

Data-driven performance assessment and forecasting of solar water heaters using artificial intelligence techniques

M Magdy^{1,*}, R A Ibrahim¹, M Mansouri², and N E Zakzouk¹

¹ Arab Academy for Science, Technology & Maritime Transport (AASTMT), Egypt

² University of Stavanger, Norway

* mohammedmagdey99@gmail.com

Abstract. Solar water heating systems (SWHS) provide a convenient means of reducing energy consumption for both residential and commercial applications, space conditioning and domestic hot water production. Performance assessment and forecasting of solar heaters are essential for evaluating energy conversion efficiency, identifying performance degradation and predicting future thermal energy availability and required maintenance. However, conventional physics-based assessment may not fully capture the nonlinear and transient behaviour of real solar heaters, motivating the development of data-driven forecasting models which outweigh the former regarding accuracy, flexibility, computational efficiency, and reduced dependence on physical parameters. In this work, a data-driven model for an SWHS is developed, trained and evaluated based on a MATLAB/Simulink model and selected physical response trends. The model was first used to generate 1097 operating points covering a wide range of environmental and operational conditions. These operating points were defined by eight input variables, comprising three meteorological variables and five design and operating variables with collector efficiency selected as the prediction target. Second, several models were created to forecast collector efficiency including, Multiple Linear Regression (MLR), Polynomial Regression (PR), and machine learning (ML) techniques including Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT), and Extreme Gradient Boosting (XGBoost), in addition to an Artificial Neural Networks (ANN) model. The performance of these methods was analysed using different performance metrics. The results indicated the highest forecasting ability for SVR, followed by ANN, demonstrating the superiority of non-linear ML models in capturing SWH performance characteristics compared to conventional regression models.

1. Introduction

Energy is one of the key elements to modern living standards, economic development and environmental sustainability. Nevertheless, the increasing residential and industrial energy demand continues to drive global consumption, which is still predominantly supplied by fossil fuels, resulting in environmental and climate change [1,2]. A substantial proportion of energy use in residential and industrial sectors is associated with space conditioning and domestic hot-water production; therefore, improving the efficiency and renewable contribution of these applications is essential for reducing their environmental impact [3,4]. Among renewable resources, solar energy is particularly attractive as it can be converted into either electricity or useful heat [5–7]. Accordingly, integrating solar thermal collectors into water- and space-heating systems is a practical approach to reducing grid electricity dependence and fossil-based energy while lowering associated CO₂ emissions [8].

A typical application is the solar water heating system (SWHS) which converts incident solar radiation into heat, transfers it to a circulating fluid, and stores the energy as hot water [9,10]. In an indirect double-loop configuration, an antifreeze solution circulates between the collector field and a heat exchanger while remaining separated from the potable-water circuit [9,11]. Powering the circulation pumps from a photovoltaic (PV) array can also make the system independent of the electrical grid [11]. Although solar water-heating is a mature technology, its performance is strongly affected by

operating and climatic conditions. Flat-plate collector efficiency decreases approximately linearly with the ratio of the mean collector-fluid-to-ambient temperature difference to incident irradiance [3,12]; therefore, efficiency declines as the required water temperature increases or solar irradiance decreases. Other collector quantities, such as the plate temperature and temperature-dependent heat-loss coefficients, additionally require iterative solution within the physics-based model. These coupled dependencies make system performance difficult to estimate accurately over a wide range of conditions using simple analytical approximations alone, thus physics-based simulation is considered as a viable option to evaluate system behaviour and support design assessment [3,13]

Accordingly, accurate performance assessment and forecasting of SWHs are essential for evaluating energy conversion efficiency, predicting future thermal energy availability, identifying performance degradation and supporting predictive maintenance and intelligent control. It further enables better utilization of solar resources and improves the reliability, economic viability, and overall sustainability.

Although detailed physics-based models provide accurate system predictions, they can be computationally demanding, particularly when several evaluations are required for parametric analysis or optimization. Data-driven models, on the other hand, learn the actual behaviour from operational data, allowing the capture of nonlinearities and real-world effects that may be difficult, expensive, or impossible to represent accurately in a physics-based forecasting model. Recent studies have applied neural networks, machine learning and other data-driven methods to predict the thermal performance of solar collectors and water-heating systems under different climatic, operating, and design conditions [14,15]. More recently, design-oriented approaches have emerged, including physics-informed hyper-optimization of flat-plate collectors [16] and thermodynamics-informed neural networks for predicting suitable collector-design conditions [17]. However, most available studies focus on predicting the performance of a predefined collector configuration rather than modelling the complete indirect solar water-heating system or systematically comparing several regression families under a common preprocessing, tuning, and evaluation protocol.

In this study, a comprehensive design dataset for a SWHS is generated from a MATLAB/Simulink model, and data-driven surrogate models are then developed and evaluated to assess the performance of the considered SWHS and forecast its collector efficiency. A detailed MATLAB/Simulink model of an indirect flat-plate SWHS with PV-driven circulation is used as the reference model for generating system-response data under different operating and environmental conditions, and the study proceeds in the following steps:

- The model is used to generate 1097 design points from eight meteorological, operating, and design inputs and twelve corresponding system outputs.
- Before using the generated data for forecasting, the Simulink implementation is verified against the expected collector-efficiency characteristic and selected physical response trends.
- The resulting dataset is then characterized statistically, and seven data-driven models - including Multiple Linear Regression (MLR), Polynomial Regression (PR), Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT), and Extreme Gradient Boosting (XGBoost), in addition to an Artificial Neural Networks (ANN) model - are trained and evaluated for predicting collector efficiency.
- The forecasting results were compared with those obtained from the MATLAB/Simulink reference model to evaluate the accuracy, robustness, and generalization capability of the proposed data-driven approach.

2. Solar water-heating system under consideration

Modelling and simulation of the considered solar water-heating system is detailed in the following subsections.

2.1. System configuration and modelling

In this study, a dataset, for a SWHS, is generated based on a MATLAB/Simulink model and subsequently data-driven surrogate models are developed for predicting the heater collector efficiency η_{col} . The considered SWHS consists of an indirect (double-loop) solar water-heating system with PV-

driven circulation, shown schematically in Figure 1, and is implemented in MATLAB/Simulink using the Renewable Energy Desalination Systems (REDS) library [18]. In the primary loop, an array of flat-plate collectors heats a 30% ethylene-glycol mixture, which is circulated to a counter-flow heat exchanger of effectiveness ϵ_{hex} , which provides freeze and corrosion protection and prevents scale deposition inside the absorber tubes, at the cost of the additional temperature drop across the heat exchanger. The secondary loop transfers the heat to potable water held in an insulated storage tank of volume V , which delivers water to the user at the required temperature $T_{stg,o}$ while make-up water enters the tank at $T_{stg,i}$ to replace the consumed volume. Flow is circulated by two pumps, P , supplied by a stand-alone PV array with battery storage, so that the plant requires no grid connection.

During simulation and under specified operating and design conditions, the model determines system performance and sizing outputs that include, collector efficiency, collector and total field areas, number of collectors, heat-exchanger area, mean plate temperature, heat-loss coefficients, and pumping power. Among these outputs, collector efficiency η_{col} is selected as the prediction targets in this study as it is the main component responsible for converting incident solar radiation into useful heat. Consequently, accurate prediction can support system performance assessment across changing weather and operating conditions [19]. All quantities in the proceeding sections refer to a system's daily operating horizon (OH), of 8 hours.

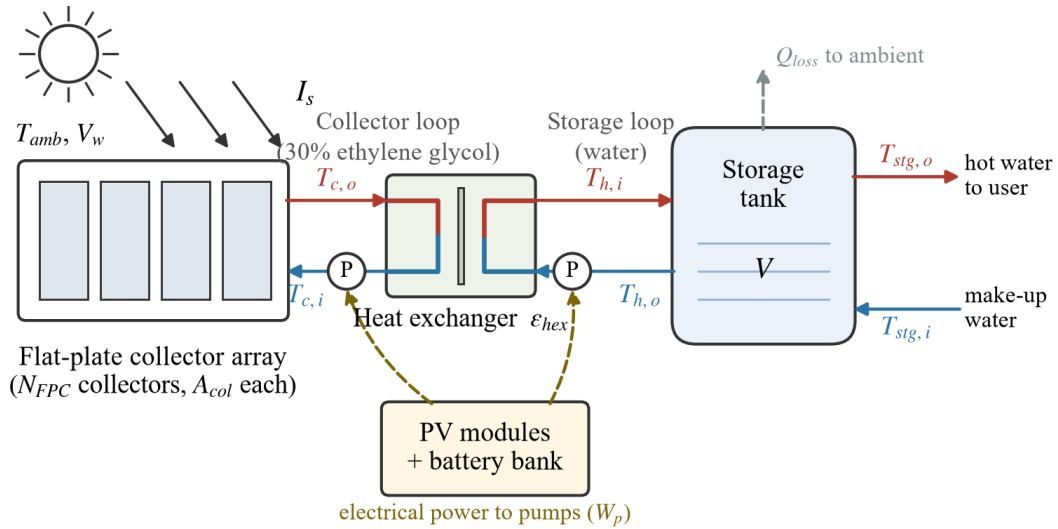


Figure 1. Schematic of the considered indirect flat-plate collector SWHS with PV-driven pumping

2.2. Storage tank subsystem

The average thermal load Q_{stg} required by the storage tank over the daily operating horizon is calculated from the volume of water raised from the make-up temperature, $T_{stg,i}$, to the delivery temperature, $T_{stg,o}$, over the OH (1)[20]:

$$Q_{stg} = \frac{\rho_w V c_{p,w} (T_{stg,o} - T_{stg,i})}{3600 \times OH} \quad (1)$$

where ρ_w and $c_{p,w}$ are the density and specific heat capacity of water and OH is the daily operating horizon. The temperature of the water leaving the tank towards the heat exchanger, $T_{h,o}$ is then obtained from the energy balance on the storage-tank side as (2):

$$T_{h,o} = T_{h,i} + T_{stg,i} - T_{stg,o} \quad (2)$$

where $T_{h,i}$ is the temperature of the water returning from the heat exchanger to the tank.

The model also evaluates the storage-side mass flow rate, the tank geometry and the thermal loss to the surroundings through a series conduction-convection resistance with an external film coefficient obtained from a forced-convection correlation for crossflow over a cylinder, together with the tank-side pumping power. These quantities are only recorded but are not considered as prediction targets in this study; therefore, their governing relations are omitted.

2.3. Heat exchanger subsystem

The exchanger is characterized by its effectiveness, treated here as a design specification rather than an outcome. Applying the effectiveness definition to the glycol side gives the collector-loop temperatures directly as presented in (3-4)[20]:

$$T_{c,o} = T_{h,o} + \frac{T_{h,i} - T_{h,o}}{\varepsilon_{hex}} \quad (3)$$

$$T_{c,i} = T_{c,o} - \varepsilon_{hex}(T_{c,o} - T_{h,o}) \quad (4)$$

where $T_{c,o}$ and $T_{c,i}$ are the collector-loop outlet and inlet temperatures, respectively, while $T_{h,i}$ and $T_{h,o}$ are the water-side outlet and inlet temperatures of the heat exchanger, respectively.

The heat-exchanger surface area is also calculated from the number of transfer units for a counter-flow configuration and an overall heat-transfer coefficient expressed as a cubic function of the mean glycol temperature. Since the exchanger area is recorded as an output but is not among the prediction targets considered in this study, the corresponding relations are not considered.

2.4. Flat-plate collector subsystem

The instantaneous collector efficiency η_{col} is described by the linear loss characteristic of the commercial collector given by (5)[20]:

$$\eta_{col} = 0.78 - 3.4 \frac{T_{m,c} - T_{amb}}{I_s}, \quad T_{m,c} = \frac{T_{c,o} + T_{c,i}}{2} \quad (5)$$

where $T_{m,c}$ is the mean collector-fluid temperature, T_{amb} is the ambient temperature, and I_s is the equivalent peak solar irradiance. Equation (5) is the linear form of the collector efficiency curve prescribed by the collector test standard [21,22] and used throughout the solar thermal literature. The constants 0.78 and 3.4 W m⁻² K⁻¹ represent the optical efficiency and overall collector loss coefficient, respectively.

The useful thermal load carried by the collector loop is calculated as (6)

$$Q_{thl} = \frac{\rho_{eg} V_{loop} c_{p,eg} (T_{c,o} - T_{c,i})}{3600 \times OH} \quad (6)$$

in which ρ_{eg} , $c_{p,eg}$ and V_{loop} are the density, specific heat capacity and circulating volume of the glycol mixture respectively. The required collector aperture area is then obtained from (7):

$$A_{col} = \frac{1000 Q_{thl}}{\eta_{col} I_s} \quad (7)$$

where A_{col} is the required collector aperture area.

Following the collector performance and sizing relations presented above, the model further evaluates the absorber heat-loss characteristics needed to determine the plate temperature and associated loss coefficients. The absorber top-loss coefficient is calculated using the Klein correlation for a flat-plate collector with N_G glazings. The wind heat transfer coefficient and the auxiliary groups are given in (8) as:

$$h_w = 5.7 + 3.8V_w, \quad f = (1 + 0.089h_w - 0.1166h_w\varepsilon_p)(1 + 0.07866N_G) \quad (8)$$

where V_w is the wind speed, h_w is the external convective heat-transfer coefficient, ε_p is the absorber-plate emittance, and N_G is the number of glass covers. f is a dimensionless auxiliary factor in the Klein correlation accounting for the effects of wind convection, absorber-plate emittance, and the number of

glass covers. In the following relations, absolute temperatures are required; therefore, $T_{mp,K} = T_{mp} + 273.15$ and $T_{amb,K} = T_{amb} + 273.15$. The temperature difference is unchanged $T_{mp} - T_{amb}$ whether expressed in °C or K. The remaining auxiliary parameters C and e , with collector tilt angle β can be represented by (9):

$$C = 520(1 - 5.1 \times 10^{-5} \beta^2), \quad e = 0.43(1 - 100/T_{mp,K}) \quad (9)$$

The top-loss coefficient U_l is given by (10):

$$U_l = \left[\frac{N_G}{\frac{C}{T_{mp,K}} \left(\frac{T_{mp} - T_{amb}}{N_G + f} \right)^e} + \frac{1}{h_w} \right]^{-1} + \frac{\sigma(T_{mp,K} + T_{amb,K})(T_{mp,K}^2 + T_{amb,K}^2)}{(\varepsilon_p + 0.00591 N_G h_w)^{-1} + \frac{2N_G + f - 1 + 0.133\varepsilon_p}{\varepsilon_g} - N_G} \quad (10)$$

in which σ is the Stefan-Boltzmann constant ε_p and ε_g are absorber plate and of the glass cover emittances, respectively. The overall coefficient adds to the edge and back losses (11):

$$U_{total} = U_l + U_e + U_b \quad (11)$$

where U_l , U_e , and U_b denote the top, edge, and back loss coefficients, respectively.

The mean absorber-plate temperature T_{mp} is subsequently obtained from the collector energy balance (12):

$$T_{mp} = \frac{A_{col} I_s / 1000 - Q_{thl}}{A_{col} U_{total} / 1000} + T_{amb} \quad (12)$$

where U_e and U_b are the edge- and back-loss coefficients, respectively.

From the plate temperature, the model further evaluates the heat-removal factor, the collector efficiency factor, the fin efficiency and the flow factor, and the PV subsystem sizes the module array and battery bank to cover the combined pumping demand over the OH. These quantities are recorded in the dataset generated but are not among the prediction targets considered in this study.

3. Simulink implementation and dataset generation

The subsystems modeled earlier are implemented in MATLAB R2017a as embedded MATLAB function blocks inside a single Simulink model [18], connected by tagged signal routing, with lookup tables supplying the water, glycol and air property correlations and six memory blocks closing the plate-temperature loop discussed above. [Figure 2 shows the top level of the Simulink model with its four subsystems, together with the flat-plate collector subsystem, in which the embedded MATLAB function block evaluates \(5\)–\(12\) from the constant parameter blocks and the temperatures routed from the heat-exchanger subsystem.](#)

(a)



(b)

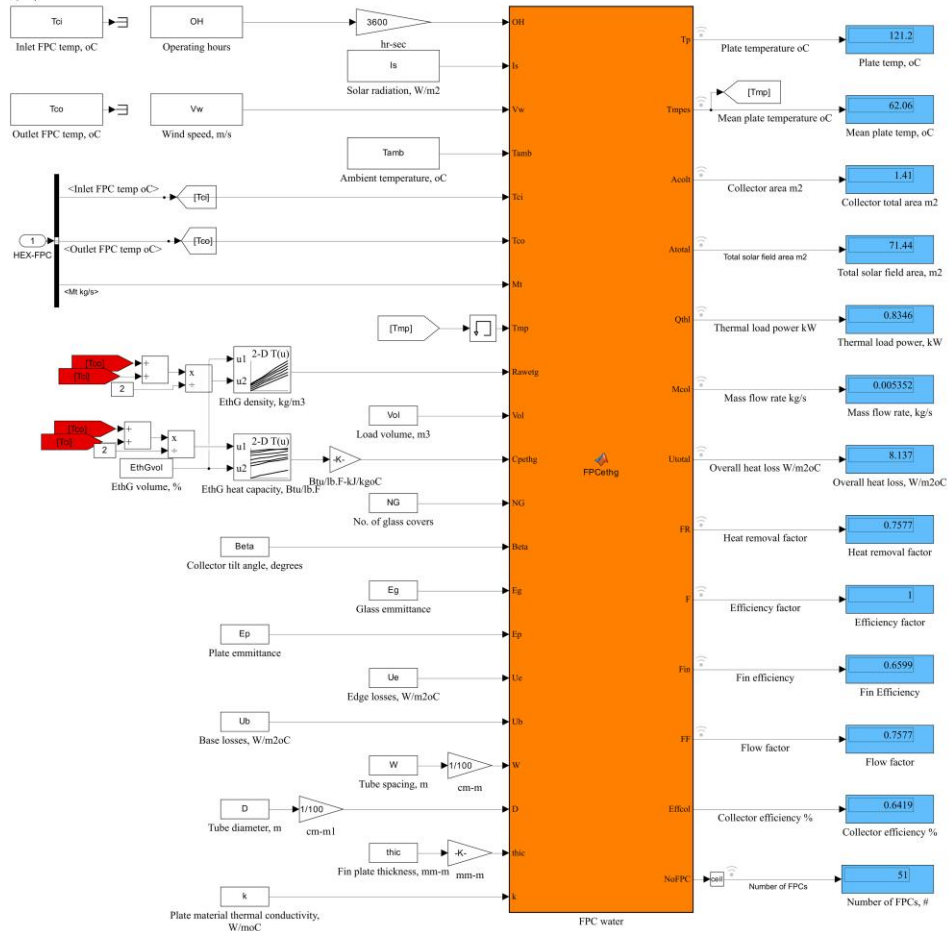


Figure 2. Simulink implementation of the considered SWHS: (a) top level with the collector, heat-exchanger, storage-tank and PV subsystems; (b) flat-plate collector subsystem, with the displayed outputs at the reference design condition.

3.1. Considered SWHS parameters

The parameters of the collector, exchanger and tank subsystems are held constant throughout the study and are listed in Table 1; corresponding to a single-glazed copper-absorber collector. The Simulink model receives eight input variables comprising three meteorological conditions namely, solar

irradiance, ambient temperature, and wind speed, and five system design and operating specifications, including storage-tank volume, heat-exchanger effectiveness, make-up water temperature, required delivery temperature, and heat-exchanger return temperature. Based on these inputs, the model determines the corresponding system performance and sizing outputs.

As shown in Table 2, the meteorological data were taken from the NASA POWER daily archive [23] for a site at latitude 30.85° N, longitude 28.95° E on the Egyptian Mediterranean coast, covering the period between 1 January 2023 to 1 January 2026, yielding 1097 consecutive daily records. The remaining inputs are design specifications, sampled independently over the ranges given in Table 2, except that the make-up temperature was coupled to the ambient temperature and the heat-exchanger return temperature to the delivery temperature, so that only physically consistent combinations enter the dataset. During simulation, the model determines twelve system performance and sizing outputs under the specified meteorological and operating, among which, η_{col} was selected as the prediction target in this study.

Table 1. Parameters of the collector, exchanger and tank subsystems held fixed during dataset generation.

Subsystem	Parameter	Symbol	Value	Unit
Collector	Operating horizon	OH	8	h
	Collector-loop fluid volume	V_{loop}	0.15	m ³
	Ethylene glycol fraction	-	30	%
	Number of glass covers	N_G	1	-
	Tilt angle	β	30	deg
	Glass cover emittance	ε_g	0.88	-
	Absorber plate emittance	ε_p	0.95	-
	Edge loss coefficient	U_e	0.5	W m ⁻² K ⁻¹
	Back loss coefficient	U_b	0.3	W m ⁻² K ⁻¹
	Tube spacing	W	0.10	m
	Tube diameter	D	0.01	m
	Fin thickness	δ	0.5	mm
	Plate conductivity (copper)	k	385	W m ⁻¹ K ⁻¹
Heat exchanger	Pump efficiency	η_p	0.7	-
	Exchanger diameter	D_{hex}	0.02	m
Storage tank	Tank diameter	D_t	2.0	m
	Insulation thickness	t_{ins}	0.05	m
	Insulation conductivity	k_{ins}	0.04	W m ⁻¹ K ⁻¹

Table 2. The eight model inputs, their source or basis, and their range over the 1097 design points.

Input	Symbol	Unit	Source or basis	Min	Max	Mean	SD
Solar irradiance	I_s	W m ⁻²	NASA POWER	244.1	1309.4	886.6	278.6
Ambient temperature	T_{amb}	°C	NASA POWER, T2M	11.33	31.05	21.58	4.75
Wind speed	V_w	m s ⁻¹	NASA POWER, WS10M	1.13	10.94	4.76	1.52
Storage tank volume	V	m ³	Institutional store size	2.00	24.85	8.62	5.15
HEX effectiveness	ε_{hex}	-	Typical counter-flow solar loop	0.600	0.899	0.748	0.087
Make-up temperature	$T_{stg,i}$	°C	Tracks the seasonal mains temperature	8.18	20.97	14.83	2.68
Delivery temperature	$T_{stg,o}$	°C	Brackets the 60 °C service standard	45.01	64.96	55.04	5.74
HEX return temperature	$T_{h,i}$	°C	Set above the delivery temperature	50.44	79.50	65.07	6.37

3.2. Verification of the Simulink model

In order to verify the performance of the simulated Simulink output model with the physics-based mathematical model, it is necessary to check that the collector model reproduces the expected physical relationship represented in (5). The collector-efficiency verification plot found in Figure 3 was generated by varying solar irradiance and ambient temperature over representative values within their investigated ranges, while the remaining model inputs were maintained at a fixed reference condition. **Ten irradiance levels, from 300 to 1200 W m⁻² in steps of 100 W m⁻², were combined with four ambient-temperature values of 15, 20, 25 and 30 °C, resulting in 40 operating points.** For each combination, the Simulink model was executed and the corresponding collector efficiency, η_{col} , and mean collector-fluid temperature, $T_{m,c}$, were recorded. The reduced temperature, $\frac{T_{m,c} - T_{amb}}{I_s}$ was then calculated for each operating point and plotted against the simulated collector efficiency. The resulting points were compared with the theoretical linear efficiency characteristic of (5).

As presented in Figure 3, the solid straight line presents the theoretical collector-efficiency characteristic given by (5), while the scattered points correspond to the 40 operating conditions obtained from the Simulink model. The close agreement between the simulated points and the theoretical line confirms that the collector subsystem reproduces the expected efficiency relationship, **with the simulated efficiencies ranging from about 28% to 70%.**

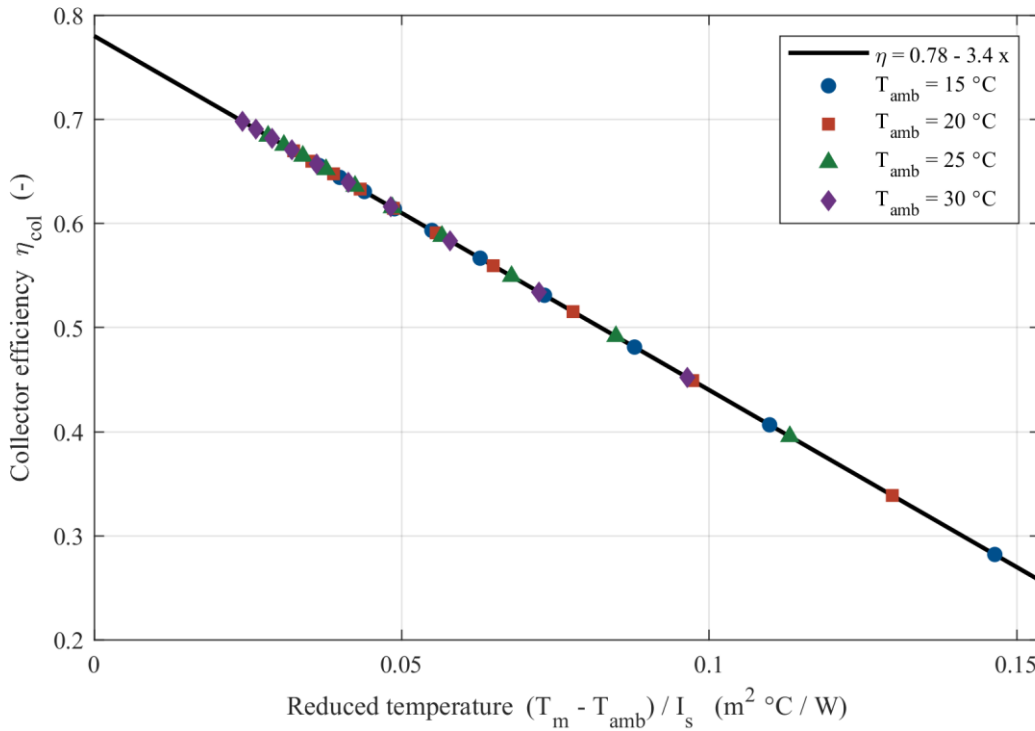


Figure 3. η_{col} against reduced temperature for different irradiance and ambient temperature combinations

3.3. Structure of the inputs-output mapping

The input-output relationship for η_{col} is inherently nonlinear i.e. although the η_{col} relation can be expressed explicitly, it involves coupled nonlinear variables such as solar irradiance, heat-exchanger effectiveness, and the different temperature variables. These relationships make direct evaluation of η_{col} over a wide range of operating conditions increasingly cumbersome, particularly when repeated calculations are required. This motivates the use of machine-learning surrogate models to provide fast and accurate predictions of η_{col} over the considered operating range while preserving the underlying physical relationships of the system.

It is worth noting that by substituting (2) -(4) into (5), η_{col} can be expressed explicitly as a function of six of the eight model inputs, where V_w and V do not appear in this relation, indicating that they have no direct physical influence on η_{col} . This is clearly shown in Figure 4 which plots η_{col} against different system inputs including solar irradiance, ambient temperature, wind speed, tank volume, heat exchanger effectiveness, make-up water temperature, delivered water temperature and heat exchanger hot-side inlet temperature. Wind speed and storage-tank volume have essentially no direct effect on η_{col} . In contrast, solar irradiance, heat-exchanger effectiveness, and ambient temperature have positive effects, with solar irradiance producing the largest variation. The heat-exchanger hot-side inlet temperature has the strongest negative effect, while the make-up water temperature causes only a slight decrease in η_{col} .

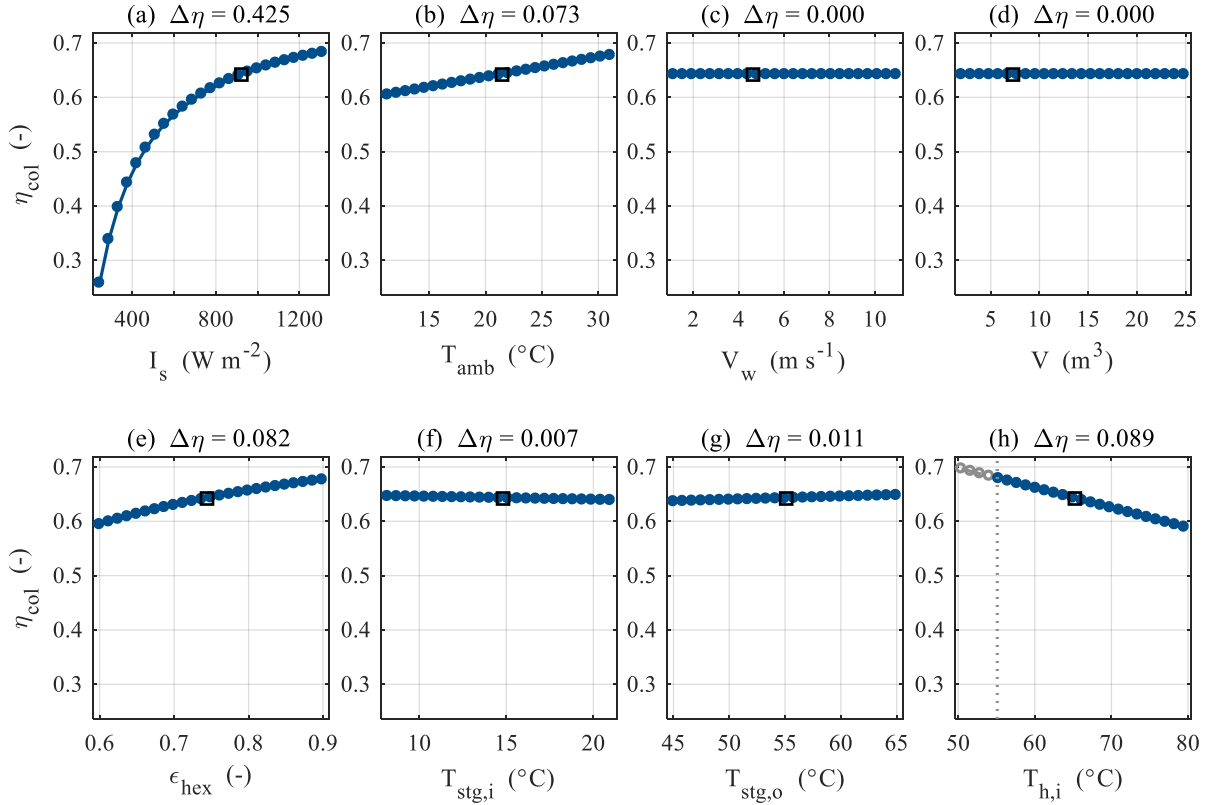


Figure 4. η_{col} against each system input, with other seven held at reference design value.

4. Generated dataset distribution

In this section, the statistical distributions of the generated dataset are presented, first for the eight model inputs and then for the prediction target.

4.1. Distribution of the inputs

The marginal distribution of each of the eight inputs is shown in Figure 5, where the solid line marks the median and the dotted lines mark the first and third quartiles. The three meteorological inputs inherit the shape of the local climate: the irradiance is broad and slightly left-skewed (skewness -0.18) with a pronounced seasonal spread as presented in Figure 5(a), the ambient temperature is close to symmetric over 11.3 to 31.1°C in Figure 5(b), and the wind speed is mildly right-skewed (0.55) with 13 values beyond the interquartile fence as shown in Figure 5(c). The tank volume is deliberately right-skewed (1.09) so that small domestic-scale and large institutional-scale tanks are both represented in Figure 5(d). Finally, the four temperature and effectiveness specifications are close to uniform by construction as shown in Figure 5(e-h).

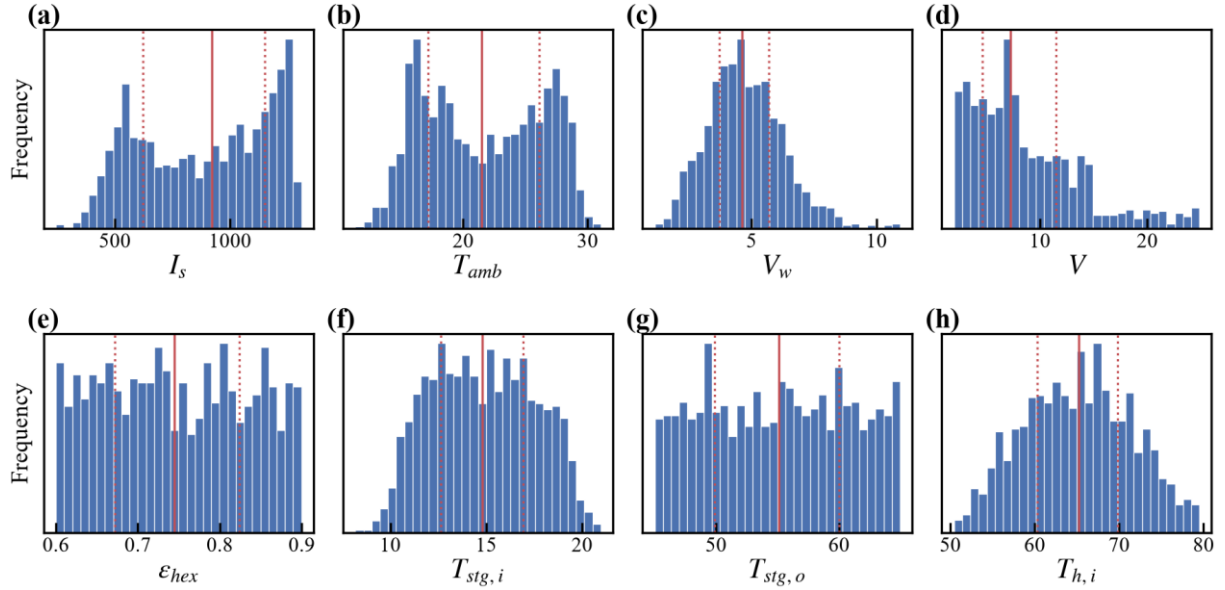


Figure 5. Marginal distributions of the model inputs over the 1097 design points: (a) I_s , (b) T_{amb} , (c) V_w , (d) V , (e) ϵ_{hex} , (f) $T_{stg,i}$, (g) $T_{stg,o}$, and (h) $T_{h,i}$.

4.2. Distribution of the targets

Figure 6 shows distribution of η_{col} over the 1097 simulated design points which is noticed to be left-skewed (skewness = -1.42, excess kurtosis = 2.74), with a dense concentration near 0.65 and a lower tail associated with low-irradiance conditions. The minimum value of 0.015 occurs under low irradiance when the mean glycol temperature is considerably higher than the ambient temperature.

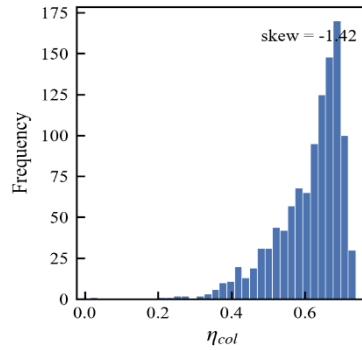


Figure 6. Distribution of the output η_{col}

5. Development and selection of data-driven forecast models

Seven supervised regression algorithms were evaluated namely: MLR, PR, radial-basis-function SVR, DT, RF, XGBoost, and ANN. These algorithms represent linear, nonlinear, kernel-based, tree-based, ensemble, and neural-network modelling approaches for a fair comparison of forecast performance among different regression families. All models used the same eight meteorological, operating, and design variables as inputs, with collector efficiency considered the primary prediction target. The complete modelling workflow is presented in Figure 7 which summarizes all steps carried out in this work.

The dataset was partitioned into training and testing partitions using an 80:20 ratio, corresponding to 877 and 220 records out of the total of 1097, respectively. Input variables were standardized to zero mean and unit variance for MLR, PR, SVR, and ANN, with the scaling parameters determined using the training partition only for models except tree-based models. Model hyperparameters were optimized through grid search with five-fold cross-validation applied only to the training partition. PR used a fixed

second-order expansion. For SVR, the selected parameters were the regularization parameter $C = 100$, which controls the penalty assigned to prediction errors; the radial-basis-function kernel coefficient $\gamma = 0.05$, which controls the kernel width and model sensitivity to nearby samples; and the error-tolerance margin $\varepsilon = 0.01$, within which prediction errors are not penalized. DT used a maximum depth of 10, a minimum of two samples per leaf, and five samples per split. RF comprised 200 trees with a maximum depth of 16 and a minimum of one sample per leaf. XGBoost used 600 boosting rounds, a maximum depth of 3, a learning rate of 0.05, and row and column subsampling fractions of 0.8. Finally, the selected ANN contained two hidden layers with 128 and 64 neurons with a hyperbolic tangent activation function. All models were implemented using Scikit-learn, except XGBoost, which was implemented using the XGBoost library.

Following hyperparameter selection, each model was retrained on the complete training partition and evaluated on the unseen test data. Predictive performance was assessed using the coefficient of determination R^2 , root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

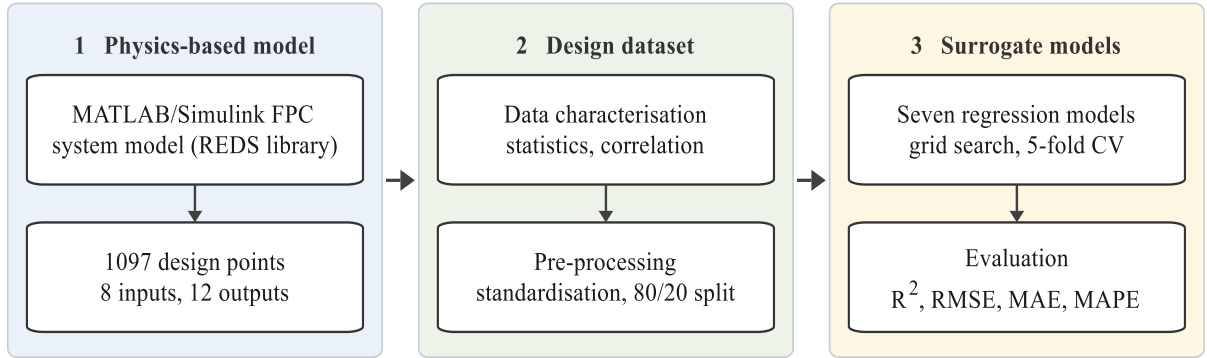


Figure 7. Workflow of the proposed physics-informed forecast framework

6. Results and discussion

The evaluation metrics for the test-set forecast performance of the seven models are given in Table 3 while Figure 8 shows the same result graphically for easier interpretation and comparison between all tested models. SVR presents the most accurate model in every metric, reaching an R^2 of 0.9981, RMSE of 0.0037, MAPE of 0.42% and MAE of 0.00220. The ANN model comes second, with R^2 of 0.9979, RMSE of 0.0038, MAPE of 0.49% and MAE of 0.00269. The second-order PR and RF follow in the third and fourth positions, scoring an R^2 of 0.9867 and 0.9763, and an MAPE of 1.319 and 1.705, respectively. Next, the XGBoost model achieves an R^2 of 0.9743, 1.240 for MAPE, 0.0135 for RMSE, and 0.00631 for MAE. Finally, MLR and single DT come in sixth and seventh position respectively. These results verify the effectiveness of SVR model combining high forecast accuracy, compared to linear regression models, with relatively less implementation complexity compared to ensemble ML and neural networks.

6.1. Predicted points against simulated values

Figure 9 shows parity plots for all 220 test points using the seven considered models. Both SVR and ANN demonstrate the best agreement with the parity line over the entire efficiency range, in line with their highest R^2 values. PR and RF also agree well, although with a bit more scatter, which is larger for RF. XGBoost tracks the parity line closely over most of the range but departs from it at the lowest efficiencies. The DT predictions clearly demonstrate clustering and dispersion, which is in accordance with the piecewise-constant property of tree-based regression. Conversely, MLR shows a clear systematic curvature with overprediction at low and high efficiencies and underprediction over the mid range, indicating low prediction generalizability among all studied models.

Table 3. Evaluation Metrics for the test-set forecast performance of the seven models, ranked by R^2

Model	R^2	RMSE	MAE	MAPE (%)	Train time (s)	Test time (ms)
SVR	0.9981	0.00373	0.00220	0.419	1.0378	6.66
ANN	0.9979	0.00383	0.00269	0.489	1.0958	1.04
PR	0.9867	0.00974	0.00747	1.319	0.0031	0.47
RF	0.9763	0.01300	0.00949	1.705	0.5247	18.18
XGBoost	0.9743	0.01354	0.00631	1.240	0.1517	1.70
MLR	0.8985	0.02691	0.02206	3.829	0.0013	0.31
DT	0.8982	0.02694	0.01834	3.349	0.0033	0.26

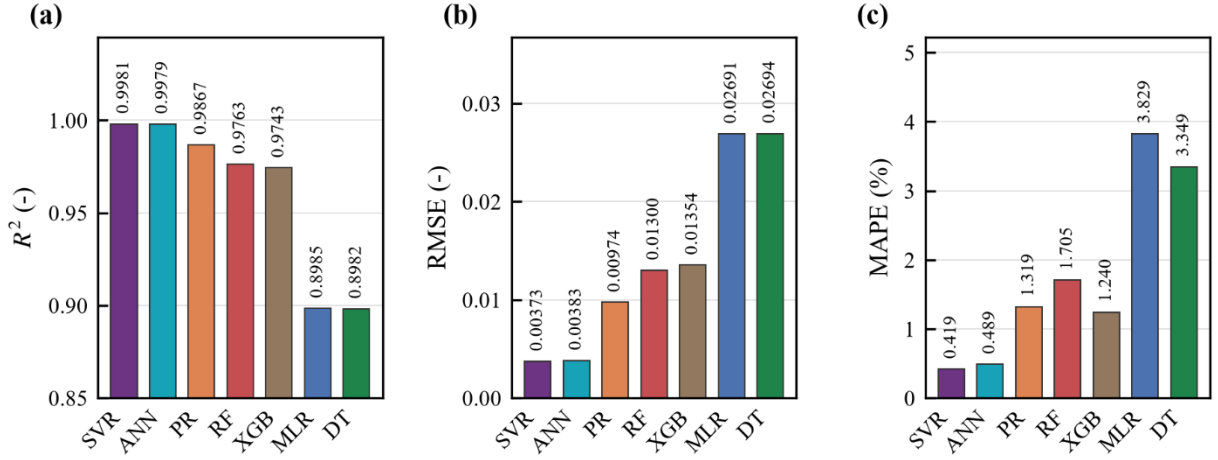


Figure 8. Test-set performance of the seven models for η_{col} : (a) R^2 , (b) RMSE, (c) MAPE

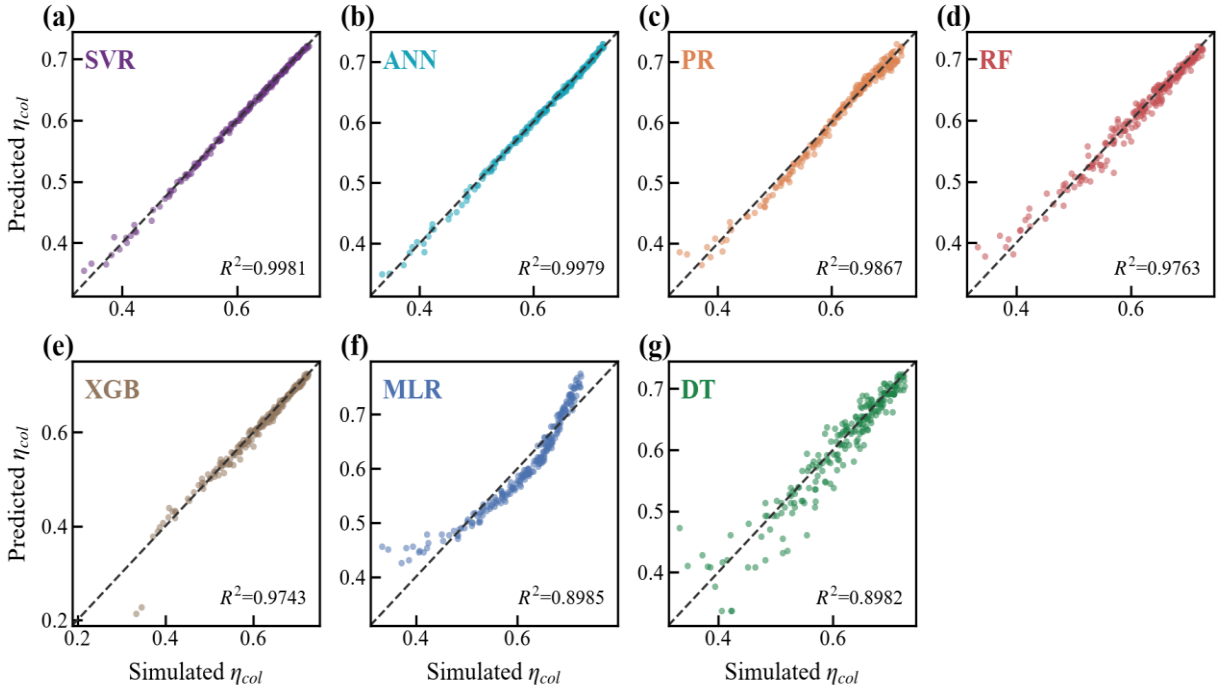


Figure 9. η_{col} actual versus predicted plots (a) SVR, (b) ANN, (c) PR, (d) RF, (e) XGBoost, (f) MLR, and (g) DT.

6.2. Proposed model against related work

Table 4 summarizes recent studies that forecast the performance of SWHS using data-driven techniques, together with the position of the proposed work [24–27]. The comparison clearly demonstrates the advantages of the proposed work. Unlike previous studies that rely on individual prediction models, multiple experimentally derived configurations, detailed transient thermal simulations, or computationally intensive hybrid optimization frameworks, the proposed approach establishes a unified benchmarking framework encompassing seven regression families for an indirect FPC solar water-heating system with PV-driven circulation [24–26]. Using 1097 simulated design points, the proposed SVR model achieves a high prediction accuracy of $R^2 = 0.9981$, $RMSE = 0.0037$, and $MAPE = 0.42\%$. Although some previous studies report marginally higher R^2 values, these results are obtained using considerably more complex frameworks, such as ANFIS [25]. In contrast, the proposed SVR achieves competitive accuracy without requiring hybrid optimization, evolutionary algorithms, or complex algorithmic coupling, resulting in a favourable balance between accuracy and implementation complexity. Consequently, the proposed work offers not only accurate SWHS performance prediction but also a simpler, computationally efficient, and practically deployable modelling framework suitable for system design, performance assessment, optimization, and future real-time control applications.

Table 4. Recent studies forecasting SWHS performance, and the position of the proposed work.

Ref.	System and rating	Forecasting technique	Accuracy	Implementation complexity
[24]	FPC with parallel tubes and copper-nanotube nanofluid	ANN and algebraic correlation	$R^2 = 0.949\text{--}0.999$ (ANN); $R^2 = 0.886\text{--}0.953$ (correlation)	Moderate – high ANN requires training and tuning; correlation is relatively simple
[25]	V-trough thermosiphon SWH with five 1-m risers beneath a 0.2-m absorber and twisted-tape inserts; 162 measured points	ANFIS, GLR, RT, and SVM	ANFIS: $R^2 = 0.9997$, RMSE = 0.45; SVM: $R^2 = 0.961$	High – multiple ML models and an ANFIS framework require parameter tuning and training
[26]	2-m ² FPC thermosiphon SWHS with 12 risers and single glazing	Nodal transient thermal model	Not reported; model not experimentally validated	High – detailed transient thermal modeling with multiple design parameters
[27]	Domestic parabolic-dish SWH experimental unit	ANN-GA and ANN-GWO hybrid models	$R^2 = 0.9969$, MAPE = 0.69% (ANN-GA); $R^2 = 0.97$, MAPE = 1.23% (ANN-GWO);	Very high – ANN coupled with evolutionary optimization algorithms increases framework complexity and computational requirements
Proposed work	Indirect FPC SWHS with photovoltaic-driven circulation; 1097 simulated design points	MLR, PR, SVR, DT, RF, XGBoost, and ANN	SVR: $R^2 = 0.9981$, $RMSE = 0.0037$, $MAPE = 0.42\%$; ANN: $R^2 = 0.9979$	Low–moderate – accurate prediction is achieved with SVR without hybrid optimization or complex algorithmic coupling

7. Conclusions

SWHS are commonly utilized as one of the most convenient applications of solar energy due to its economic and environmental benefits. This paper presents a data driven study for SWHS performance forecast based on MATLAB/Simulink modeling of an indirect flat-plate SWHS with PV-driven circulation. First, simulation implementation was verified against the mathematically expected collector-efficiency characteristic and selected physical response trends. Second, the verified Simulink model was repeatedly executed under 1097 combinations of climatic, operating, and design conditions to generate a comprehensive dataset. The dataset comprised eight input variables, including solar

irradiance, ambient temperature, wind speed, storage-tank volume, heat-exchanger effectiveness, make-up water temperature, delivery temperature, and heat-exchanger return temperature, and collector efficiency was selected as the prediction target. Finally, the generated dataset was subsequently characterized, pre-processed, and used to train and compare seven data-driven models, among which SVR provided the best overall error and fitting performance, closely followed by the ANN model, the second-order PR and RF, while XGBoost, MLR and DT showed progressively lower accuracy. Overall, the results demonstrate that ML surrogate models can reproduce the collector-efficiency response of the physics-based solar water-heating model with high accuracy while substantially reducing the computational effort required for repeated evaluations.

Acknowledgment

The authors would like to acknowledge the AIESI project (NORPART-2021/10403) and funding received from the Norwegian Directorate for Higher Education and Skills (HK-dir) through the NORPART scheme. This publication reflects only the authors' views, and the HK-dir is not liable for any use that may be made of the information contained therein.

References

- [1] Özcan Y and Deniz E 2023 Solar thermal waste heat energy recovery in solar distillation systems by using thermoelectric generators. *Eng. Sci. Technol. Int. J.* **40**, 101362.
- [2] Kanyarusoke K 2020 Problems of engineering entrepreneurship in Africa: a design optimization example in solar thermal engineering. *Eng. Sci. Technol. Int. J.* **23**(2), 345–356.
- [3] Özden E and Kaya M N 2025 Comprehensive analysis of key design parameters affecting the efficiency of flat plate solar thermal collectors. *Case Stud. Therm. Eng.* **73**, 106458.
- [4] Zakzouk N E, Abdelsalam A K, Helal A A and Williams B W 2014 DC-link voltage sensorless control technique for single-phase two-stage photovoltaic grid-connected system. *Proc. 2014 IEEE Int. Energy Conf. (ENERGYCON)* (Cavtat, Croatia) pp 58–64.
- [5] Ragab M M, Ibrahim R A, Desouki H and Swief R 2023 Optimal energy management applying load elasticity integrating renewable resources. *Sci. Rep.* **13**(1), 15000.
- [6] Gaber I M and Ibrahim R A 2024 Intraday solar irradiance forecasting based on hybrid machine learning methodology for photovoltaic power applications. *Future Directions in Energy Engineering* (Green Energy and Technology) (Cham: Springer) pp 11–20.
- [7] Zakzouk N E, El Dyasty A, Ahmed A and El Safty S M 2018 Power flow control of a standalone photovoltaic-fuel cell-battery hybrid system. *Proc. 7th Int. Conf. on Renewable Energy Research and Applications (ICRERA)* (Paris) pp 431–436.
- [8] Buker M S and Riffat S B 2015 Building integrated solar thermal collectors – a review. *Renew. Sustain. Energy Rev.* **51**, 327–346.
- [9] Al-Mamun M R *et al* 2023 State-of-the-art in solar water heating (SWH) systems for sustainable solar energy utilization: a comprehensive review. *Sol. Energy* **264**, 111998.
- [10] Naveenkumar R, Venkateshkumar R, Mohanavel V, Franklin C, Ismail S O, Ravichandran M, Kannan S and Soudagar M E M 2025 Recent developments in solar water heaters and solar collectors: a review on experimental and neural network analyses. *Results Eng.* **25**, 104394.
- [11] Fadzlin W A, Hasanuzzaman M, Rahim N A, Amin N and Said Z 2022 Global challenges of current building-integrated solar water heating technologies and its prospects: a comprehensive review. *Energies* **15**(14), 5125.
- [12] Xiao S, Zhang Y, Xia K and Long J 2022 Influence of operating conditions on solar energy utilization efficiency of flat plate solar collector. *Renew. Wind Water Sol.* **9**(1), 2.
- [13] Tiwari A K, Gupta S, Joshi A K, Raval F and Sojitra M 2021 TRNSYS simulation of flat plate solar collector based water heating system in Indian climatic condition. *Mater. Today Proc.* **46**, 5360–5365.
- [14] Hu X, Guo L, Wang J and Liu Y 2025 Computational fluid dynamics and machine learning integration for evaluating solar thermal collector efficiency-based parameter analysis. *Sci. Rep.*

- 15(1), 24528.
- [15] Du B, Lund P D and Wang J 2022 Improving the accuracy of predicting the performance of solar collectors through clustering analysis with artificial neural network models. *Energy Rep.* **8**, 3970–3981.
 - [16] Han J, Mesgarpour M, Asirvatham L G, Wongwises S, Ahn H S and Mahian O 2023 A hyper-optimisation method based on a physics-informed machine learning and point clouds for a flat plate solar collector. *J. Therm. Anal. Calorim.* **148**(13), 6223–6242.
 - [17] Cáceres M, Avila C and Rivera E 2024 Thermodynamics-informed neural networks for the design of solar collectors: an application on water heating in the highland areas of the Andes. *Energies* **17**(19), 4978.
 - [18] REDS Library 2024 Solar water heater system – FPC with PV, Renewable Energy and Desalination Simulation Library. Available: <https://www.redslibrary.com/product-page/solar-water-heater-system-fpc-pv> (accessed 25 August 2026).
 - [19] Alawi O A, Kamar H M, Salih S Q, Abba S I, Ahmed W, Homod R Z, Jamei M, Shafik S S and Yaseen Z M 2024 Development of optimized machine learning models for predicting flat plate solar collectors thermal efficiency associated with Al₂O₃-water nanofluids. *Eng. Appl. Artif. Intell.* **133**, 108158.
 - [20] Duffie J A and Beckman W A 2013 *Solar Engineering of Thermal Processes* 4th edn (Hoboken, NJ: Wiley).
 - [21] Kalogirou S A 2004 Solar thermal collectors and applications. *Prog. Energy Combust. Sci.* **30**(3), 231–295.
 - [22] ISO 9806:2017 Solar energy – Solar thermal collectors – Test methods (Geneva: International Organization for Standardization).
 - [23] NASA 2026 Prediction Of Worldwide Energy Resources (POWER) project. Available: <https://power.larc.nasa.gov/> (accessed 20 July 2026).
 - [24] Mausam K, Singh S, Ghosh S K and Singh R P 2024 Thermal performance modelling of solar flat plate parallel tube collector using ANN. *Energy* **303**, 131940.
 - [25] Saravanan A, Sree S R, Reddy M S, Elumalai P V, Karthik K, Cheeli A K and Hasan N 2024 Thermal performance prediction of a V-trough solar water heater with a modified twisted tape using ANFIS, G.L.R., R.T. and SVM models of machine learning. *Sci. Rep.* **14**(1), 27206.
 - [26] Batiana A, Tiendrebeogo S, Tubreoumya G, Sawadogo E and Sara B 2026 Theoretical optimization of performance indicators for the flat-plate solar collector of a thermosiphon solar water heater. *Sci. J. Energy Eng.* **14**(1), 21–31.
 - [27] Kabengele K T, Olayode I O and Tartibu L K 2026 Performance prediction of a domestic parabolic dish solar water heater using ANN-GA and ANN-GWO hybrid models. *Discov. Appl. Sci.*, doi:10.1007/s42452-026-09198-5.