum	50.9	49.1	50.7	40.7	54.8	39.7	27.8	589.9	5.0	5.0	7.0	7.1	99.7	0.93

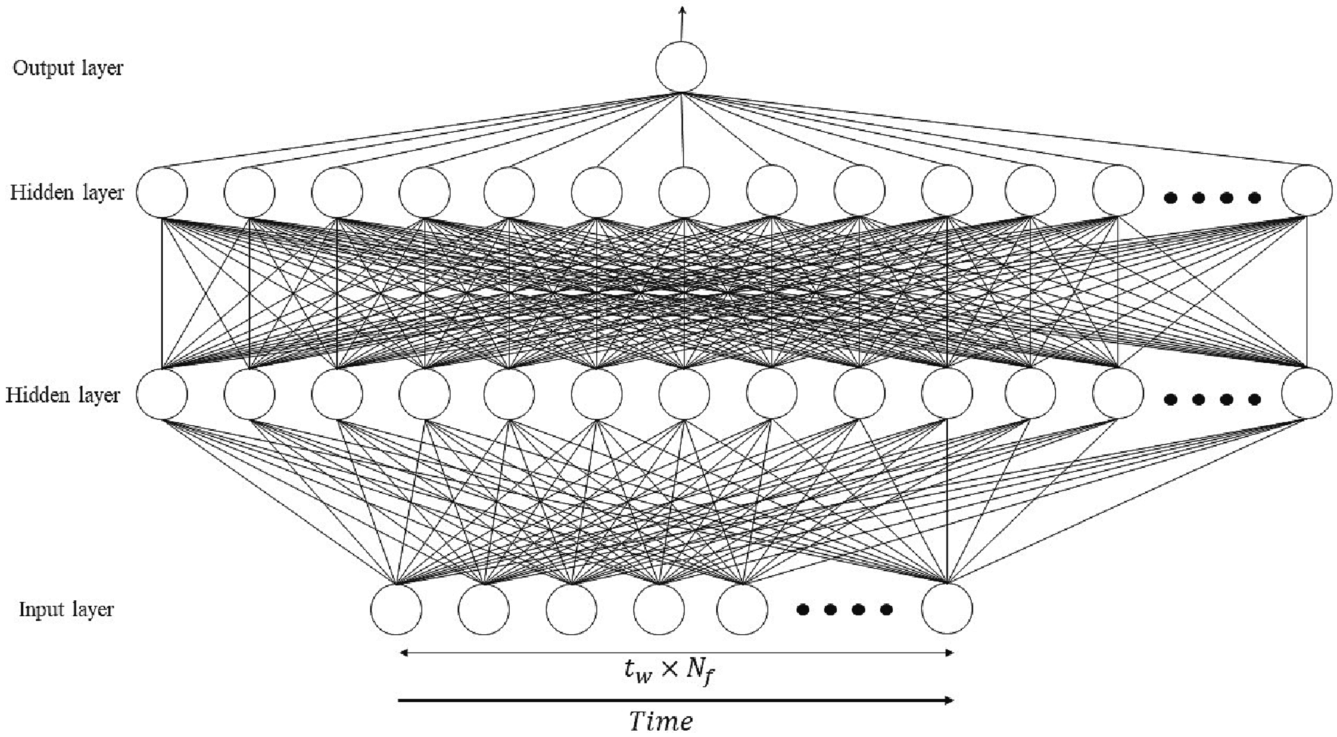


Fig. 4. Structure of MLP model.

showed that the multi-layer perceptron neural network had great performance prediction capability and a reasonable computational cost. Porrazzo et al. [34] have employed an artificial neural network to adaptively control the feed flow rate of solar-driven membrane distillation system, with results showing 17% freshwater productivity.

Previous investigations have shown that meeting two requirements is necessary to achieve precise modelling of solar-driven DCMD systems: (1) All components of the solar-driven DCMD system such as solar collectors, storage tanks and MD modules should be considered in the modelling simultaneously [16]. (2) dynamic models should be applied to properly capture the transient performance of the system due to the intermittent nature of solar energy [20,22]. When considering these two important points in modelling solar-driven DCDM systems, performance diagnoses of these systems using existing dynamic mathematical models becomes too complex and computationally expensive. This stems from solving many heat and mass transfer equations for various system components (DCMD module, solar collectors and storage tanks) and the time-varying performance of the system. A recent study by authors [35] has revealed that machine/deep learning methods, despite their effectiveness in analyzing desalination technologies, have not been sufficiently applied to performance modelling of solar-driven DCMD systems. The literature shows that few studies have applied machine learning models to characterize different MD modules [35–38]. However, these studies have not met the two above-mentioned requirements. Their focus was solely on modelling the stand-alone MD modules, and they have not applied data-driven methods to model the integrated solar-driven DCMD system. Furthermore, data-driven approaches have been used to model the performance of MD modules under steady-state conditions.

To address these shortcomings, this study aims to apply a novel methodology based on machine learning approaches for accurate modelling of an integrated solar-driven DCMD desalination system. To this end, two time-series neural network models, namely multilayer perceptron (MLP) and long short-term memory (LSTM) have been employed. To apply this methodology, an experimental set-up was designed and fabricated, whereby the performance of the system was investigated over 20 days under differing operational and weather conditions. The experimental tests were designed based on the face-centered composite design (FCCD) approach [39] in order to appropriately set different flow rates in the system. Unlike the previous studies in terms of application of data-driven methods for modelling of stand-alone MD modules, more experimental data was collected for developing the data-driven models [36,37].A dataset consisting of 160 experimental samples was then collected in which the effect of different parameters (temperatures at different points, ambient temperature, solar intensity and mass flow rates) on three performance indicators (permeate flux, efficiency and GOR) was comprehensively captured. The K-fold cross-validation method was used for accurate hyper-parameter tuning. This method provides a better estimate of model performance by using all available data for both training and validation. This can help avoid biases that might be present when using a single validation set, especially in cases where the size of the dataset is relatively small, as seen in previous studies on data-driven modelling of stand-alone MD modules [35]. The predictive performance of the developed models was then comprehensively assessed in terms of various statistical criteria. Further, the accuracy of both models was compared by selecting different sample sizes of datasets. Overall, the novelty of this study can be summarized as follows:

- Applying a novel methodology based on machine learning approaches for accurate and non-complex dynamic modeling of an integrated solar-driven DCMD desalination system for the first time.
- An FCCD-designed dataset was used to evaluate the solar-driven DCMD system over 20 days with varying operational and weather conditions. This is in contrast to previous studies that used data-driven methods for the analysis of other types of solar desalination technologies with limited experimental data [30,32,35,40].
- Applying the K-fold cross-validation method for accurate hyper-parameter tuning and proper evaluation of the developed neural networks based on the unseen test dataset.

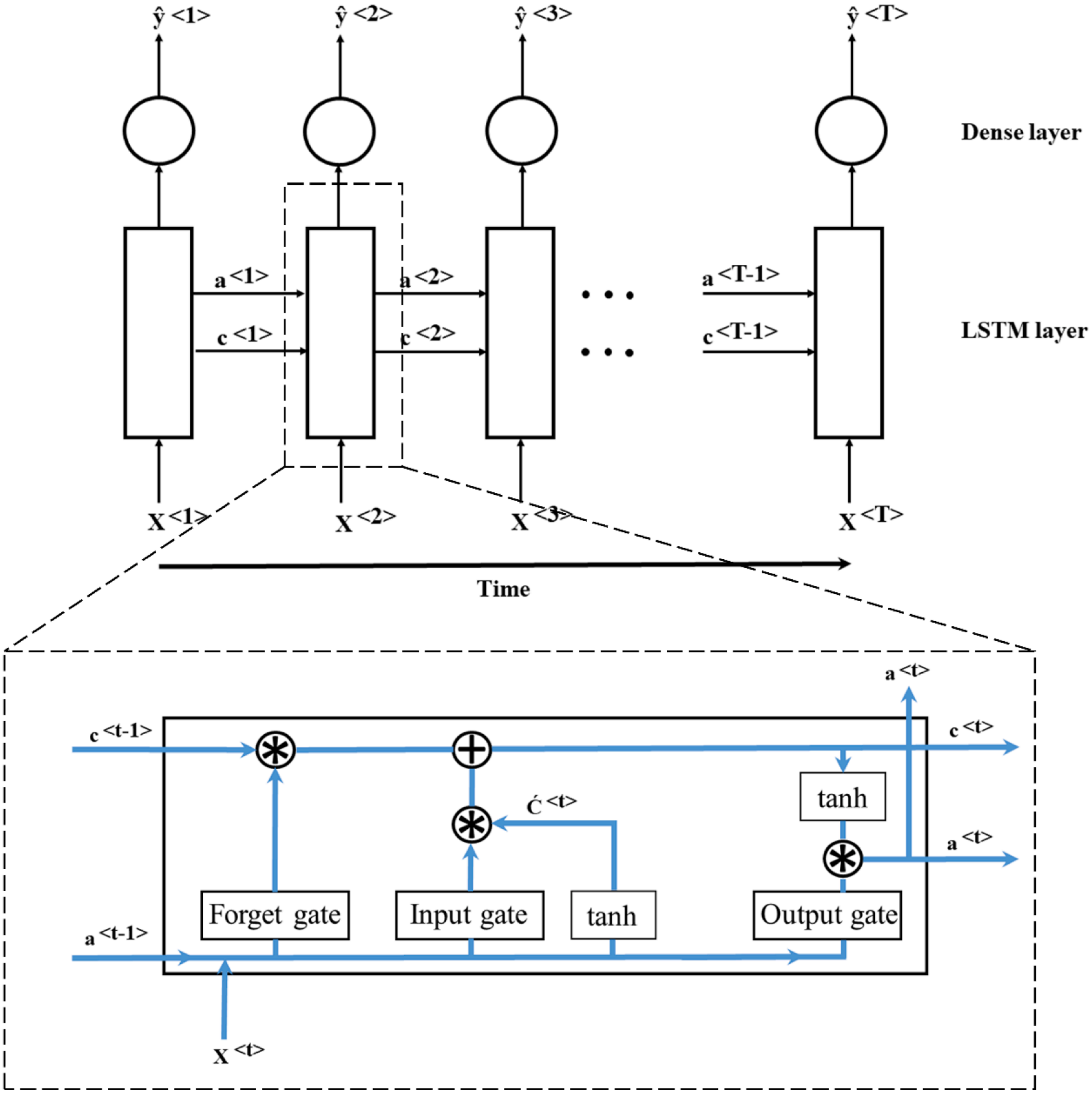


Fig. 5. The layout of the LSTM model.

2. Experimental rig and data acquisition

The photos and schematic diagram, as well as the components of the fabricated solar-driven DCMD system, are shown in Fig. 1 and, Fig. 2 respectively. The system is comprised of three main sections, namely the solar loop, feed water loop and coolant water loop.

Regarding the solar loop, a heat pipe evacuated tube solar collector, featuring as an efficient absorber and conductor of solar energy, was employed to absorb the solar energy. The absorbed heat was then transferred to the evaporator section of the heat pipes using aluminium fins inside the evacuated tubes. Via evaporation of ethanol as the working fluid of the heat pipes, the vapor moves towards the condenser section and transfers heat to the solar working fluid. The heated solar working fluid (state T4) then enters a coiled heat exchanger located in the feed water storage tank and heats up the saline feed water. The solar working fluid (state T2) is then pumped back to the solar collector. Table 1 illustrates the characteristics of the main components of the solar loop.

A tubular hollow fiber DCMD module with a total membrane area of 0.118 m^2, sourced from the Memsift company, was used in this study. The properties of the DCMD module are listed in Table 2. The hot saline feed water is pumped to the lumen side of the DCMD module (state T0) in the feed water loop. After undergoing heat and mass transfer inside the DCMD module, it returns to the feed storage tank (state T1). In the coolant loop, cold water is pumped from the coolant storage tank to the shell side of the DCMD module (state T3). After releasing its heat to the air-cooled condenser, the water is then pumped back to the coolant tank (state T5). Due to the vapor pressure difference between the feed and coolant flows in the DCMD module, vaporization occurs inside the hollow fiber membranes. As a result, the vapor permeates through the hydrophobic membrane and is then condensed in the coolant water loop. Lastly, the distilled water is added to the coolant storage tank. The characteristics of the air-cooled condenser are shown in Table 3. It is worth noting here that tap water was chosen as the coolant water, and feed water with a salinity of 3.5% (almost equivalent to seawater salinity) was synthesized and used for the tests. A conductivity meter (Multi 3410, WTW) was used to monitor the salinity of the feed water. Three valves and flowmeters (0 to 0.068 kg/s, Omega) were utilized to regulate and monitor the flow rates. For a better understanding of the effects of mass flow rates on the performance of the system, T-Class thermocouples (-185 to 300 °C, TC Measurement & Control Pty Ltd) were used to monitor the temperatures at different points of the system (T0-T5, Fig. 2). Furthermore, a pyranometer (0 to 2000 W/m^2) was used to monitor solar radiation during the experimental tests.

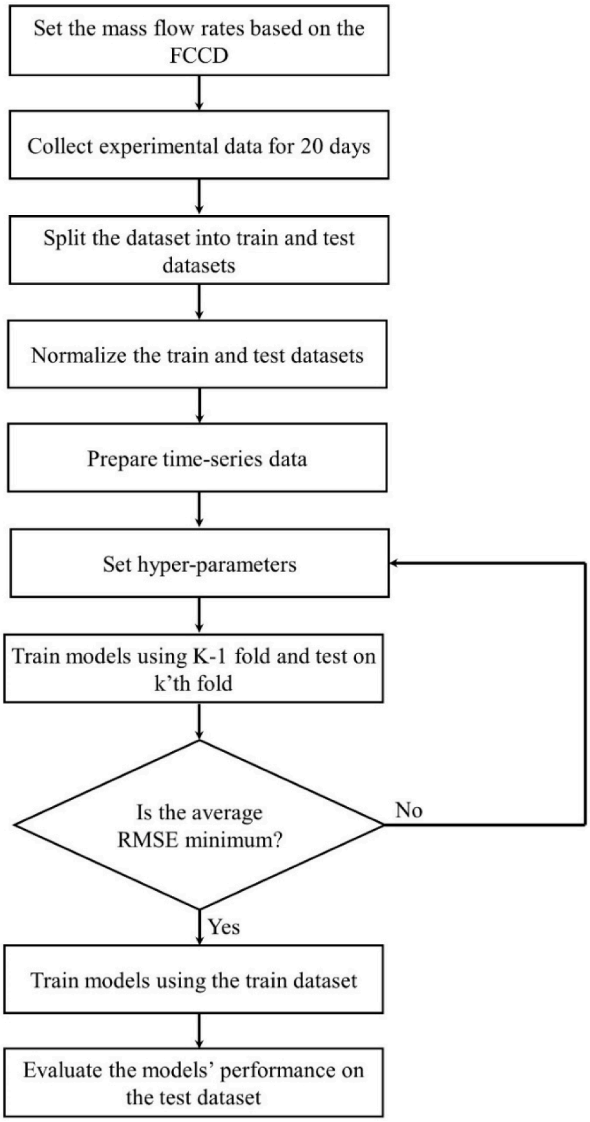


Fig. 6. The development procedure of machine learning models.

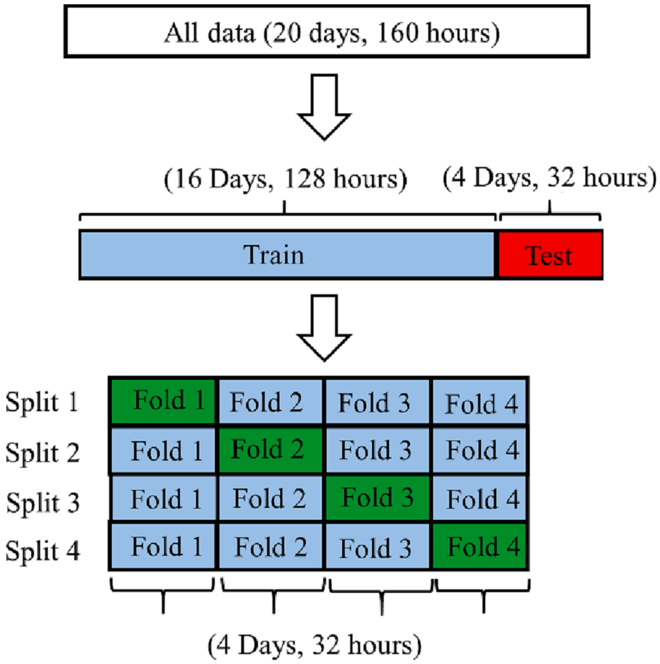


Fig. 7. K-fold cross-validation method.

3. Modelling methodology

The performance of a solar-driven DCMD may be influenced by solar working flow rate, feed flow rate and coolant flow rate. To understand the impact of these parameters, the face-centered composite design (FCCD) was applied to set the number of experimental tests and mass flow rates [41]. The application of the FCCD method can reduce the time and cost of the data collection process while providing sufficient information about system characteristics [35]. However, this method may not sufficiently capture the entire dynamic range of the system. As a result, including more gradual or continuous changes in the mass flow rates in the dataset design can help capture a broader range of frequencies, including both low and high frequencies. Using FCCD design, 20 experimental points were designed, categorized into central, axial and factorial points (Fig. 3). Table 4 illustrates the values of mass flow rates over the different days of experiments as designed by the FCCD. After determination of the mass flow rates for each day, experimental tests were conducted for 20 days between 9:30 AM and 4:30 PM in April and May 2022 under Perth weather conditions. The dynamic performance of the solar-driven DCMD system was then evaluated using two neural networks, namely multilayer perceptron (MLP) and long short-term memory (LSTM). Eleven input influential parameters were then selected to develop the MLP and LSTM models, including temperatures at different points (T0 to T5), ambient temperature, solar intensity and mass flow rates (Q_{solar}, Q_{feed}, and $Q_{coolant}$). Further, permeate flux, efficiency and GOR were considered as three targets of the machine learning methods. A summary of the experimental dataset is shown in Table 5. Permeate flux was calculated based on the collected freshwater in the coolant storage tank per membrane effective area. The efficiency of the system was also calculated from [12]:

$$\eta = \frac{(M_d{*}h_{fg})/3600}{AI + W_{pumps} + W_{fan}} \tag{1}$$

where M_d is produced hourly freshwater, h_{fg} refers to latent heat of evaporation of water, A is the area of solar collector, I represents solar intensity, W_{pumps} is the sum of power consumed by the three water pumps (a total of 201 W) and W_{fan} refers to the air fan power (35 W). To convert hour to second, the conversion factor of 3600 is used. Equation (1) represents the ratio of the latent heat of vaporization of the distilled water to the total energy input of the system, which includes the solar input energy and the electrical energy consumption of the fan and pumps [13,42,43]. The objective of this equation is to evaluate the effectiveness of active solar desalination systems by maximizing freshwater productivity while minimizing the total input energy required for the system. In fact, this performance indicator includes both thermal and electrical input energies. However, the performance of solar-driven DCMD systems can also be analysed using specific electrical-based performance indicators, such as specific energy consumption, thereby considering the distinction between the quality of energy inputs.

GOR is another key performance indicator of solar-driven desalination systems and is calculated as [10,12]:

$$GOR = \frac{(M_d{*}h_{fg})/3600}{\dot{m}_{feed}.C_p.(T0 - T1)} \tag{2}$$

Here, $\dot{m}_{feed}$ is the mass flow rate of feed water, and C_p represents the specific heat of feed water. $T0$ and $T1$ also show feed water inlet and outlet temperatures, respectively.

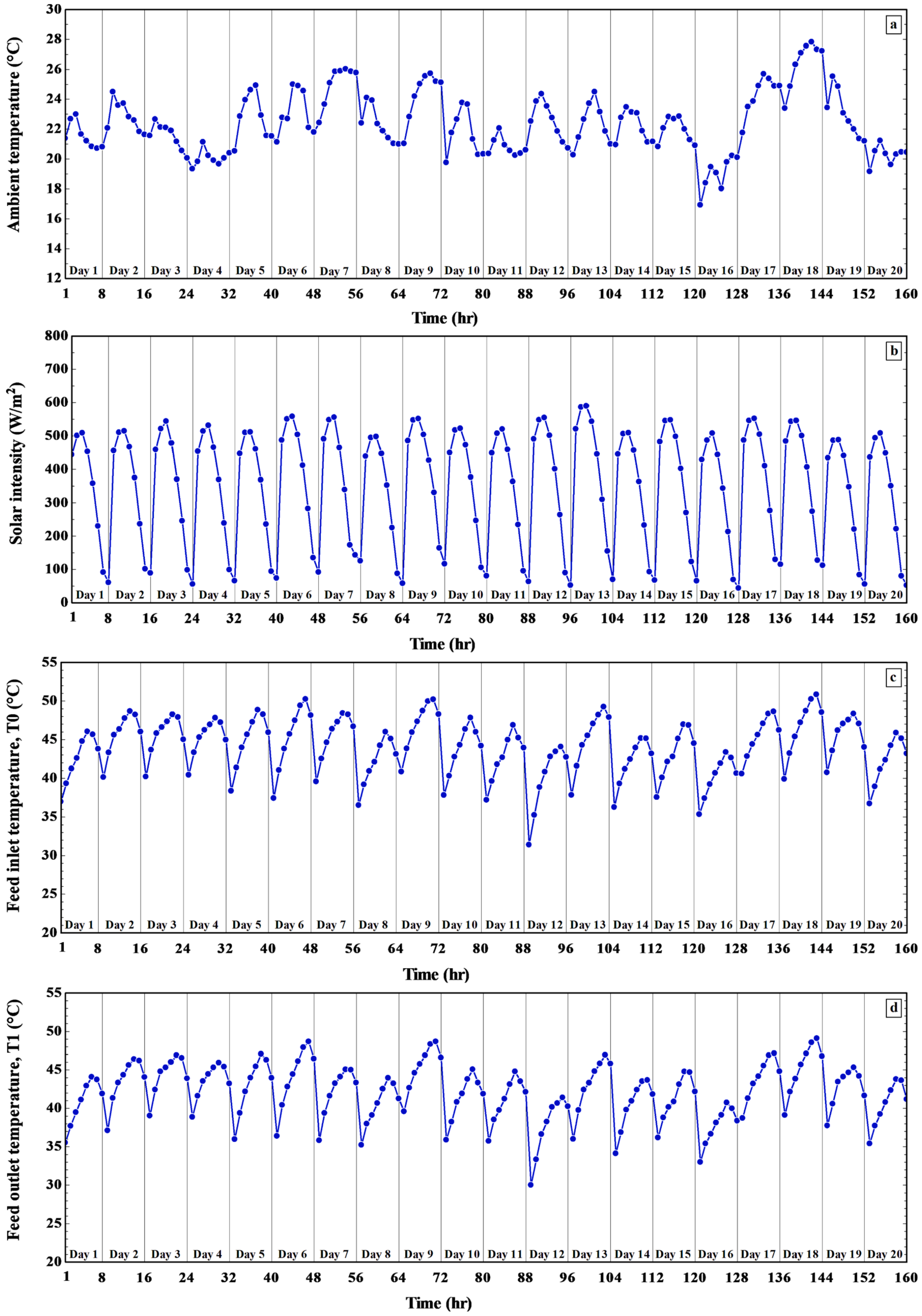


Fig. 8. The variations of influential parameters and three performance indicators over the 20 days of experiments.

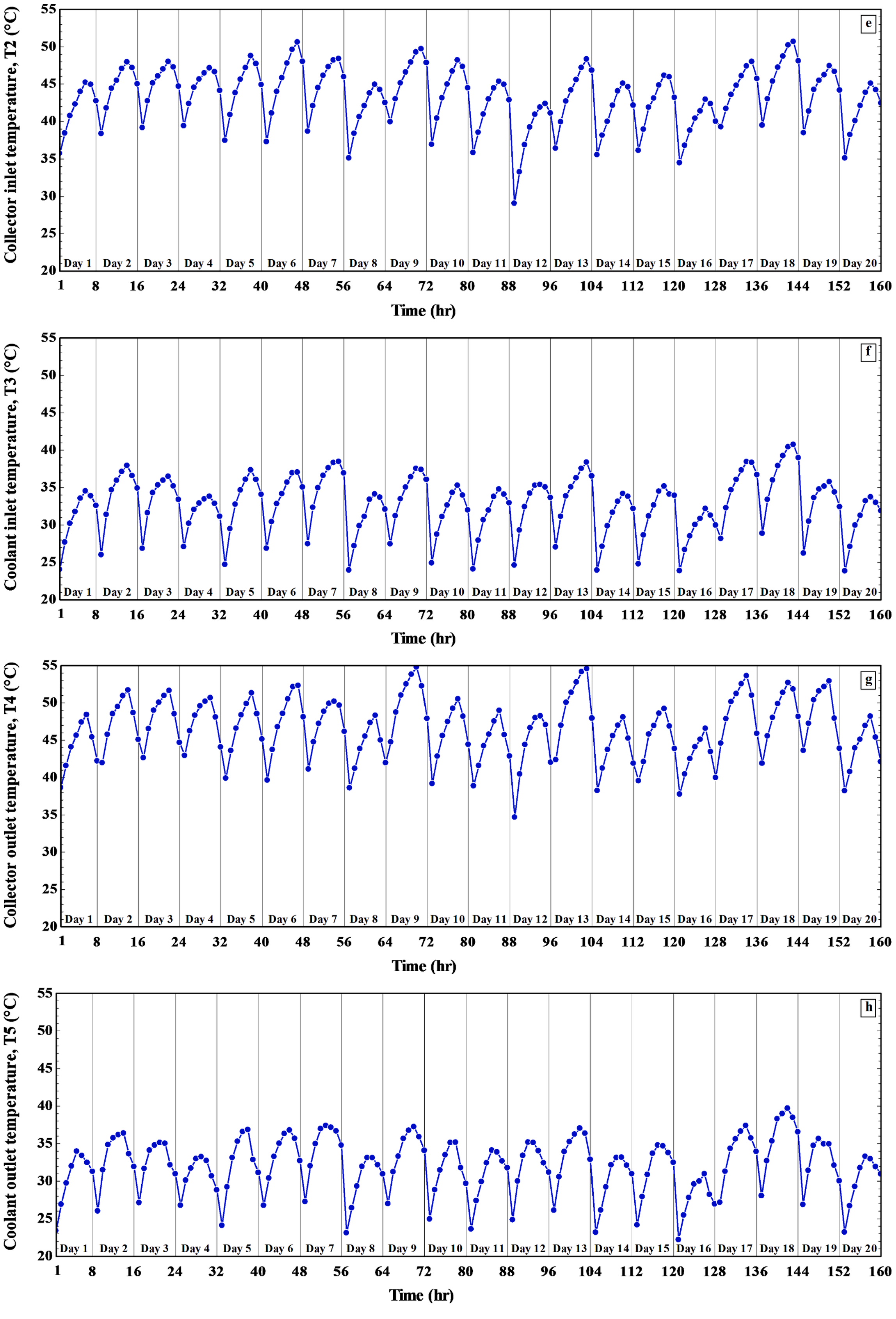


Fig. 8. (*continued*).

3.1. Multi-layer perceptron network (MLP)

Multi-layer perceptron network (MLP) is one of the artificial intelligence techniques that has been developed based on imitating the human brain. This model has been broadly applied for performance diagnoses of various energy systems due to its great ability to analyse

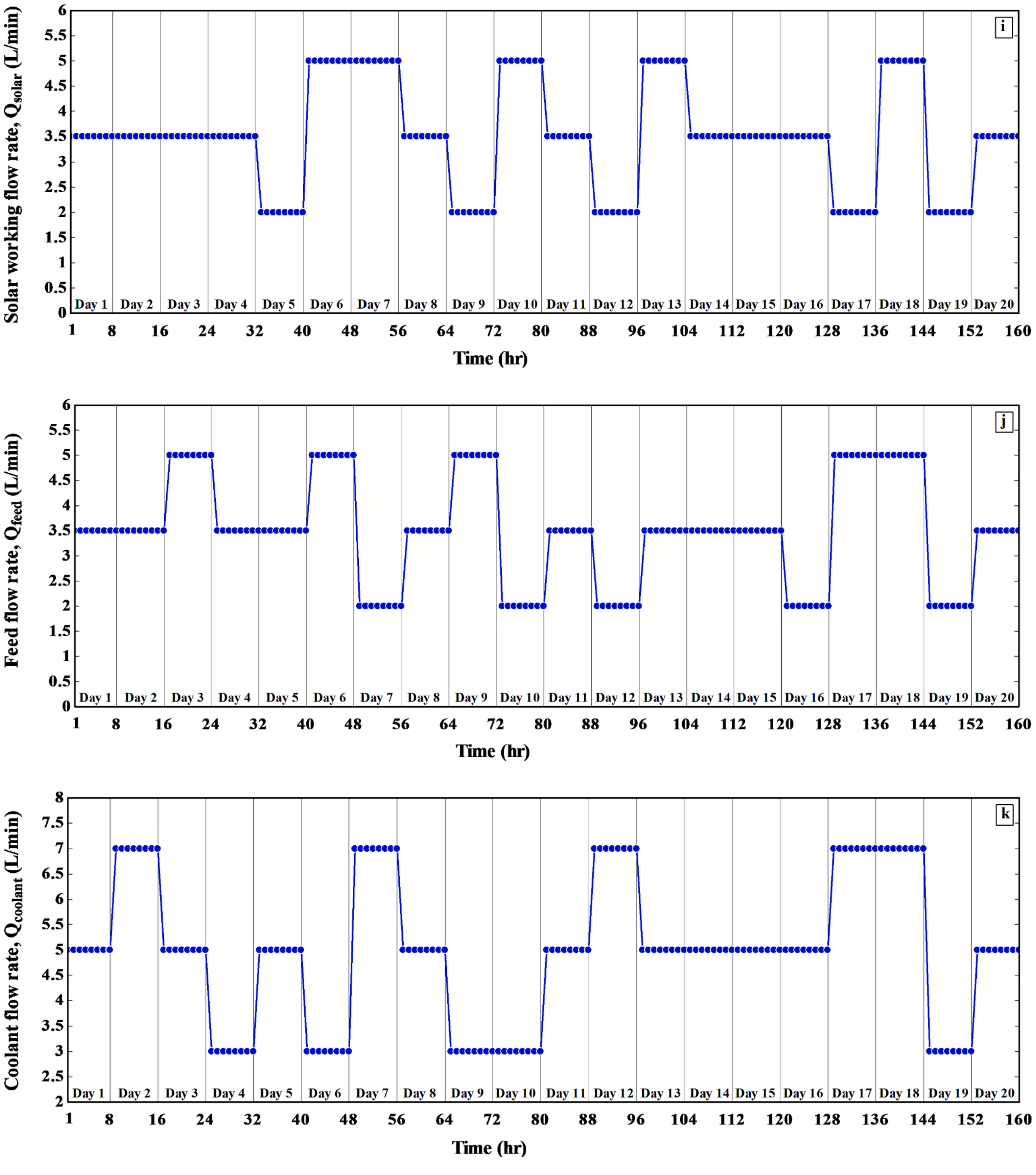


Fig. 8. (*continued*).

non-linear and complex systems [29,35,44]. As shown in Fig. 4, the MLP consists of input, hidden and output layers. This model, without detailed information about a system, can determine the relationship between inputs and outputs using processing units called neurons. The number of neurons in the input and output layers is equal to the number of inputs and outputs, respectively. However, the number of neurons in hidden layers and the number of hidden layers are unfixed, where MLP models can be constructed by stacking more hidden layers with many neurons. The design and training of the MLP model are composed of feed-forward propagation and backpropagation. In the former, the input data is fed to the input layer and after assignment by random initial weights and biases is transferred to the hidden layer. The weighted input data and biases are summed up inside the neurons of the hidden layer. The activation function is then applied to add non-linearity to the data which is used for the next hidden layer. Eventually, the weighted input data and bias is transferred into the output layer and by applying a linear activation function prediction occurs. In the backpropagation step, weights and biases are determined backwards (from the output layer to the input layer) using a gradient descent optimization algorithm [29].

3.2 Long short-term memory network (LSTM)

Long short-term memory network (LSTM) is a modified variant of recurrent neural networks. This robust machine learning model solves the gradient vanishing problem with recurrent neural networks and benefits from capturing long-term dependency in sequential data [45]. Fig. 5 depicts the schematic diagram of the developed LSTM model as well as the LSTM unit. The LSTM model has the benefit of storing

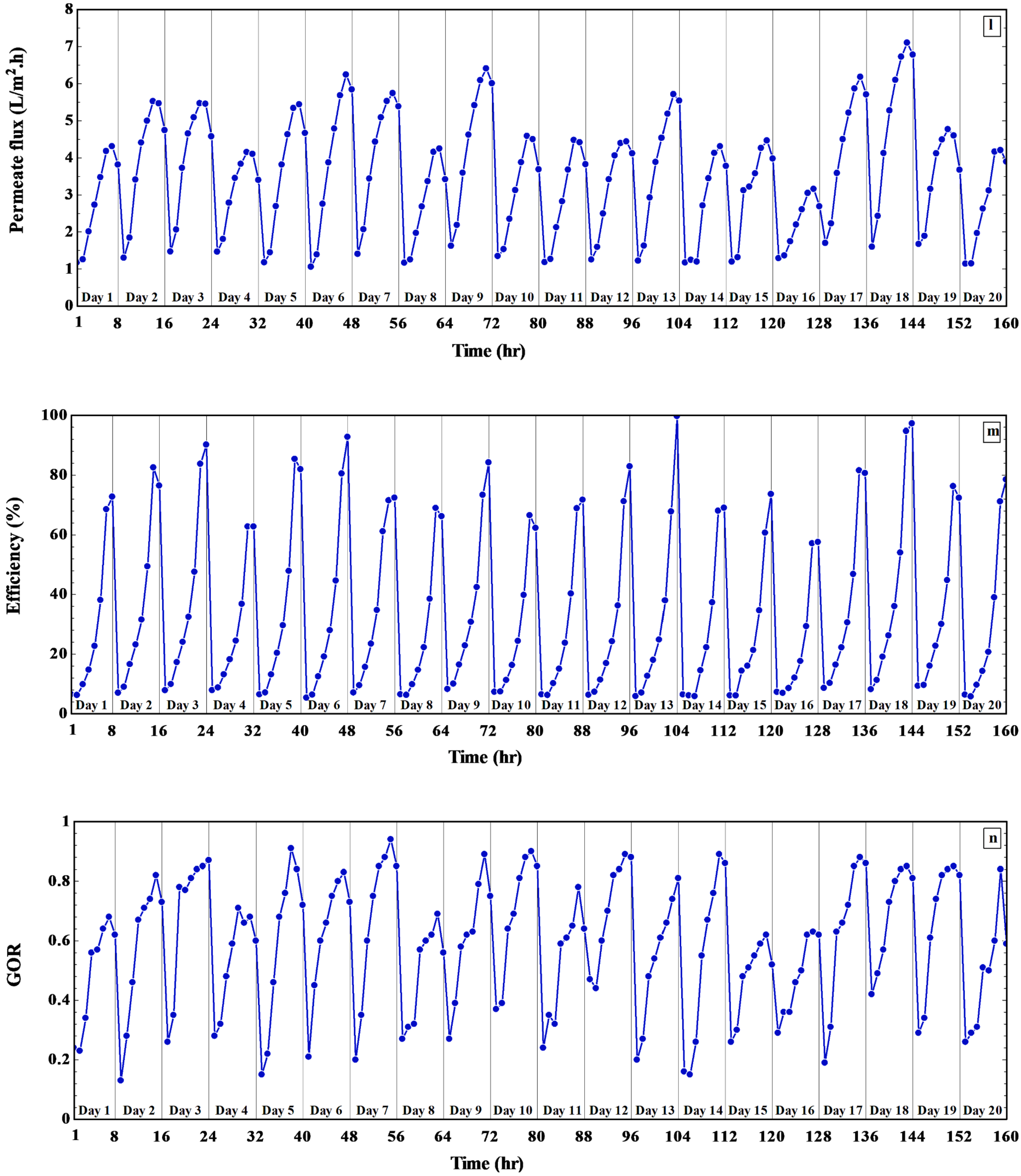


Fig. 8. (*continued*).

information throughout the sequential data using a memory cell. Moreover, updating the information from the previous timestep (t-1) to the current timestep (t) can be controlled by three gates namely forget, input and output gates. Each gate holds its weights (W, U) and bias (b), and uses a logistic activation function. As the output valve of the logistic activation function lies in the range of [0,1], updating the information flow can be effectively performed inside the LSTM units. When the output value is zero the information is not passed, while the output value is 1 the whole information is transferred. Each LSTM unit has three inputs including the memory cell of the previous timestep (c $^{<t-1>}$), the activation function output of the previous timestep (a $^{<t-1>}$) and current input data ($X^{<t>}$). Further, each LSTM unit owns two outputs: the memory cell of the current timestep (c $^{<t>}$) and the activation function output of the current timestep ($a<t>$). The forget gate (Γ_f) is utilized to remove the trivial information from the previous timestep:

$$\Gamma_f = \sigma\left(W_f X^{<t>} + U_f a^{<t-1>} + b_f\right) \tag{3}$$

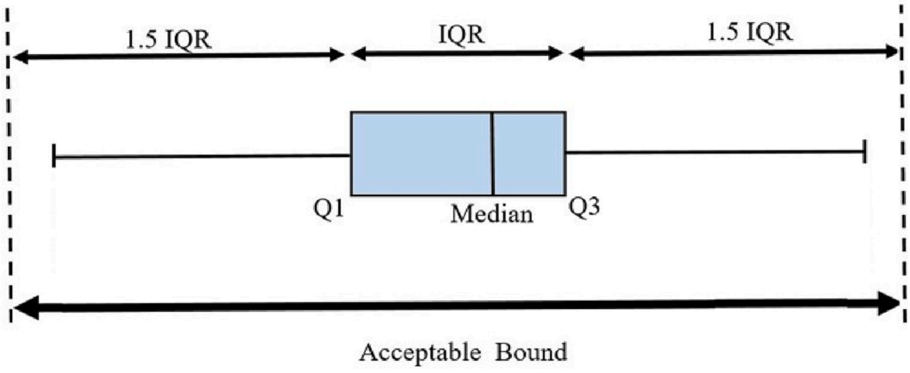


Fig. 9. Interquartile range method for outlier detection.

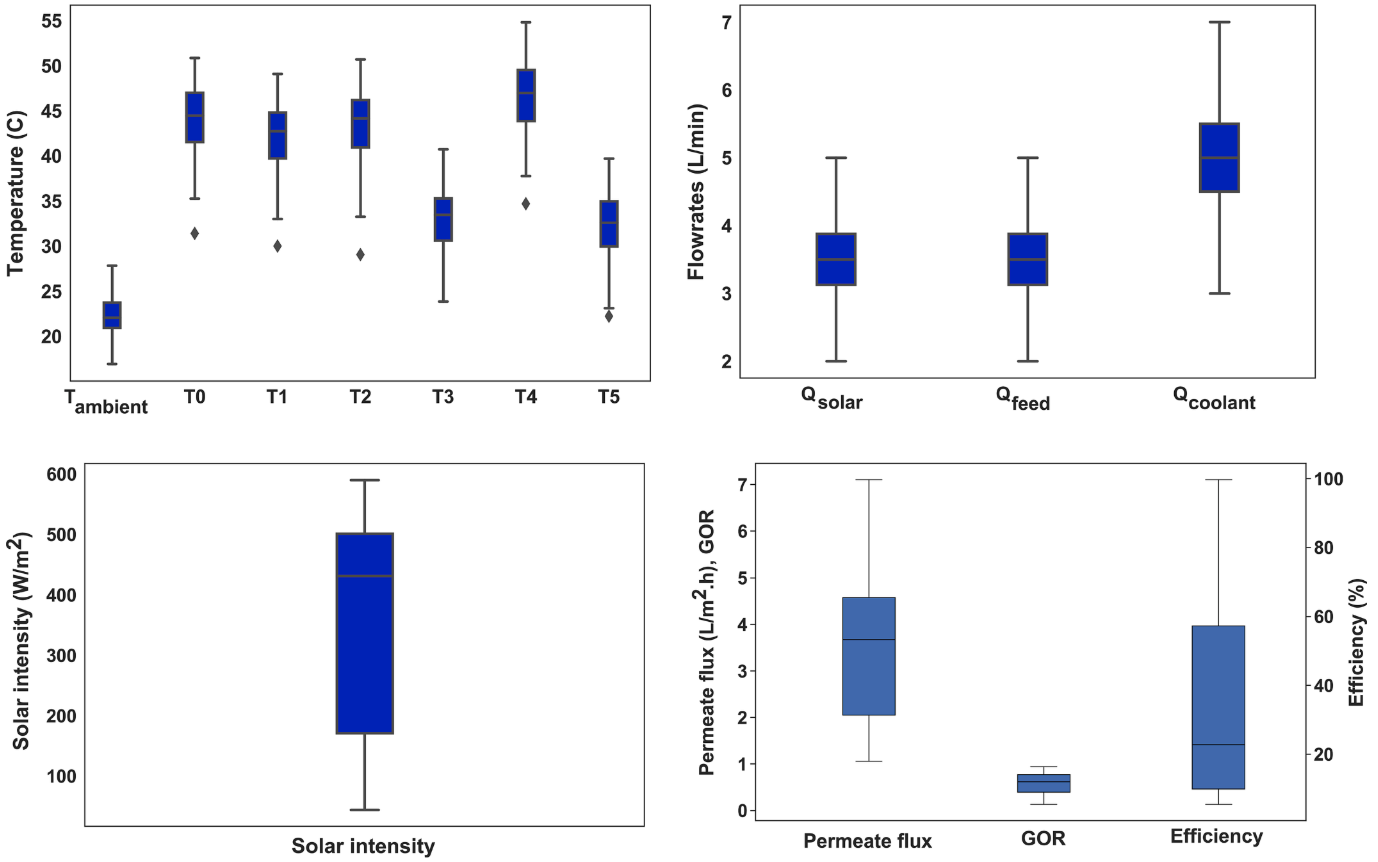


Fig. 10. Distribution of input parameters and performance indicators for the detection of outliers using the IQR method.

The input gate (Γ_i) determines the information which needs to be updated wherein potential cell sate ($\tilde{C}^{<t>}$) is generated:

$$\Gamma_i = \sigma\left(W_i X^{<t>} + U_i a^{<t-1>} + b_i\right) \tag{4}$$

$$\tilde{C}^{<t>} = tanh\left(W_c X^{<t>} + U_c a^{<t-1>} + b_c\right) \tag{5}$$

Here σ and *tanh* represent the logistic and hyperbolic tangent activation functions, respectively. As shown in Fig. 5, the memory cell at current timestep ($c^{<t>}$) is calculated as:

$$c^{<t>} = \Gamma_f * c^{<t-1>} + \Gamma_i * \tilde{C}^{<t>} \tag{6}$$

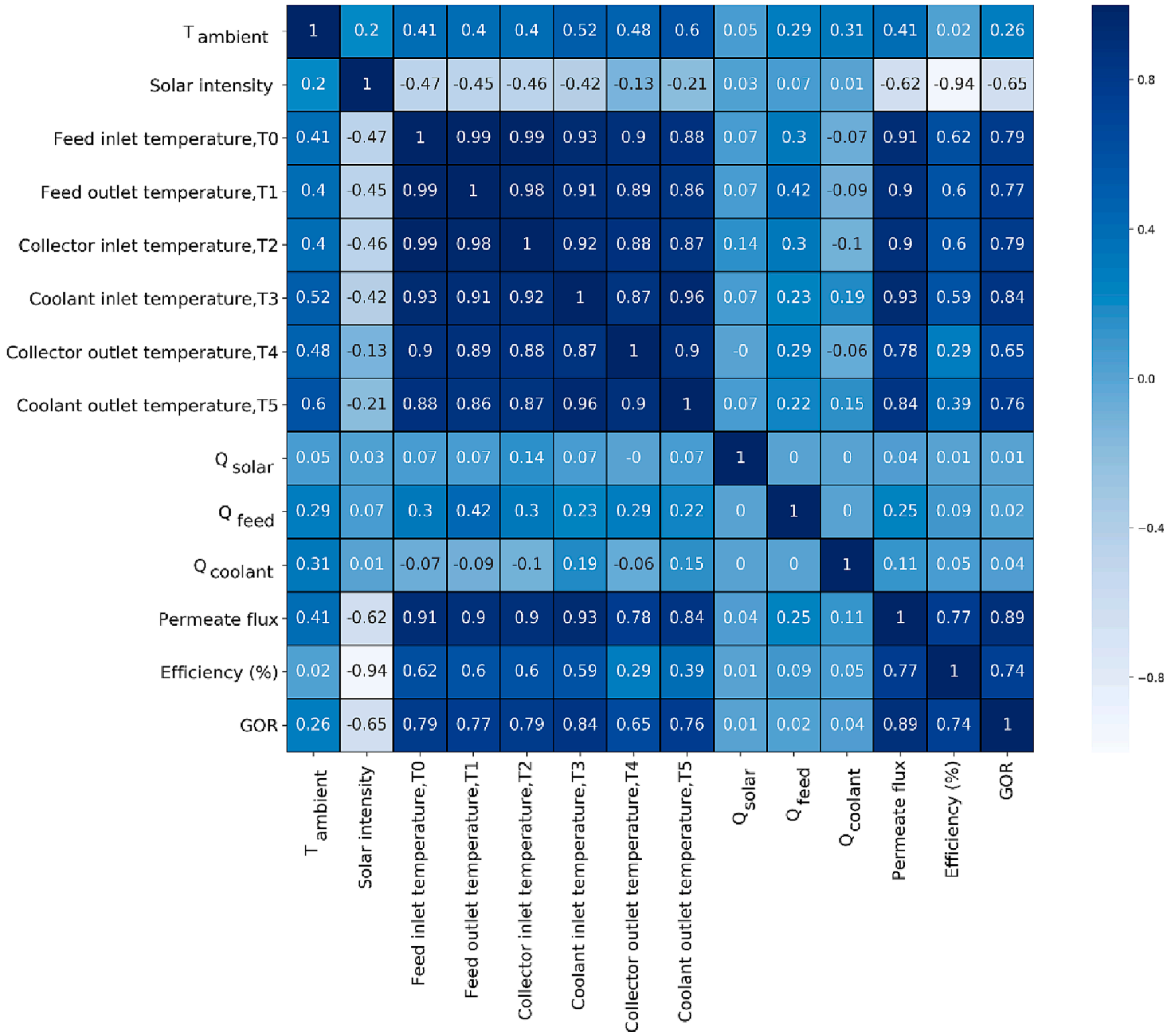
Lastly, the useful information in the memory cell of current timestep (c < t >) is used by the output gate (Γ_o) to calculate the activation function output of current timestep (a < t >):

$$\Gamma_o = \sigma\left(W_o X^{<t>} + U_o a^{<t-1>} + b_o\right) \tag{7}$$

$$a^{<t>} = \Gamma_o * \tanh\left(c^{<t>}\right) \tag{8}$$

3.3. Development of the MLP and LSTM models

Time-series input data was utilized in this study to develop MLP and LSTM models. The models used input data at the current time step (t) with lagged input data (t-1, t-2…) to predict each performance indicator (permeate flux, efficiency, and GOR). The number of lagged timesteps is a hyper-parameter and should be accurately determined in the hyper-parameter tuning procedure. To prepare the time-series data for both models, a Python code was developed, where a sliding window of $t_w \times N_f$



	$T_{ambient}$	Solar intensity	Feed inlet temperature,T0	Feed outlet temperature,T1	Collector inlet temperature,T2	Coolant inlet temperature,T3	Collector outlet temperature,T4	Coolant outlet temperature,T5	Q_{solar}	Q_{feed}	$Q_{coolant}$	Permeate flux	Efficiency (%)	GOR
$T_{ambient}$	1	0.2	0.41	0.4	0.4	0.52	0.48	0.6	0.05	0.29	0.31	0.41	0.02	0.26
Solar intensity	0.2	1	-0.47	-0.45	-0.46	-0.42	-0.13	-0.21	0.03	0.07	0.01	-0.62	-0.94	-0.65
Feed inlet temperature,T0	0.41	-0.47	1	0.99	0.99	0.93	0.9	0.88	0.07	0.3	-0.07	0.91	0.62	0.79
Feed outlet temperature,T1	0.4	-0.45	0.99	1	0.98	0.91	0.89	0.86	0.07	0.42	-0.09	0.9	0.6	0.77
Collector inlet temperature,T2	0.4	-0.46	0.99	0.98	1	0.92	0.88	0.87	0.14	0.3	-0.1	0.9	0.6	0.79
Coolant inlet temperature,T3	0.52	-0.42	0.93	0.91	0.92	1	0.87	0.96	0.07	0.23	0.19	0.93	0.59	0.84
Collector outlet temperature,T4	0.48	-0.13	0.9	0.89	0.88	0.87	1	0.9	-0	0.29	-0.06	0.78	0.29	0.65
Coolant outlet temperature,T5	0.6	-0.21	0.88	0.86	0.87	0.96	0.9	1	0.07	0.22	0.15	0.84	0.39	0.76
Q_{solar}	0.05	0.03	0.07	0.07	0.14	0.07	-0	0.07	1	0	0	0.04	0.01	0.01
Q_{feed}	0.29	0.07	0.3	0.42	0.3	0.23	0.29	0.22	0	1	0	0.25	0.09	0.02
$Q_{coolant}$	0.31	0.01	-0.07	-0.09	-0.1	0.19	-0.06	0.15	0	0	1	0.11	0.05	0.04
Permeate flux	0.41	-0.62	0.91	0.9	0.9	0.93	0.78	0.84	0.04	0.25	0.11	1	0.77	0.89
Efficiency (%)	0.02	-0.94	0.62	0.6	0.6	0.59	0.29	0.39	0.01	0.09	0.05	0.77	1	0.74
GOR	0.26	-0.65	0.79	0.77	0.79	0.84	0.65	0.76	0.01	0.02	0.04	0.89	0.74	1

Fig. 11. Heat map illustration of Pearson's correlation coefficients.

extracts the input data for the prediction of each target. The input data with a shape of $t_w \times N_f$ is then fed into the models, where t_w and N_f represent the number of timesteps and features, respectively. It is worth mentioning that the size of t_w in the sliding window affects the number of predictions. For instance, when $t_w = 2$ and $N_f = 11$, current time step input data and one lagged input data are fed into the model in a shape of 2×11, and totally Datasetsize-1 predictions are made via the developed models. To feed the MLP model, the shape of input data should be a flatten vector. Therefore, each input sample was reshaped to a vector with $t_w \times N_f$ elements [26].

Fig. 6 illustrates the modelling procedure employed in this study. The original dataset was split into train and test datasets. Accordingly, 80% of the data (16 days) was chosen for training and the remaining 20% (4 days) was allocated for testing the predictive performance of the developed models. Moreover, the K-fold cross-validation method (with $K = 4$) was applied to achieve the optimum hyper-parameters and increase the generalization capability of the models. As shown in Fig. 7, the training dataset was transformed into four different splits. Each model was trained using three folds, and the remaining fold (green fold) was used to evaluate the performance of the developed models. This procedure was repeated four times, and eventually, the average performance of the models was used to accurately set their hyper-parameters. To apply MLP and LSTM models, scikit-learn and Keras modules of Python language were employed, respectively [46,47]. Further, an Adam optimizer was used to train the models and minimize the loss function of the models (mean square error) [33]. To decrease the convergence time and increase the accuracy of the models, min–max normalization method was employed whereby input features have been scaled into [0,1] range as follows:

$$X_{scaled} = \frac{X - X_{min}}{X_{\max} - X_{\min}} \tag{9}$$

Here, X_{scaled} shows the scaled data. X, $X_{\min}$, and $X_{\max}$ also represent the actual data, minimum, and maximum data, respectively. Finally, the performance of the developed models was evaluated based on different statistical indicators, including mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE) and coefficient of determination (R^2) [48]:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}_i| \tag{10}$$

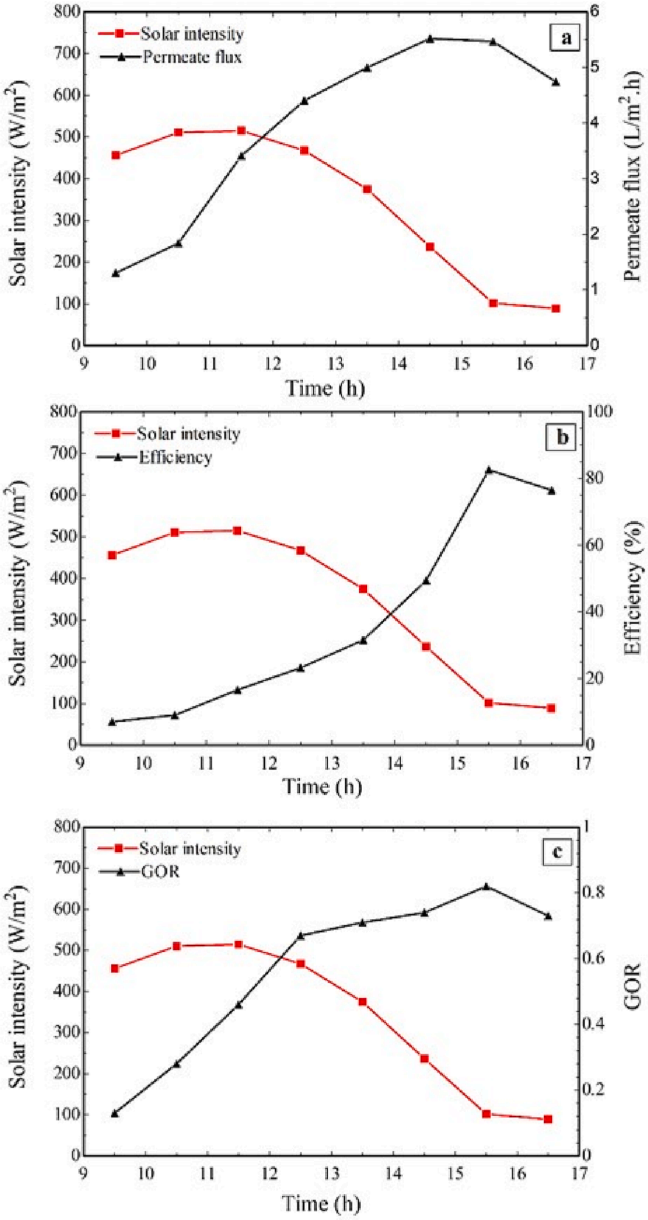


Fig. 12. The variations of solar intensity with permeate flux, efficiency, and GOR for the 2nd day of the experiment.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(y_i - \widehat{y}_i\right)^2}{N}} \tag{11}$$

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N} \left| \frac{\left(\widehat{y}_i - y_i\right)}{y_i} \right| * 100 \tag{12}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{N} \left(y_i - \overline{y}\right)^2} \tag{13}$$

where N shows the number of samples. y_i, $\widehat{y}_i$, and $\overline{y}$ also represent the actual data, the predicted value, and the mean value of the experimental data, respectively.

4. Results and discussion

This section examines the variations of input parameters and three performance indicators (permeate flux, efficiency and GOR) of the solar-driven DCMD system over the 20-day experiment. Pre-processing is also applied to determine the outliers and determine the effects of influencing parameters on the three performance indicators. The hyper-parameter tuning process based on the K-fold cross-validation method is then outlined in detail. The ability of the MLP and LSTM machine learning models for predicting each indicator is comprehensively examined. Eventually, the effect of dataset size on the predictive ability of the two developed models is investigated.

The performance of the solar-driven DCMD system can be affected by several varying parameters, including temperatures at different points (T0 to T5), weather conditions (ambient temperature and solar intensity) and mass flow rates (Q_{solar}, Q_{feed}, and Q_{coolant}). Fig. 8 illustrates the variations of these influential parameters along with variations in three performance indicators over the 20 days of experiments. Fig. 8(c-h) shows that temperatures are highly changing throughout the experiments. It can be seen from Fig. 8(c) that the maximum feed inlet temperature on the 18th day is 50.85 °C, while the maximum feed inlet temperature on the 12th day is 44.07 °C. This can be attributed to variations in weather conditions shown in Fig. 8(a, b) or the effects of mass flow rates (Fig. 8(i-k)). As mentioned before, mass flow rates were chosen based on the FCCD design, and their variations during the 20 days of experiments are shown in Fig. 8(i-k). As shown in Fig. 8(l), permeate flux significantly changed on different days of experiments. The maximum permeate flux on the 18th day was 7.103 L/m^2.h, whereas this value was only 3.155 L/m^2.h on the 16th day. It can be inferred from Fig. 8(l-n) that the maximum values of efficiency and GOR occur at the final hours of each day. At this time, the permeate flux is still significant while solar intensity as the main heat source of the system owns low values. Furthermore, Fig. 8(m, n) shows that GOR follows a more stochastic behaviour for each day of experiments compared to efficiency. The reason is that efficiency is directly dependent on solar intensity which almost followed a similar trend over the 20 days of experiments (Fig. 8(b)). However, GOR is mainly affected by feed mass flow rate (Q_{feed}) and the temperature difference between the inlet and outlet of the membrane module (T0-T3). This leads to more significant variations in GOR on each day the of experiment.

4.1. *Pre-processing of experimental data*

Pre-processing of data plays an important role in the development of machine learning models. This can lead to better data interpretation and the detection of suspected data samples. The first step in the pre-processing analysis is the detection of outliers, meaning data samples that are largley different from the rest of the data samples. The detection and then removing these abnormal data samples can increase the accuracy of the machine learning models. The interquartile range (IQR) method has been applied in this study to determine the outliers. In this process, the variations of each input parameter or performance indicator over the 20 days of experiments can be drawn using a boxplot, as shown in Fig. 9. Q1 represents the first quartile of the dataset, which shows 25% of the data samples are restricted between the Q1 and minimum. The median is the second quartile and shows the center point of the dataset. Q3 is the third quartile of the dataset, representing the 25% of data samples are limited between Q3 and maximum. The IQR is defined by subtraction of Q3 from Q1 and the acceptable bound is determined based on lower and upper limits, where the lower limit = Q1–1.5 IQR and upper limit = Q3 + 1.5 IQR. Each data sample outside of the acceptable bound is considered an outlier.

Fig. 10 illustrates the distribution of inputs and performance indicators based on the IQR outlier detection method. It is evident that there are only a few outliers for *T0*, *T1*, *T2*, *T4*, and *T5*. Since all these outliers are below the lower limit, they may be derived from different reasons such as cloudy weather in some hours. Although removing these few outliers may result in the development of more accurate models, they were not removed from the dataset as they may provide valuable

Table 6
Selection of the best architecture of the MLP model.

Architecture	Timestep	Number of layers	Number of neurons in each layer	MAE (L/m^2.h)	RMSE (L/m^2.h)	MAPE (%)	R^2
Permeate flux							
1	2	1	50	0.192	0.258	8	0.968
2	3	1	50	0.168	0.24	7.3	0.970
3	4	1	50	0.172	0.224	7.14	0.976
4	4	1	100	0.134	0.194	5.608	0.983
5	4	1	150	0.127	0.169	4.869	0.984
6	4	1	200	0.122	0.151	5.042	0.987
7	4	2	150	0.054	0.080	1.874	0.997
8	4	3	150	0.043	0.044	1.248	0.998
Efficiency							
1	2	1	50	11.115	13.042	62.04	0.739
2	3	1	50	10.436	13.249	62.43	0.774
3	4	1	50	10.616	13.173	61.69	0.749
4	3	1	100	9.437	11.081	58.87	0.817
5	3	1	150	9.221	10.343	56.69	0.845
6	3	1	200	8.565	9.983	53.18	0.858
7	3	2	200	1.543	2.193	7.387	0.993
8	3	3	200	0.722	0.962	3.43	0.998
GOR							
1	2	1	50	0.043	0.053	8.993	0.918
2	3	1	50	0.040	0.061	7.061	0.926
3	4	1	50	0.038	0.052	8.204	0.931
4	4	1	100	0.029	0.043	5.627	0.953
5	4	1	150	0.026	0.042	5.840	0.962
6	4	1	200	0.025	0.036	4.439	0.954
7	4	2	200	0.021	0.029	4.659	0.966
8	4	3	200	0.021	0.030	3.214	0.969

Table 7
Selection of the best architecture of the LSTM model.

Architecture	Timestep	Number of layers	Number of neurons in each layer	MAE (L/m^2.h)	RMSE (L/m^2.h)	MAPE (%)	R^2
Permeate flux							
1	2	1	50	0.161	0.207	6.381	0.975
2	3	1	50	0.129	0.177	5.283	0.985
3	4	1	50	0.117	0.172	4.797	0.983
4	4	1	100	0.107	0.148	4.411	0.988
5	4	1	150	0.107	0.134	3.777	0.988
6	4	1	200	0.093	0.130	4.280	0.992
7	4	2	200	0.095	0.122	3.173	0.990
8	4	3	200	0.103	0.145	3.672	0.992
Efficiency							
1	2	1	50	4.306	6.052	29.09	0.939
2	3	1	50	4.330	3.924	20.560	0.967
3	4	1	50	4.150	5.370	16.109	0.949
4	4	1	100	3.998	4.249	15.85	0.957
5	4	1	150	3.263	5.394	19.32	0.958
6	4	1	200	3.280	4.271	15.35	0.976
7	4	2	200	2.260	3.088	12.01	0.977
8	4	3	200	2.33	2.53	9.68	0.990
GOR							
1	2	1	50	0.046	0.059	9.69	0.911
2	3	1	50	0.036	0.053	7.46	0.915
3	4	1	50	0.035	0.050	7.06	0.939
4	4	1	100	0.031	0.043	5.99	0.954
5	4	1	150	0.028	0.040	5.45	0.947
6	4	1	200	0.024	0.035	4.75	0.957
7	4	2	200	0.025	0.040	6.21	0.951
8	4	3	200	0.032	0.041	4.98	0.942

information during an unexpected period.

The correlation between input parameters and their effects on the performance of the solar-driven DCMD system can be investigated using Pearson's correlation analysis [49]. As shown in Fig. 11, mass flow rates have much lower effects on the three performance indicators compared to temperatures at different points (T0 to T5) and solar intensity. Further, it can be inferred from Fig. 11 that temperatures at different points are highly correlated with each other. This is because of the interaction effects of three closed loops of the system on each other (i.e., solar loop, feed loop and coolant loop). Moreover, there is a negative

Table 8
The best hyper-parameters for MLP and LSTM models.

Indicator	MLP			LSTM		
	Timestep	Number of layers	Number of neurons in each layer	Timestep	Number of layers	Number of neurons in each layer
Permeate flux	4	3	150	4	2	200
Efficiency	3	3	200	4	3	200
GOR	4	3	200	4	1	200

Table 9
Performance of each model for the prediction of each indicator on the cross-validation dataset.

Indicator	MLP				LSTM			
	MAE	RMSE	MAPE	R^2	MAE	RMSE	MAPE	R^2
Permeate flux	0.043	0.044	1.248	0.998	0.095	0.122	3.173	0.99
Efficiency	0.722	0.962	3.43	0.998	2.33	2.53	9.68	0.99
GOR	0.021	0.030	3.214	0.969	0.024	0.035	4.75	0.957

Table 10
Performance of the models for the prediction of each target on train and test datasets (values in brackets represent the LSTM model).

Indicator	Train				Test			
	MAE	RMSE	MAPE	R^2	MAE	RMSE	MAPE	R^2
Permeate flux	0.046 [0.080]	0.064 [0.119]	1.652 [3.270]	0.998 [0.993]	0.204 [0.260]	0.261 [0.294]	6.950 [8.952]	0.977 [0.971]
Efficiency	0.782 [1.005]	1.180 [1.420]	3.111 [4.305]	0.998 [0.997]	1.803 [4.991]	2.848 [7.298]	5.453 [12.968]	0.990 [0.937]
GOR	0.041 [0.048]	0.059 [0.065]	8.849 [11.083]	0.921 [0.904]	0.046 [0.051]	0.062 [0.064]	8.963 [9.691]	0.907 [0.900]

correlation between solar intensity and the three performance criteria. The underlying reason is that a higher freshwater productivity has been recorded in the system at the late hours of the day when solar intensity is close to its lowest values. To further clarify this, the variations of solar intensity with three performance indicators of the system for the second day of the experiment are shown in Fig. 12. It can be seen that the maximum solar intensity happened around 11:30 while the maximum permeate flux, efficiency, and GOR were achieved at 14:30, 15:30, and 15:30, respectively. As the studied system is an indirect solar-driven DCDM system [8], the maximum solar intensity values show their effects on the performance of the system after 3–4 h. This is primarily due to warming up the water inside the feed storage tank. This can be also justified by negative correlations between solar intensity and temperatures at different points (T0-T5), as shown in Fig. 11. This shows that including lagged input data in machine learning models can improve their accuracy and provide valuable information for developing more accurate performance predictive models.

4.2. Hyper-parameter tuning using K-fold cross-validation method

Accurate hyper-parameter tuning not only can result in better training of machine learning models, but can also significantly enhance their generalization capability. Hence, the K-fold cross-validation method has been applied in this study. To develop MLP and LSTM models, three main hyper-parameters, namely timestep, the number of hidden layers and the number of neurons in each hidden layer have been considered. The performance of the developed models was then evaluated on the cross-validation dataset. The number of timesteps, hidden layers and neurons has been changed in the (2-3-4), (1-2-3), and (50-100-150-200) ranges, respectively. By applying a stepwise search method, the optimum value of timestep is initially estimated when models have been developed using 1 hidden layer and 50 neurons. Afterward, the effect of increasing the number of neurons on the performance of the models was estimated when timestep was kept at its optimum value and models were constructed using 1 hidden layer. Finally, deeper models with more hidden layers were investigated when timesteps and the number of neurons were kept fixed at their optimum values. Table 6 and Table 7 indicate the hyper-parameter tuning process of MLP and LSTM models, respectively. It can be inferred that both MLP and LSTM models had a better generalization capability by including more lagged data (3 or 4). The optimum values of hyper-parameters for the prediction permeate flux, efficiency and GOR are summarized in Table 8. The predictive performance of each model on the cross-validation dataset for the prediction of each performance indicator is also shown in Table 9. Both models exhibit great accuracy on cross-validation data samples with MAPE $< 10\%$ and R2 > 0.95. It can be also inferred that MLP outperforms LSTM model on the cross-validation dataset. As shown in Table 6 and Table 7, neural networks with fewer neurons (i.e., 50 neurons) also demonstrated acceptable predictive performance, with most R^2 values exceeding 0.9. However, to achieve the best performance, more complex models with a higher number of neurons were selected as the final models. It is worth noting that despite using a higher number of neurons, the models displayed excellent generalization capabilities, as confirmed by the results of the cross-validation analysis presented in Table 9.

4.3. Predictive performance analysis

After training the machine learning models with optimum hyper-parameters, the predictive accuracy of both models was evaluated on test data samples that were not used in the models during their development. Table 10 illustrates the performance of the MLP and LSTM models on train and test datasets for the prediction of each indicator. Higher values of R^2 and lower values of errors (MAE, RMSE, and MAPE) represent the better performance of the models. The results show that MLP has better predictive performance on both training and test datasets

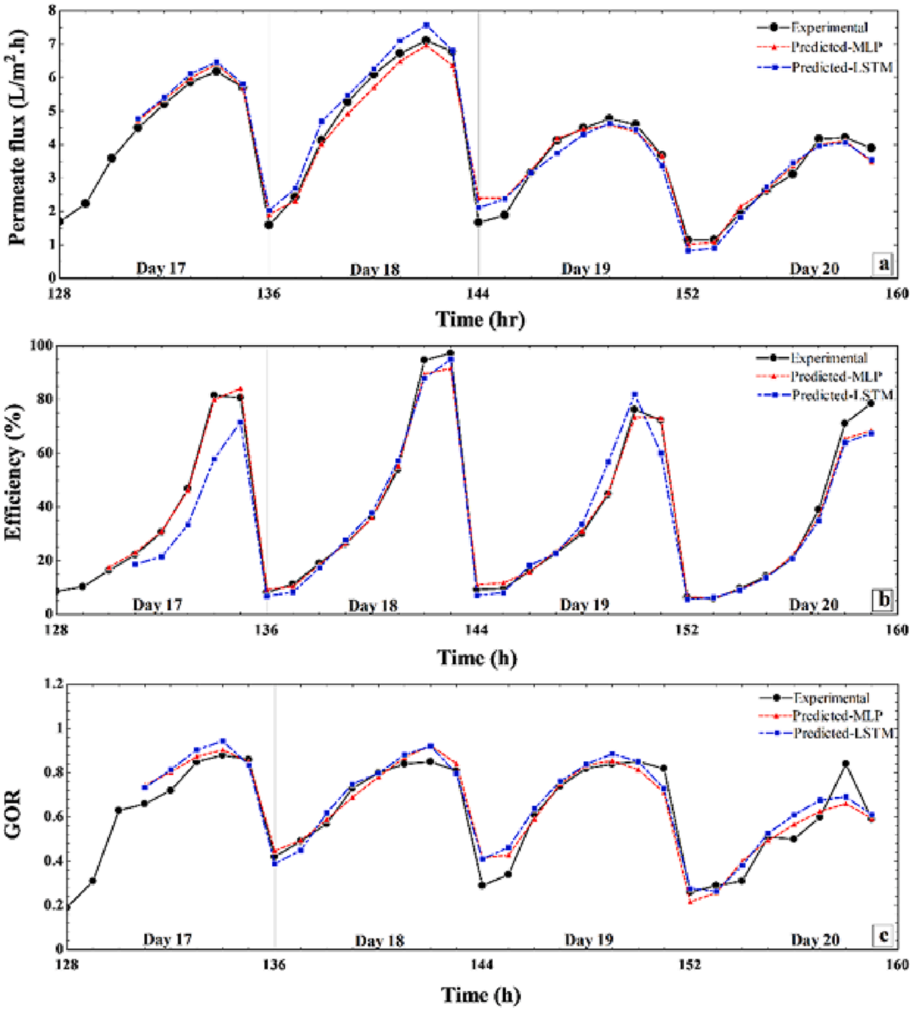


Fig. 13. Experimental data versus models’ prediction on test dataset: (a). permeate flux, (b). efficiency, (c). GOR.

than the LSTM model for the prediction of all three indicators. This superiority is highlighted more in the prediction of system efficiency, where R^2_{test} and $MAPE_{test}$ for the MLP model are 0.99 and 5.453% while these values are 0.93 and 12.96% for the LSTM model, respectively. Furthermore, Table 10 shows that both models have lower predictive performance for the prediction of GOR based on the R^2_{test}, compared to the prediction of permeate flux and efficiency.

Fig. 13 depicts the capability of MLP and LSTM models for the performance prediction of the solar-driven DCMD system using the test dataset. It is apparent that both models almost accurately predict the dynamic behaviour of the solar DCMD system. As shown in Fig. 13(a), the maximum deviations from the experimental values for both models happen at the first hour of operation (9:30 AM). The machine learning models use input data from the current timestep (9:30 AM) as well as lagged input data from the previous day to predict the system’s performance at 9:30 AM. As a result, both MLP and LSTM models tend to overpredict at the first hour of operation. This is primarily because there is a significant difference in permeate flux between the late hours of the previous day and the first hour of the current day (9:30 AM). The better predictive performance of the MLP model compared to the LSTM model for the prediction of efficiency can be seen in Fig. 13(b). The LSTM model tends to underpredict the experimental efficiencies for some days, e.g., 17th and 20 th days. Thus, it can be observed from Fig. 13(c) that both models exhibit a lower accuracy for the trend prediction of GOR compared to the efficiency of the system. This is mainly due to the more fluctuating behavior of GOR compared to efficiency.

To better diagnose the performance of machine learning models for the modeling of the solar-driven DCMD system, relative deviations are also graphically shown in Fig. 14. As depicted in Fig. 14(a), the MLP model outperformed the LSTM model on the training dataset with more data samples close to zero deviation. Moreover, the maximum relative deviations of both models on the test dataset occur at low permeate fluxes (between 1 and 2 $L/m^2.h$), which relates to the first hours of the days (see Fig. 13(a)). It can be seen from Fig. 14(b) that the MLP model with more test data samples concentrated around zero has much better predictions compared to the LSTM for the prediction of efficiency. This also justifies the listed results in Table 10. Despite the application of the K-fold cross-validation method for accurate hyper-parameter tuning, Fig. 14(c) shows that both models have higher values of relative deviations for the prediction of GOR on both training and test datasets. This is expected that increasing the number of data samples may result in performance enhancement of the models for more accurate prediction of GOR.

4.4. *Influence of dataset size*

The predictivity robustness of the studied models can be significantly affected by the size of the dataset. The development of accurate machine learning models with small-sized datasets can also considerably lower the time and cost associated with the data acquisition process. Hence, the effect of dataset size on the performance of the MLP and LSTM models for the performance prediction of the solar-driven DCMD system

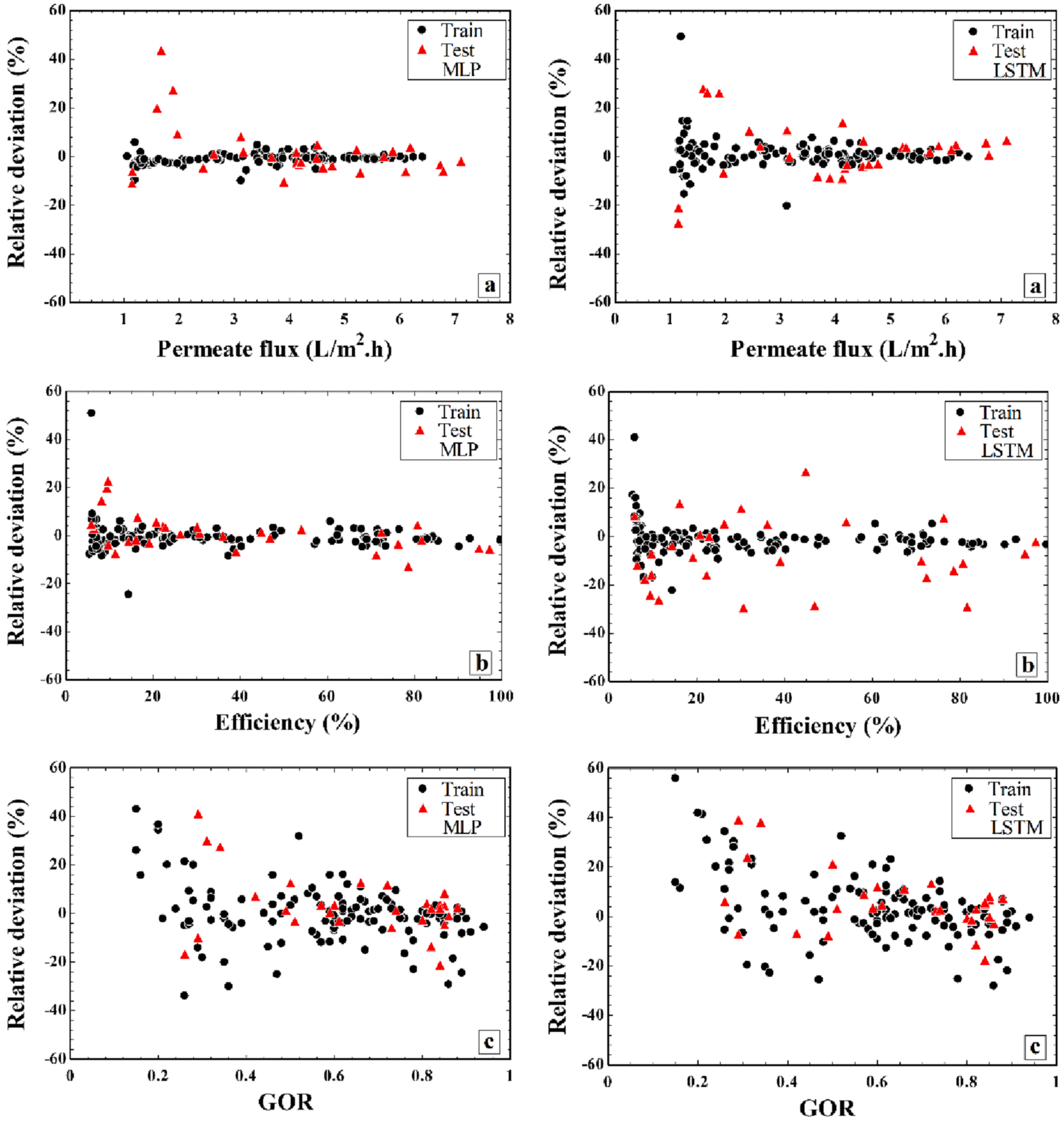


Fig. 14. Relative deviations of predicted values from experimental data for the prediction of three indicators: (a) permeate flux, (b) efficiency, (c) GOR.

has been investigated in this study. Fig. 15 compares the accuracy of the models trained and tested with different sizes of the dataset. Accordingly, 80% of the dataset was used for training and 20% was used for testing the performance of the studied models. It can be seen that the performance of both models is enhanced by increasing the number of data samples. Further, MLP exhibits better performance than the LSTM model for different sizes of the dataset. Fig. 15(a, b) demonstrates that increasing the number of data samples from 40 to 80 leads to a significant improvement in the performance of both models for predicting permeate flux and efficiency. However, increasing the number of samples from 80 to 160 only results in a slight improvement. In contrast, Fig. 15(c) illustrates that the performance of MLP and LSTM models for the prediction of GOR indicator improves significantly with increasing the number of samples from 40 to 160. It can be inferred that increasing the number of samples can improve the predictive performance of the studied models in terms of the GOR indicator which followed a more stochastic behaviour for each day of experiments.

5. Conclusions

In this study, the capability of two neural networks was comprehensively investigated for dynamic performance modelling of a solar-driven DCMD system for the first time. The models were developed based on data collected over 20 days from a small-scale solar-driven DCMD system that was experimented with three mass flow rates: solar working flow rate, feed flow rate and coolant flow rate. These mass flow rates were set appropriately based on the faced-centered-composite design approach. The effects of different input parameters (ambient temperature, solar intensity, feed inlet temperature (T0), feed outlet temperature (T1), collector inlet temperature (T2), coolant inlet temperature (T3), collector outlet temperature (T4), coolant outlet temperature (T5), solar working flow rate (Q_{solar}), feed flow rate (Q_{feed}) and coolant flow rate ($Q_{coolant}$)) on three performance indicators (permeate flux, efficiency and GOR) were then captured. The advanced K-fold cross-validation approach was applied to properly tune the hyper-parameters and the adequacy of the studied models was then

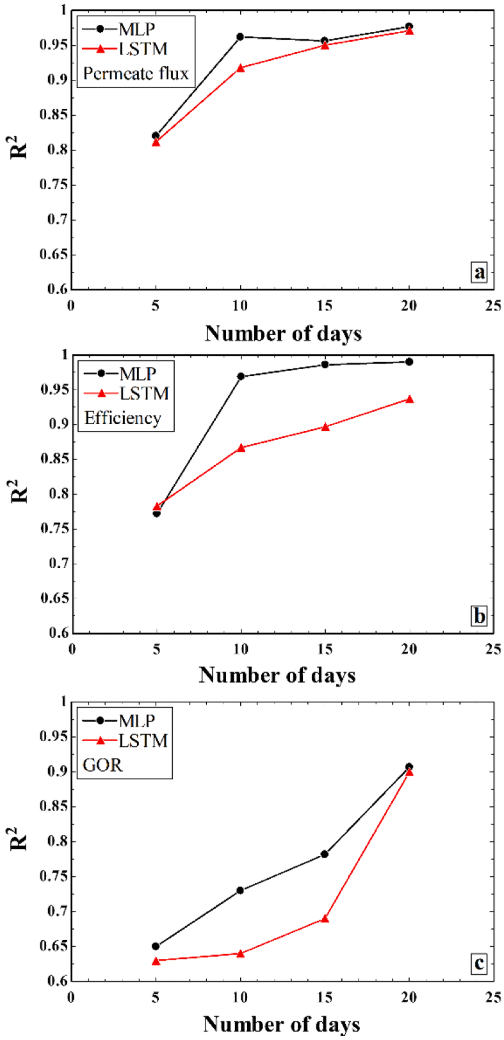


Fig. 15. Effect of dataset size on the predictive performance of MLP and LSTM models for the prediction of (a) permeate flux, (b) efficiency, and (c) GOR.

examined based on the unseen test dataset. The robustness of the developed models for dynamic performance modelling of the solar-driven DCMD system was comprehensively compared and the following main conclusions were acquired:

- Analysis of the dataset showed that the highest efficiencies and GORs happened at the late hours of each day, where permeate flux remained significant whereas solar intensity had low values.
- Mass flow rates had insignificant effects on the performance of the solar-driven DCMD system in comparison with temperatures at different points (T0 to T5), and solar intensity.
- Solar intensity as the main driving force of the system was negatively correlated with the three performance indicators. This was mainly due to higher freshwater productivity at late hours of the day when solar intensity is low.
- The MLP model outperformed the LSTM model for dynamic modelling of the solar-driven DCMD system. This superiority was highlighted in the prediction of efficiency of the system, where R^2_{test} and $MAPE_{test}$ for the MLP model were 0.99 and 5.453% while these values reached 0.93 and 12.96% for the LSTM model.
- Statistical criteria and relative deviation analysis revealed that both models had lower predictive performance for the prediction of GOR, compared to the prediction of permeate flux and efficiency mainly due to its abrupt variations.
- The generalization capability of both models was enhanced by increasing the number of data samples. Moreover, increasing the number of data samples significantly improved the accuracy of the models for the prediction of GOR.

This study highlights the significant role of data-driven methods in accurately modeling solar-driven DCMD systems. To enhance the robustness of these models, it is highly recommended for future studies to apply data-driven models with larger datasets that incorporate more gradual and continuous variations of manipulated inputs.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Razeghi, M., Hajinezhad, A., Naseri, A., Noorollahi, Y., Farhan Moosavian, S., 2023. Multi-criteria decision-making for selecting a solar farm location to supply energy to reverse osmosis devices and produce freshwater using GIS in Iran. Sol. Energy 253, 501–514.

[2] Rahimpour, M.R., Kazerooni, N.M., Parhoudeh, M., 2019. Chapter 8 - Water treatment by renewable energy-driven membrane distillation. In: Basile, A., Cassano, A., Figoli, A. (Eds.), Current Trends and Future Developments on (Bio-) Membranes. Elsevier, pp. 179–211.

[3] Chen, L., Xu, P., Wang, H., 2020. Interplay of the factors affecting water flux and salt rejection in membrane distillation: A state-of-the-art critical review. Water 12, 2841.

[4] Qtaishat, M.R., Banat, F., 2013. Desalination by solar powered membrane distillation systems. Desalination 308, 186–197. https://doi.org/10.1016/j.desal.2012.01.021.

[5] Ghaffour, N., Soukane, S., Lee, J.G., Kim, Y., Alpatova, A., 2019. Membrane distillation hybrids for water production and energy efficiency enhancement: A critical review. Appl. Energy 254, 113698. https://doi.org/10.1016/j.apenergy.2019.113698.

[6] Soomro, M.I., Kim, W.-S., 2018. Parabolic-trough plant integrated with direct-contact membrane distillation system: Concept, simulation, performance, and economic evaluation. Sol. Energy 173, 348–361. https://doi.org/10.1016/j.solener.2018.07.086.

[7] Miladi, R., Frikha, N., Kheiri, A., Gabsi, S., 2019. Energetic performance analysis of seawater desalination with a solar membrane distillation. Energ. Conver. Manage. 185, 143–154. https://doi.org/10.1016/j.enconman.2019.02.011.

[8] González, D., Amigo, J., Suárez, F., 2017. Membrane distillation: Perspectives for sustainable and improved desalination. Renew. Sustain. Energy Rev. 80, 238–259. https://doi.org/10.1016/j.rser.2017.05.078.

[9] Zhao, Q., Zhang, H., Hu, Z., Hou, S., 2020. A solar driven hybrid photovoltaic module/direct contact membrane distillation system for electricity generation and water desalination. Energ. Conver. Manage. 221, 113146 https://doi.org/10.1016/j.enconman.2020.113146.

[10] Shafieian, A., Khiadani, M., 2019. A novel solar-driven direct contact membrane-based water desalination system. Energ. Conver. Manage. 199, 112055 https://doi.org/10.1016/j.enconman.2019.112055.

[11] Bamasag, A., Alqahtani, T., Sinha, S., Ghaffour, N., Phelan, P., 2020. Experimental investigation of a solar-heated direct contact membrane distillation system using evacuated tube collectors. Desalination 487, 114497. https://doi.org/10.1016/j.desal.2020.114497.

[12] Kabeel, A.E., Abdelgaied, M., El-Said, E.M.S., 2017. Study of a solar-driven membrane distillation system: Evaporative cooling effect on performance enhancement. Renew. Energy 106, 192–200. https://doi.org/10.1016/j.renene.2017.01.030.

[13] Abdelgaied, M., Kabeel, A.E., Sathyamurthy, R., 2020. Improving the performance of solar powered membrane distillation systems using the thermal energy storage mediums and the evaporative cooler. Renew. Energy 157, 1046–1052. https://doi.org/10.1016/j.renene.2020.05.123.

[14] Hejazi, M.-A.-A., Bamaga, O.A., Al-Beirutty, M.H., Gzara, L., Abulkhair, H., 2019. Effect of intermittent operation on performance of a solar-powered membrane distillation system. Sep. Purif. Technol. 220, 300–308. https://doi.org/10.1016/j.seppur.2019.03.055.

[15] Bindels, M., Medaer, B., Gebrehiwot, M., Nelemans, B., 2021. Approximation of dynamic membrane distillation processes applied to concentration of aqueous sucrose solutions. Desalination 503, 114951. https://doi.org/10.1016/j.desal.2021.114951.

[16] Duong, H.C., Xia, L., Ma, Z., Cooper, P., Ela, W., Nghiem, L.D., 2017. Assessing the performance of solar thermal driven membrane distillation for seawater desalination by computer simulation. J. Membr. Sci. 542, 133–142. https://doi.org/10.1016/j.memsci.2017.08.007.

[17] Chang, H., Wang, G.-B., Chen, Y.-H., Li, C.-C., Chang, C.-L., 2010. Modeling and optimization of a solar driven membrane distillation desalination system. Renew. Energy 35, 2714–2722. https://doi.org/10.1016/j.renene.2010.04.020.

[18] Eleiwi, F., Ghaffour, N., Alsaadi, A.S., Francis, L., Laleg-Kirati, T.M., 2016. Dynamic modeling and experimental validation for direct contact membrane distillation (DCMD) process. Desalination 384, 1–11. https://doi.org/10.1016/j.desal.2016.01.004.

[19] Shim, W.G., He, K., Gray, S., Moon, I.S., 2015. Solar energy assisted direct contact membrane distillation (DCMD) process for seawater desalination. Sep. Purif. Technol. 143, 94–104. https://doi.org/10.1016/j.seppur.2015.01.028.

[20] Karam, A.M., Alsaadi, A.S., Ghaffour, N., Laleg-Kirati, T.M., 2017. Analysis of direct contact membrane distillation based on a lumped-parameter dynamic predictive model. Desalination 402, 50–61.

[21] Ali, E., Saleh, J., Orfi, J., Najib, A., 2020. Developing and validating linear dynamic models for direct contact membrane distillation during start-up over wide operating conditions. Comput. Chem. Eng. 134, 106678 https://doi.org/10.1016/j.compchemeng.2019.106678.

[22] Ali, E., Orfi, J., Najib, A., 2020. Nonlinear dynamic modeling and validation of a direct contact membrane distillation for water desalination. Appl. Therm. Eng. 179, 115719 https://doi.org/10.1016/j.applthermaleng.2020.115719.

[23] Alnaqi, A.A., Moayedi, H., Shahsavar, A., Nguyen, T.K., 2019. Prediction of energetic performance of a building integrated photovoltaic/thermal system thorough artificial neural network and hybrid particle swarm optimization models. Energ. Conver. Manage. 183, 137–148. https://doi.org/10.1016/j.enconman.2019.01.005.

[24] Ahmad, M.W., Reynolds, J., Rezgui, Y., 2018. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. J. Clean. Prod. 203, 810–821. https://doi.org/10.1016/j.jclepro.2018.08.207.

[25] Unterberger, V., Lichtenegger, K., Kaisermayer, V., Gölles, M., Horn, M., 2021. An adaptive short-term forecasting method for the energy yield of flat-plate solar collector systems. Appl. Energy 293, 116891. https://doi.org/10.1016/j.apenergy.2021.116891.

[26] Correa-Jullian, C., Cardemil, J.M., López Droguett, E., Behzad, M., 2020. Assessment of Deep Learning techniques for Prognosis of solar thermal systems. Renew. Energy 145, 2178–2191. https://doi.org/10.1016/j.renene.2019.07.100.

[27] Behzadi, A., Sadrizadeh, S., 2023. A rule-based energy management strategy for a low-temperature solar/wind-driven heating system optimized by the machine learning-assisted grey wolf approach. Energ. Conver. Manage. 277, 116590 https://doi.org/10.1016/j.enconman.2022.116590.

[28] Heidari, A., Khovalyg, D., 2020. Short-term energy use prediction of solar-assisted water heating system: Application case of combined attention-based LSTM and time-series decomposition. Sol. Energy 207, 626–639. https://doi.org/10.1016/j.solener.2020.07.008.

[29] Elsheikh, A.H., Sharshir, S.W., Abd Elaziz, M., Kabeel, A.E., Guilan, W., Haiou, Z., 2019. Modeling of solar energy systems using artificial neural network: A comprehensive review. Sol. Energy 180, 622–639. https://doi.org/10.1016/j.solener.2019.01.037.

[30] Zarei, T., Behyad, R., 2019. Predicting the water production of a solar seawater greenhouse desalination unit using multi-layer perceptron model. Sol. Energy 177, 595–603. https://doi.org/10.1016/j.solener.2018.11.059.

[31] Elsheikh, A.H., Panchal, H., Ahmadein, M., Mosleh, A.O., Sadasivuni, K.K., Alsaleh, N.A., 2021. Productivity forecasting of solar distiller integrated with evacuated tubes and external condenser using artificial intelligence model and moth-flame optimizer. Case Stud. Therm. Eng. 28, 101671 https://doi.org/10.1016/j.csite.2021.101671.

[32] Elsheikh, A.H., Katekar, V.P., Muskens, O.L., Deshmukh, S.S., Elaziz, M.A., Dabour, S.M., 2021. Utilization of LSTM neural network for water production forecasting of a stepped solar still with a corrugated absorber plate. Process Saf. Environ. Prot. 148, 273–282. https://doi.org/10.1016/j.psep.2020.09.068.

[33] Salem, H., Kabeel, A.E., El-Said, E.M.S., Elzeki, O.M., 2022. Predictive modelling for solar power-driven hybrid desalination system using artificial neural network regression with Adam optimization. Desalination 522, 115411. https://doi.org/10.1016/j.desal.2021.115411.

[34] Porrazzo, R., Cipollina, A., Galluzzo, M., Micale, G., 2013. A neural network-based optimizing control system for a seawater-desalination solar-powered membrane distillation unit. Comput. Chem. Eng. 54, 79–96. https://doi.org/10.1016/j.compchemeng.2013.03.015.

[35] Behnam, P., Faegh, M., Khiadani, M., 2022. A review on state-of-the-art applications of data-driven methods in desalination systems. Desalination 532, 115744. https://doi.org/10.1016/j.desal.2022.115744.

[36] Gil, J.D., Ruiz-Aguirre, A., Roca, L., Zaragoza, G., Berenguel, M., 2018. Prediction models to analyse the performance of a commercial-scale membrane distillation unit for desalting brines from RO plants. Desalination 445, 15–28. https://doi.org/10.1016/j.desal.2018.07.022.

[37] Khayet, M., Cojocaru, C., 2012. Artificial neural network modeling and optimization of desalination by air gap membrane distillation. Sep. Purif. Technol. 86, 171–182. https://doi.org/10.1016/j.seppur.2011.11.001.

[38] Cao, W., Liu, Q., Wang, Y., Mujtaba, I.M., 2016. Modeling and simulation of VMD desalination process by ANN. Comput. Chem. Eng. 84, 96–103. https://doi.org/10.1016/j.compchemeng.2015.08.019.

[39] Ruiz-Aguirre, A., Andrés-Mañas, J.A., Fernández-Sevilla, J.M., Zaragoza, G., 2017. Modeling and optimization of a commercial permeate gap spiral wound membrane distillation module for seawater desalination. Desalination 419, 160–168. https://doi.org/10.1016/j.desal.2017.06.019.

[40] Moustafa, E.B., Hammad, A.H., Elsheikh, A.H., 2022. A new optimized artificial neural network model to predict thermal efficiency and water yield of tubular solar still. Case Stud. Therm. Eng. 30, 101750 https://doi.org/10.1016/j.csite.2021.101750.

[41] Mäkelä, M., 2017. Experimental design and response surface methodology in energy applications: A tutorial review. Energ. Conver. Manage. 151, 630–640. https://doi.org/10.1016/j.enconman.2017.09.021.

[42] Elzahaby, A.M., Kabeel, A.E., Bassuoni, M.M., Elbar, A.R.A., 2016. Direct contact membrane water distillation assisted with solar energy. Energ. Conver. Manage. 110, 397–406. https://doi.org/10.1016/j.enconman.2015.12.046.

[43] Rastegar, S., Kargarsharifabad, H., Rahbar, N., Shafii, M.B., 2020. Distilled water production with combination of solar still and thermosyphon heat pipe heat exchanger coupled with indirect water bath heater – Experimental study and thermoeconomic analysis. Appl. Therm. Eng. 176, 115437 https://doi.org/10.1016/j.applthermaleng.2020.115437.

[44] Sun, Y., Haghighat, F., Fung, B.C.M., 2020. A review of the-state-of-the-art in data-driven approaches for building energy prediction. Energ. Buildings 221, 110022. https://doi.org/10.1016/j.enbuild.2020.110022.

[45] Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12 (2011) 2825-2830.

[47] F. Chollet, Keras: The python deep learning library, Astrophysics source code library, (2018) ascl: 1806.1022.

[48] Faegh, M., Behnam, P., Shafii, M.B., Khiadani, M., 2021. Development of artificial neural networks for performance prediction of a heat pump assisted humidification-dehumidification desalination system. Desalination 508, 115052.

[49] Nguyen, N.Q., Bui, L.D., Doan, B.V., Sanseverino, E.R., Cara, D.D., Nguyen, Q.D., 2021. A new method for forecasting energy output of a large-scale solar power plant based on long short-term memory networks a case study in Vietnam. Electr. Pow. Syst. Res. 199, 107427 https://doi.org/10.1016/j.epsr.2021.107427.