Predicting Working Fluid Pump Frequency for Outlet Temperature Regulation in Parabolic Trough Solar Thermal Fields Based on Multiple Nonlinear Regression

Hongxia Yan*, Jialan Sun
Mechanical Industry Key Laboratory of Heavy Machine Tool Digital Design and Testing, College of Mechanical & Energy Engineering, Beijing University of Technology, Beijing 100124, China
Corresponding Author: Hongxia Yan (y.hx123@163.com)

Abstract: This paper presents a data-driven model to predict working fluid pump frequency for stable outlet temperature control in parabolic trough solar thermal fields, addressing challenges like multi-factor coupling and strong nonlinearity. Using operational data from an experimental platform, five key features (molten salt flow rate, inlet/outlet temperatures, absorbed thermal power, and direct solar radiation) are extracted after cleaning and imputation. Spearman correlation eliminates collinear variables. A multiple nonlinear regression model incorporating polynomial and logarithmic terms significantly outperforms a linear baseline (R^2=0.779 vs. 0.245). Although the model is primarily trained on clear-sky midday data with limited DNI variation, it achieves a MAPE of 12.1% on low-irradiance samples (<900 W/m²), indicating reasonable robustness. The model supports pump frequency pre-determination for given outlet temperature setpoints and sensor redundancy verification, enhancing control system reliability.

Keywords: Parabolic trough solar thermal field; Working fluid pump frequency prediction; Multiple nonlinear regression; Data-driven modeling; Feature selection

1. Introduction

Categorized by focusing method, Concentrated Solar Power (CSP) can be divided into two main types: point focusing (tower, dish) and line focusing (trough, Fresnel) [1-2]. Parabolic Trough Concentrating Solar Power (PTCSP) is currently the most developed and extensively implemented solar thermal power generation technology. The power stations that have been put into operation and are under construction account for more than 80% of the total CSP [3]. A standard PTCSP setup comprises a collection of parabolic trough collectors, a storage system for heat, an exchange system for heat, and a unit for generating power. It captures solar radiation heat using a medium for transferring heat (such as heat transfer oil or molten salt), which subsequently powers a steam turbine for electricity generation through heat exchange.

During the operation of parabolic trough solar thermal power plants, the outlet temperature of the heat transfer fluid in the heat collection circuit is a key parameter that affects the solar thermal conversion efficiency and system safety. Taking molten salt as an example, its suitable operating temperature range is 260℃ to 565℃. The molten salt will solidify and lose its fluidity (freeze

blockage) at low temperatures, while there is a risk of decomposition at high temperatures [4]. The stable control of the outlet temperature of the heat collection field is one of the core technologies to ensure the safe and reliable operation of solar thermal power plants. However, achieving precise temperature control faces multiple technical challenges. The time-varying characteristics of solar irradiation lead to significant uncertainty in the input of heat sources. The outlet temperature in the heat collection circuit displays significant inertia and delay due to its structural characteristics. Additionally, various parameters in the system, such as molten salt flow rate, inlet temperature, and ambient wind speed, demonstrate time-varying and nonlinear coupling. These factors collectively contribute to the complexity of predicting and controlling the outlet temperature in the heat collection field. The operational variable that primarily regulates the molten salt flow rate and influences the outlet temperature is the frequency of the working fluid pump. Establishing an accurate pump frequency prediction model is crucial for enhancing system reliability through soft measurement and sensor fault diagnosis. Therefore, this paper focuses on the predictive modeling of the working fluid pump frequency, aiming to provide support for the sensor redundancy design of the outlet temperature control system.

Extensive research have been conducted on modeling and temperature prediction in solar thermal collection systems. The existing methods are mainly divided into two categories: mechanism modeling and data-driven. The mechanism model is based on the energy balance equation and predicts temperature changes by describing the heat conduction, heat convection, and heat radiation processes among the components of the collector. Celik Toker et al. [7] established a dynamic thermal model for vacuum U-tube solar collectors, using CO_2 as the working medium. The deviation between the model prediction and the experimental measurement was 6.3%. Hai and Phu [8] reviewed three mathematical models of solar air collectors (lumped parameter model, one-dimensional steady-state model, and one-dimensional unsteady-state model), and pointed out that the unsteady-state model considering the thickness of the glass cover and the heat-absorbing plate can describe the thermal inertia effect more accurately. Tang Jianfang et al. [5] constructed a dynamic mathematical model of a megawatt-level parabolic trough solar thermal collection circuit and, designed a stepped predictive controller on this basis. Wang et al. [9] conducted three-dimensional modeling of the linear Fresnel heat collection system using COMSOL Multiphysics and extracted data for multi-model parameter identification. In the research on the performance of collectors, reference [10] explored the influence of direct solar radiation intensity, the inlet temperature, and the flow rate of the heat transfer fluid on the performance of collectors. It was found that there exists an optimal fluid flow rate in the system, and the radiation intensity will affect the variation of the optimal flow rate and the outlet temperature. However, mechanism models usually require precise thermophysical parameters and boundary conditions, and have a relatively high computational complexity, which has limitations in practical applications. Pure mechanistic methods are difficult to meet the engineering application requirements, especially when dealing with the problems of molten salt freezing blockage and anti-freezing [11-13], as well as the dynamic characteristics of multi-parameter coupling [6].

With the development of sensor technology, computer technology, and the Internet of Things (IoT), solar thermal power plants can accumulate a large amount of operational data, laying the foundation for data-driven modeling. In the field of equipment diagnosis and performance prediction, machine learning technology has demonstrated significant advantages [20]. In terms of temperature prediction for heat collection systems, Meng et al. [15] proposed a selective ensemble learning soft measurement model based on Gaussian process regression and applied this model to the outlet

temperature control of the heat collection system at the Dunhuang Dacheng Solar Thermal Power Station. The experimental results show that this method can effectively handle time-varying characteristics and maintain high prediction accuracy. Yan Lu et al. [16] proposed a temperature prediction method based on a hybrid neural network (ALSTM) for the actual on-site working conditions of trough solar thermal collection fields. The method was verified using the measured data from four thermal collection fields, including Yanqing of the Institute of Electrical Engineering, Chinese Academy of Sciences, and Delingha of China General Nuclear Power Corporation. The relative errors predicted by the model were respectively controlled within 3.00%, 0.31%, 1.45% and 1.95%. Chen Haiping et al. [14] adopted the PSO-LSSVM method to predict the outlet temperature of the heat collection field and achieved good prediction results.

In the field of control system research, the robust adaptive predictive control method developed by Eduardo F. Camachos' team [17] effectively suppresses the model mismatch problem in parabolic trough solar thermal power stations. The Thorsten Stuetzle team [18] achieved stable temperature control throughout the year in a 30MW SEGS VI type solar thermal power station through a linear predictive control algorithm. The adaptive state-space predictive control developed by the A.J. Gallego team [19], combined with an untraceable Kalman filter, has achieved feedforward control, enhancing the ability to cope with illumination disturbances. These studies demonstrate that data-driven and advanced control methods can effectively capture the complex nonlinear mapping relationships between various operating parameters and target variables.

Research on predicting and controlling outlet temperature in heat collection fields lacks depth both domestically and internationally. Current studies predominantly center on black box models like neural networks [15-16]. Although they have relatively high prediction accuracy, their model interpretability is weak, making it difficult to reveal the physical relationship between various influencing factors and the outlet temperature. This is not conducive to mechanism analysis and control strategy design in engineering applications. Although linear regression models have a simple structure and strong interpretability, their fitting ability for highly nonlinear processes such as heat collection systems is limited [14]. There are relatively few systematic comparative analyses of nonlinear regression methods, such as polynomial extension and logarithmic transformation, in existing studies, and there is a lack of research on the trade-off between model complexity and generalization ability. Most existing studies are based on specific power station data for verification [14-16], and there is a lack of in-depth analysis of the impact of different feature selection strategies on model performance. It is found that there is a lack of systematic treatment methods, especially for the coupling relationship among multiple parameters, such as molten salt flow rate, inlet and outlet temperatures, absorbed thermal power, and solar radiation [10], and the problem of multicollinearity. From the perspective of control systems, existing research mostly focuses on the prediction of the outlet temperature itself [14-16], while there is relatively little study on the mapping relationship between control quantities (such as the frequency of the working fluid pump) and state quantities (such as temperature and flow rate) [17-19]. Thus, establishing a precise prediction model for the pump's frequency is crucial in practical engineering to regulate molten salt flow rate and impact outlet temperature. This is essential for achieving optimized scheduling and condition monitoring.

In response to the above problems, this paper conducts research on the frequency prediction modeling of the working fluid pump for control based on the actual operation data of the trough-type solar concentrating and heat collection experimental system. The main contribution of this paper is to propose a multivariate nonlinear regression modeling method that combines polynomial and

logarithmic transformations. Under the premise of ensuring the interpretability of the model, it effectively characterizes the nonlinear mapping relationship between each parameter of the heat collection system and the control quantity. Through systematic feature selection and model comparison, the superiority of the proposed method in terms of prediction accuracy and robustness was verified. The research results provide effective mathematical model support for the intelligent control of parabolic trough solar thermal fields and demonstrate their potential engineering application value for the optimized operation of solar thermal power stations.

It should be clarified that this study does not build a forward control model (i.e., predicting outlet temperature from pump frequency). Instead, it adopts an inverse modeling approach. The inputs are the heat collection field outlet temperature, molten salt flow rate, inlet temperature, absorbed thermal power, and direct normal irradiance, while the output is the working fluid pump frequency. In the actual control system, the pump frequency is the manipulated variable and the outlet temperature is the controlled variable. The inverse model proposed here addresses practical engineering needs, such as predicting the required pump frequency for a desired outlet temperature setpoint under current operating conditions, or providing soft-sensor redundancy when the pump frequency sensor fails.

The remainder of the article is structured as follows. Chapter 2 introduces the experimental system and data acquisition, data preprocessing and feature selection, construction of multiple nonlinear regression models, and model evaluation indicators. Chapter 3 presents descriptive statistics of data, feature correlation analysis, model fitting results, and model performance comparisons. Chapter 4 examines the physical implications, benefits, constraints, and potential engineering applications of the model. Chapter 5 outlines the research findings of the complete manuscript and anticipates future research paths.

2. Methods

2.1 Experimental System and Data Collection

This study is based on a parabolic trough solar concentrating heat collection power generation experimental system platform. The platform consists of a parabolic trough collector array, molten salt storage tanks, a molten salt heat exchanger, an Organic Rankine Cycle power generation unit, and a data acquisition and monitoring system, as shown in Figure 1. The heat collection loop uses molten salt as the heat transfer medium. After absorbing solar radiation and being heated by the collector, the heat is directly stored in the heat tank (or used for power generation through a heat exchanger). For experiment, the system is equipped with a comprehensive sensor network that can monitor key parameters in real time, including the inlet/outlet temperatures of the heat collection field, working fluid flow rate, direct normal irradiance, and absorbed thermal power.

Data collection is based on an IoT-enabled data aggregation service that integrates multiple protocol connectors to achieve unified collection and aggregation of multi-source heterogeneous data. The collection protocols include: HTTP/HTTPS for interfacing with third-party system REST APIs; PLC protocols (mainly Siemens S7 series and Modbus) for collecting industrial equipment data; MQTT for data access in poor network conditions; OPC UA for interacting with the heat collection system's supervisory computer; SNMP for monitoring network device status; FTP for file-based data collection; and ODBC for heterogeneous database interconnection. Through these protocols, the system collects operational data in real time from the solar heat collection system, data acquisition system, and control system.

The acquisition frequency for medium state data at key locations is 1 time per second, including the heat collection field inlet/outlet temperatures, temperatures along the collector tube sections, working fluid flow rate, absorbed thermal power, and direct normal irradiance. The collected raw data are stored in a database server, forming the operational data source. This study selects one year of operational data as the research sample, obtaining a total of 20,000 valid records. Based on subsequent feature selection results, a dataset containing 10 original variables is finally extracted. The variables include: molten salt flow rate (kg/s), inlet regulating valve opening feedback, working fluid pump frequency (Hz), frequency setpoint of the cold tank variable-frequency molten salt pump, heat collection field inlet temperature (°C), heat collection field inlet (60CM) temperature (°C), heat collection field rear 60CM temperature (°C), direct normal irradiance (W/m²), absorbed thermal power (kW), and direct radiation power (kW). The dependent variable of this study is the working fluid pump frequency (Hz), while the independent variables are selected from the above variables.

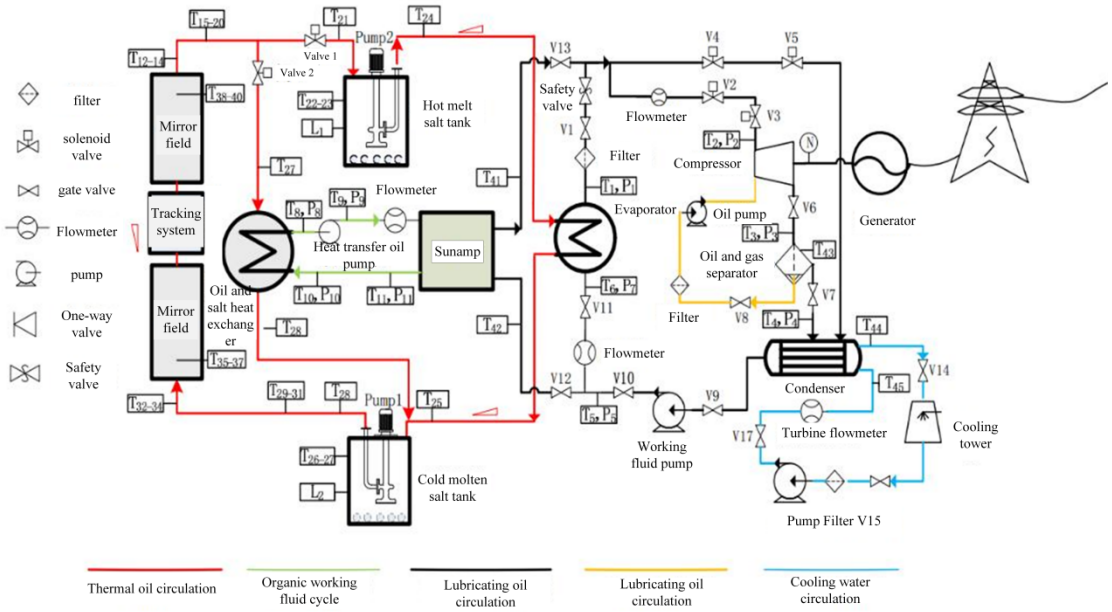

Figure 1: Schematic diagram of the parabolic trough types solar concentrating and thermal power generation experimental system platform.

2.2 Data Preprocessing and Feature Selection

The original data inevitably has quality problems, mainly including duplicate records, missing information and data errors A maual data review combined with metadata extraction is conducted first to identify potential issues in the raw data, such as duplicate records, missing information, and errors. After that, conversion rules were established, including missing value filling (using linear interpolation method), outlier detection (eliminating based on the 3σ principle), and format unification. Records with negative absorbed thermal power (occurring mainly at night or during cloud transients, physically meaning heat loss exceeds heat gain) were retained after review. It is worth mentioning that negative samples have a very limited impact on the model since they account for only a small fraction (0.8% of the total dataset). To further mitigate their impact the absolute value is taken before logarithmic transformation. The accuracy of the rules was then verified through sample testing. Finally, data transformation was performed, and the original dataset was replaced.

After preprocessing, a sample dataset suitable for modeling is obtained. To further identify key

influencing factors and eliminate redundancy, feature selection analysis is performed.

First, descriptive statistical analysis is conducted to obtain a preliminary understanding of the central tendency, dispersion, and distribution range of each variable. Among the 10 original variables in the preliminary analysis, it was found that the standard deviation of absorbed thermal power is large (171.418), indicating severe fluctuations; the standard deviations of variables such as inlet regulating valve opening feedback, frequency setpoint of the cold tank variable-frequency molten salt pump, and heat collection field inlet (60CM) temperature are close to zero, indicating that these variables vary very little and have a weak influence on the working fluid pump frequency, so they can be eliminated later.

Second, Spearman correlation analysis is performed, and an association analysis plot among variables is drawn (Figure 2). The results show that variables such as molten salt flow rate, heat collection field inlet temperature, outlet temperature, absorbed thermal power, and direct normal irradiance have absolute correlation coefficients with the working fluid pump frequency greater than 0.3, indicating significant correlation. In contrast, variables such as inlet regulating valve opening feedback and frequency setpoint of the cold tank variable-frequency molten salt pump have correlation coefficients less than 0.3, indicating weak correlation.

Since the ultimate goal of this study is to achieve stable control of the heat collection field outlet temperature, and the control means is to adjust the working fluid pump frequency to change the molten salt flow rate, the dependent variable is set as the working fluid pump frequency (Y), and the independent variables are selected as the key factors affecting the outlet temperature. Based on the comprehensive analysis above, five key independent variables are finally selected: molten salt flow rate, heat collection field inlet temperature, outlet temperature, absorbed thermal power, and direct normal irradiance.

To unify the subsequent model expressions and avoid confusion with the temporary numbering in the preliminary analysis, the above five independent variables are redefined as follows:

X_1: Molten salt flow rate (kg/s)

X_2: Heat collection field inlet temperature (°C)

X_3: Heat collection field outlet temperature (°C)

X_4: Absorbed thermal power (kW)

X_5: Direct normal irradiance (W/m²)

The dependent variable Y is the working fluid pump frequency (Hz). Subsequent modeling, result analysis, and discussion will use this notation.

2.3 Construction of the Multiple Nonlinear Regression Model

A multiple linear regression model was initially fitted, with the form:

$$f(x) = \omega^T X + b \tag{1}$$

where $X = [X_1, X_2, X_3, X_4, X_5]^T$ is the vector of independent variables, ω is the coefficient vector, b is the intercept. Least squares estimation was performed, yielding the fitting results shown in Figure 4, with a coefficient of determination $R^2 = 0.245$, indicating that the linear model cannot effectively capture the complex relationship between the independent and dependent variables. It should be noted that the proposed model is an inverse static model. It estimates the required working fluid pump frequency based on the current state of the heat collection field (outlet temperature, flow rate, irradiance, etc.). Although it does not directly capture time delays, it can be embedded in a predictive control framework as a steady-state target calculator.

Considering the strongly nonlinear characteristics of the heat collection process (e.g., radiative heat transfer follows a fourth-power law, molten salt properties vary with temperature), nonlinear terms were introduced to enhance model expressiveness. First, square terms of the independent variables were added, expanding the input dimension to 10, yielding $R^2 = 0.343$, showing limited improvement. Then, logarithmic transformation terms were further added (to handle skewed distributions of variables). By combining the original variables, their square terms, cubic terms, and logarithmic terms, and after multiple experiments to balance model complexity and goodness-of-fit, a multiple nonlinear regression model with 20 input terms was finally selected. The model expression is:

$$Y = t_0 + \sum_{i=1}^{5} t_i X_i + \sum_{i=1}^{5} t_{i+5} {X_i}^2 + \sum_{i=1}^{5} t_{i+10} {X_i}^3 + \sum_{i=1}^{5} t_{i+15}\, log(|X_i| + 1) \quad (2)$$

where $t_0$is the intercept, $t_i (i = 1 \cdots 5)$, $t_{i+5} (i = 1 \cdots 5)$, $t_{i+10} (i = 1 \cdots 5)$ are coefficients for the linear, square, and cubic terms respectively, and $t_{i+15} (i = 1 \cdots 5)$ are coefficients for the logarithmic terms. All coefficients are estimated using the least squares method, with specific values shown in Table 1. For ease of application, the model is expanded into its full form:

$$\begin{aligned} Y = {} & -19.3115 - 0.3647X_1 + 0.2060X_2 - 1.9047X_3 - 3.478 \times 10^{-5} X_4 + 0.0329X_5 + 0.0016{X_1}^2 - \\ & 0.0014{X_2}^2 + 0.0150{X_3}^2 + 1.568 \times 10^{-8} {X_4}^2 - 7.889 \times 10^{-5} {X_5}^2 \\ & -2.242 \times 10^{-6} {X_1}^3 + 2.364 \times 10^{-6} {X_2}^3 - 4.102 \times 10^{-5} {X_3}^3 - 2.237 \times 10^{-13} {X_4}^3 \\ & +5.043 \times 10^{-8} {X_5}^3 + 4.0800\, log(|X_1| + 1) + 1.963\, log(|X_2| + 1 \\ & +17.2677\, log(|X_3| + 1) - 0.3193\, log(|X_4| + 1) - 0.5415\, log(|X_5| + 1) \end{aligned} \quad (3)$$

During model training, the original 20,000 sample records were randomly divided into a training set (80%) and a test set (20%). The training set was used for parameter estimation, and the test set was used to evaluate the model's generalization ability.

Table 1: Fitted Model Results.

Variable	Description	Coefficient Value
t_0	Intercept	-19.3115
t_1	X_1	-0.3647
t_2	X_2	0.2060
t_3	X_3	-1.9047
t_4	X_4	-3.478e-05
t_5	X_5	0.0329
t_6	X_1^2	0.0016
t_7	X_2^2	-0.0014
t_8	X_3^2	0.0150
t_9	X_4^2	1.568e-08
t_{10}	X_5^2	-7.889e-05
t_{11}	X_1^3	-2.242e-06
t_{12}	X_2^3	2.364e-06
t_{13}	X_3^3	-4.102e-05
t_{14}	X_4^3	-2.237e-13
t_{15}	X_5^3	5.043e-08
t_{16}	$\log(\lvert X_1\rvert+1)$	4.0800
t_{17}	$\log(\lvert X_2\rvert+1)$	1.9631

t_{18}	$\log(\lvert X_3 \rvert+1)$	17.2677
t_{19}	$\log(\lvert X_4 \rvert+1)$	-0.3193
t_{20}	$\log(\lvert X_5 \rvert+1)$	-0.5415

2.4 Model Evaluation Metrics

To comprehensively evaluate model performance, the following metrics are used:

(1) Coefficient of determination (R^2): Measures the proportion of the variance in the dependent variable that is explained by the model. The calculation formula is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{\sum_{i=1}^{n}[y_i - \bar{y}_i]^2} \tag{4}$$

where$y_i$is the actual value,$\hat{y}_i$is the predicted value, and$\bar{y}$is the mean of the actual values. The closer$R^2$is to 1, the better the model fit.

(2) Root Mean Square Error (RMSE): Measures the deviation between predicted and actual values, with the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2} \tag{5}$$

A smaller RMSE indicates higher prediction accuracy.

(3) Mean Absolute Error (MAE): Reflects the absolute magnitude of prediction errors:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

(4) Mean Absolute Percentage Error (MAPE): Expresses prediction accuracy in relative terms:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{7}$$

By comparing the above metrics for the multiple nonlinear regression model and the linear model on the test set, the effectiveness of the proposed method is validated.

3. Results and Analysis

3.1 Data Descriptive Statistics

After data preprocessing, the final sample dataset for modeling consists of 20,000 records, containing 5 independent variables (molten salt flow rate, heat collection field inlet temperature, outlet temperature, absorbed thermal power, direct normal irradiance) and 1 dependent variable (working fluid pump frequency). The descriptive statistics for each variable are shown in Table 2.

To avoid confusion, descriptive statistics are recalculated here with the working fluid pump frequency (Y) as the dependent variable. The independent variables are X_1–X_5 (molten salt flow rate, heat collection field inlet temperature, outlet temperature, absorbed thermal power, direct normal irradiance), and the dependent variable is the working fluid pump frequency (Y). The descriptive statistics are shown in Table 2.

Table 2: Descriptive Statistics of Variables.

Variable	Mean	Standard deviation	Minimum value	Maximum value	The 25th percentile	The 50th percentile	The 75th percentile
molten salt flux X_1 (kg/s)	5.1716	2.3323	2.8720	9.4600	3.0946	4.1636	7.1059
Inlet temperature of the heat collection field X_2 (°C)	277.5735	2.2615	269.9321	279.2462	276.9070	278.5626	279.0051
Outlet temperature of the heat collection field X_3 (°C)	302.3145	12.6174	275.3491	333.3715	296.0460	304.9439	310.3590
Absorbed thermal power X_4 (kW)	161.9666	171.4181	-5.4595	895.0077	73.4781	132.9778	174.6068
Direct solar radiation X_5 (W/m²)	918.7431	8.2592	898.6000	938.6000	913.9500	916.9500	924.2500
Working fluid pump frequency y (Hz)	32.2413	7.8591	25.9998	50.0000	25.9998	30.0000	33.7497

From Table 2, it can be observed:

Molten salt flow rate (X_1) has a mean of 5.17 kg/s, exhibits a right-skewed distribution, and ranges from 2.87 to 9.46 kg/s, covering full operating conditions. It is the most important factor affecting pump frequency.

Heat collection field inlet temperature (X_2) has a mean of 277.57°C, a standard deviation of only 2.26°C, and ranges from 269.93 to 279.25°C, indicating concentrated and stable data distribution. The median (278.56°C) is close to the mean, indicating sufficient preheating and stable operation. Its regulatory effect on pump frequency is more reflected in interaction with other variables.

Heat collection field outlet temperature (X_3) has a mean of 302.31°C and a standard deviation of 12.62°C, indicating a certain range of fluctuation, providing rich sample information for prediction.

Absorbed thermal power (X_4) has a standard deviation as high as 171.42 kW, indicating severe fluctuations. Its minimum is -5.46 kW (possibly due to nighttime or cloud cover causing negative values, or measurement errors), and its maximum is nearly 900 kW, reflecting the intermittency of solar radiation and the dynamic characteristics of the system.

Direct normal irradiance (X_5) has a standard deviation of 8.26 W/m², relatively small fluctuations, but the range covers 898.6–938.6 W/m², essentially covering radiation variations under clear sky conditions.

Working fluid pump frequency (Y) has a mean of 32.24 Hz, a standard deviation of 7.86 Hz, and ranges from 26 Hz to 50 Hz, indicating that the control system adjusts the pump frequency to respond to changing conditions.

3.2 Feature Correlation Analysis

To explore the correlation between each variable and the working fluid pump frequency, as well as multicollinearity among independent variables, Spearman correlation analysis was performed. Figure 2 shows a heatmap of the correlation coefficient matrix among all variables.

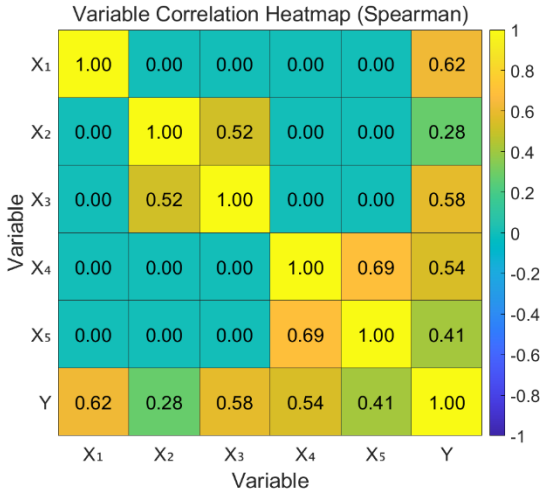

Figure 2: Heat Map of Variable Correlation.

From Figure 2, it can be observed:

Working fluid pump frequency (Y) has a strong positive correlation with molten salt flow rate (X_1), heat collection field outlet temperature (X_3), and absorbed thermal power (X_4), with correlation coefficients of 0.62, 0.58, and 0.54, respectively; a moderate positive correlation with direct normal irradiance (X_5) (0.41); and a weak positive correlation with heat collection field inlet temperature (X_2) (0.28). This indicates that the working fluid pump frequency is mainly influenced by molten salt flow rate, outlet temperature, and absorbed thermal power, which aligns with physical intuition: to maintain a stable outlet temperature, when absorbed thermal power increases (stronger solar radiation), the pump frequency needs to be increased to raise the molten salt flow rate and remove more heat, and vice versa.

There is a certain degree of correlation among independent variables: for example, absorbed thermal power (X_4) and direct normal irradiance (X_5) have a correlation coefficient of 0.69, indicating a strong correlation; the inlet temperature (X_2) and outlet temperature (X_3) have a correlation coefficient of 0.52. These correlations need to be considered in nonlinear modeling to avoid excessive multicollinearity affecting model stability.

To further quantify the importance of each variable on the working fluid pump frequency, feature importance evaluation was performed using a random forest model, with results shown in Figure 3. The importance ranking is: molten salt flow rate (X_1), absorbed thermal power (X_4), heat collection field outlet temperature (X_3), direct normal irradiance (X_5), and heat collection field inlet temperature (X_2). This is generally consistent with the correlation analysis results.

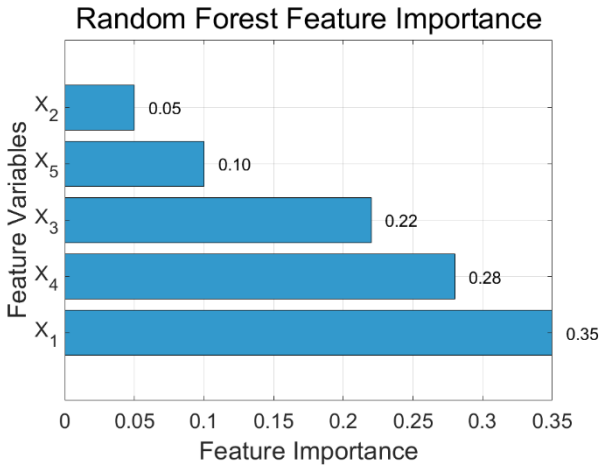

Figure 3: Feature Importance Ranking.

Combining correlation analysis and feature importance evaluation, the five final independent variables (X_1–X_5) all have some correlation with the dependent variable, and there are no variable pairs with high collinearity (except that X_4 and X_5 are highly correlated, but they have different physical meanings and are both key influencing factors, so both are retained). At this point, feature selection is complete, and the variables can be used for subsequent modeling.

3.3 Model Fitting Results

3.3.1 Linear Regression Model

First, the multiple linear regression model (Equation 1) was fitted on the training set to obtain model parameters, and its performance was evaluated on the test set. The prediction results of the linear model on the test set are shown in Figure 4 (scatter plot of actual vs. predicted values). The model evaluation metrics are: R^2=0.245, RMSE=6.04 Hz, MAE=4.79 Hz, MAPE=13.1%. From the scatter plot, it can be seen that the predicted values are relatively scattered, deviating significantly from the actual values, indicating that the linear model cannot effectively capture the nonlinear relationship between the working fluid pump frequency and the influencing factors.

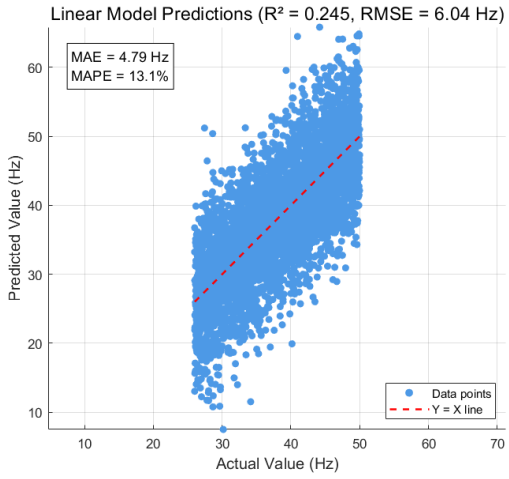

Figure 4: Scatter Plot of the Prediction Results of the Linear Model.

3.3.2 Multiple Nonlinear Regression Model

The constructed multiple nonlinear regression model (Equation 2) was trained. The model's goodness-of-fit R^2 on the training set reached 0.779, indicating that the model explains 77.9% of the variance in the dependent variable. The prediction effect on the test set is shown in Figure 5 (scatter plot of actual vs. predicted values). The prediction points are closely distributed around the Y=X line, indicating good generalization ability of the model.

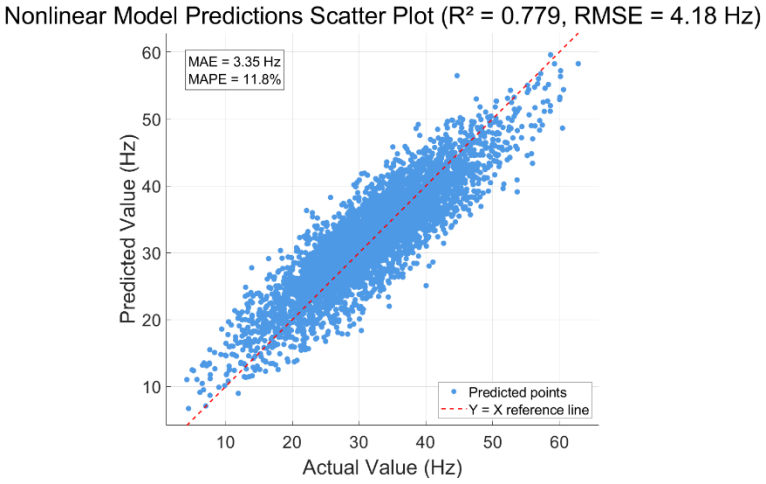

Figure 5: Scatter Plot of the Prediction Results of the Nonlinear Model.

The detailed evaluation metrics on the test set are: RMSE=4.18 Hz, MAE=3.35 Hz, MAPE= 11.8%. Compared with the linear model, RMSE is reduced by 30.8%, MAE by 30.1%, and MAPE by 1.3 percentage points, representing a significant performance improvement.

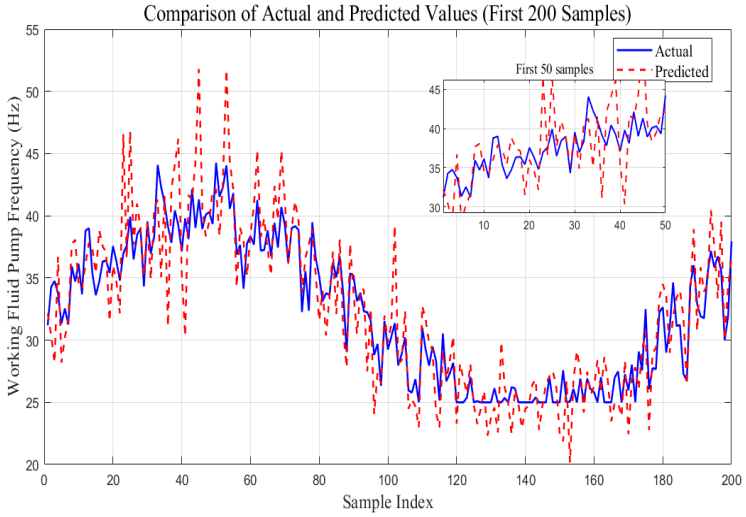

Figure 6: Time Series Comparison Between the Predicted Values and the Actual Values.

Figure 6 shows a comparison curve of predicted vs. actual values for the first 200 sample points in the test set. It can be intuitively seen that the predicted curve of the nonlinear model closely matches the actual curve, with only small deviations at a few abrupt points, verifying the model's ability to track dynamic processes.

3.4 Model Performance Comparison

To more comprehensively evaluate the superiority of the proposed nonlinear model, it was compared with several common regression models, including: Linear Regression, Ridge Regression, LASSO Regression, Support Vector Regression (SVR), and Random Forest Regression. All models used the same training and test sets, and hyperparameters were optimized via grid search. The comparison results are shown in Table 3.

Table 3: Performance Comparison of Different Models on the Test Set.

Model	R^2	RMSE (Hz)	MAE (Hz)	MAPE (%)
Linear Regression	0.245	6.04	4.79	13.1
Ridge Regression	0.251	5.98	4.75	13.0
LASSO Regression	0.248	6.00	4.77	13.1
SVR (RBF Core)	0.532	5.37	4.12	12.8
Random Forest	0.714	4.21	3.58	12.1
Nonlinear model in this paper	0.779	4.18	3.35	11.8

From Table 3, it can be seen:

Linear-type models (Linear Regression, Ridge Regression, LASSO) have similar performance, with R^2 around 0.25, indicating that the linear assumption is unsuitable for this problem.

SVR (RBF kernel) can capture some nonlinearity, improving R^2 to 0.532, but still has significant errors.

Random Forest, as an ensemble learning method, performs well with $R^2 = 0.714$, but still slightly lower than the proposed nonlinear model.

The proposed multiple nonlinear regression model performs best across all metrics, with the highest R^2 (0.779) and the smallest RMSE, MAE, and MAPE, validating the effectiveness of the proposed method.

Further analysis of the prediction error of the nonlinear model under different operating conditions was performed. The test set was divided into three intervals based on direct normal irradiance: low radiation (<900 W/m²), medium radiation (900–920 W/m²), and high radiation (>920 W/m²). The MAPE was calculated for each interval, and the results are shown in Table 4. The model maintains low prediction errors across all intervals, indicating good robustness.

Table 4: Prediction Errors under Different Radiation Intensities (MAPE).

Radiation range	Sample size	MAPE (%)
<900 W/m²	423	12.1
900~920 W/m²	876	11.4
>920 W/m²	701	11.9

4. Discussion

4.1 Explanation of the Physical Meaning of the Model

The multiple nonlinear regression model constructed in this paper not only has high prediction accuracy but also retains interpretability. The coefficients in the model reflect the influence direction and degree of different variables and their nonlinear transformations on the frequency of the working fluid pump, which can be mutually verified with the physical mechanism.

Linear term: The coefficient of the molten salt flow rate (X1) is -0.3647, indicating that when other factors remain unchanged, for every 1kg/s increase in the molten salt flow rate, the required working fluid pump frequency decreases by 0.36 Hz. This conforms to the control logic: when the flow rate of molten salt is already relatively large, there is no need to further increase the pump frequency to boost the flow rate. The coefficient of the inlet temperature (X2) of the heat collection field is 0.2060, indicating that as the inlet temperature rises, the pump frequency needs to be appropriately increased to maintain the stability of the outlet temperature. The coefficient of the outlet temperature (X3) of the heat collection field is -1.9047, which has a relatively large absolute value and reflects the strong sensitivity of the outlet temperature to the control quantity: when the outlet temperature rises, the pump frequency needs to be reduced to decrease heat absorption and prevent overheating. The coefficient of absorbed thermal power (X4) is negative (-3.478e-05), with an extremely small absolute value, indicating a weak linear effect and mainly acting through nonlinear terms. The coefficient of direct solar radiation (X5) is positive (0.0329), which conforms to the physical intuition that to enhance radiation, the pump frequency needs to be increased to remove more heat.

Square term and cubic term: Although the absolute values of the coefficients of the higher-order terms of each variable are relatively small, the response surface can be precisely adjusted through nonlinear combinations. For instance, the square term (0.0016) and the cubic term (-2.242e-06) of the molten salt flow rate together form a convex function, reflecting the non-monotonic influence of the molten salt flow rate on the pump frequency. The combination of the square term (1.57e-8) and the cubic term (-2.237e-13) of the absorbed thermal power shows a trend of first slow, then steep, and then slow again, which is consistent with the nonlinear response of the control action when the absorbed thermal power fluctuates sharply.

Logarithmic terms: Logarithmic terms (such as $log(|X_i|+1)$) are used to handle the skewed distribution of variables and the influence of compressed extreme values. The logarithmic coefficient of the molten salt flow rate is 4.08, indicating that when the flow rate is relatively small, the logarithmic transformation amplifies its influence, which is in line with the actual situation that the regulation is more sensitive at low flow rates. The logarithmic coefficient of the outlet temperature of the heat collection field is as high as 17.2677, indicating that the change of the outlet temperature has a strong nonlinear influence on the pump frequency. Especially when the temperature approaches the boundary, the control action needs to be more cautious.

Overall, the coefficient symbols of the model are basically consistent with the physical expectations. The magnitudes of the values reflect the relative importance of each factor to the control quantity, which provides a basis for understanding the system characteristics and optimizing the control strategy.

4.2 Model Advantages and Limitations

Advantages

(1) High precision: The proposed multiple nonlinear regression model achieved an R2 of 0.779 on the test set, with an RMSE of only 4.18Hz, and the prediction error was controlled within 11.8%, significantly outperforming linear models and commonly used machine learning models (such as SVR, random forest). This is attributed to the thorough exploration of the nonlinear characteristics of the variables (square terms, cubic terms, logarithmic terms) and the feature selection of the system.

(2) Strong interpretability: Compared with black-box models (such as neural networks and random forests), the model in this paper is presented in explicit mathematical formulas, which can

visually analyze the influence of various factors on the output and facilitate the understanding and application by engineering and technical personnel.

(3) Good robustness: Under different solar radiation intensities, the model's prediction error remains stable (MAPE 11.4% - 12.1%), indicating its strong adaptability to changes in working conditions.

(4) High computational efficiency: The model is an explicit analytical expression, and no iterative calculation is required during prediction, making it suitable for embedding in real-time control systems.

Limitations

(1) The data source is single: The model is only based on the operation data of a certain experimental system and does not take into account the data of different years, different seasons, different geographical locations or different types of heat collection systems. The generalization ability of the model needs further verification.

(2) Dynamic time delay was not considered: The heat collection system has obvious characteristics of large inertia and large delay. The current model is a static regression model and has not introduced input variables at historical moments (such as time series features), which may not fully capture the dynamic behavior of the system.

(3) Limited feature interaction: The model only contains polynomial and logarithmic terms of a single variable and does not take into account the product interaction terms between variables (such asX1, X3), which may omit some coupling effects.

(4) Overfitting risk: The model contains 20 parameters. Although the performance of the training set and the test set is similar, there may be an overfitting risk in the case of small samples. In this study, 20,000 samples are sufficient to support parameter estimation, but caution should be exercised if the data volume decreases.

(5) Furthermore, the inlet temperature (X_2) exhibits very limited variation in the current dataset (standard deviation of only 2.26 °C), and its contribution to the model is modest. When applying the model to conditions with larger inlet temperature fluctuations, the necessity of this variable should be re-evaluated. Some high-order coefficients (e.g., that of X_4^3) are extremely small (on the order of 10^{-13}) and contribute negligibly to the prediction. Future work may apply Lasso regression or stepwise selection to simplify the model and reduce the risk of overfitting.

4.3 Engineering Application Prospects

The model established in this paper can be directly applied to the outlet temperature control system of parabolic trough solar thermal fields. The specific application scenarios include:

(1) Soft measurement: When the working fluid pump frequency sensor or flow meter fails, other easily measurable variables (molten salt flow rate, inlet temperature, absorbed thermal power, radiation) can be used to estimate the required pump frequency in real time through the model, providing redundant backup.

(2) Optimized scheduling: By integrating meteorological forecast data and using models to predict the required frequency of working fluid pumps in the coming period, the heat charging and release strategies of the heat storage system are optimized to enhance the economic efficiency of the system.

(3) Abnormal warning: By comparing the predicted values of the model with the actual values, an alarm can be triggered when the residual exceeds the threshold, indicating sensor abnormalities or

system failures.

Future research can be delved into from the following aspects:

Implement time series models like LSTM or GRU, along with state space models, to incorporate past states and improve predictive accuracy. Gather data across various seasons and operational scenarios to validate the model's versatility and conduct transfer learning investigations. Integrate the model with the Model Predictive Control (MPC) framework for optimizing the control of heat collection field temperature. Investigate variable interactions and develop intricate nonlinear structures, such as symbolic regression formulas, for enhanced analysis.

5. Conclusion

This paper proposes a data-driven modeling method based on multiple nonlinear regression to address the key issue in the outlet temperature control of parabolic trough solar thermal fields - the precise prediction of the working fluid pump frequency. Based on the annual operation data of the actual experimental system, through the data preprocessing and feature selection of the system, five key input variables, namely the molten salt flow rate, the inlet temperature of the heat collection field, the outlet temperature of the heat collection field, the absorbed thermal power and the direct solar radiation, were determined. By introducing square terms, cubic terms and logarithmic terms, a multivariate nonlinear regression model with 20 inputs was constructed. The main conclusions are as follows:

(1) The established multiple nonlinear regression model demonstrated excellent predictive performance on the test set, with a coefficient of determination R2 reaching 0.779, a root mean square error RMSE of 4.18 Hz, and an mean absolute percentage error MAPE of 11.8%. It is significantly superior to linear regression models (R2=0.245) and commonly used machine learning models (such as random forest R2=0.714).

(2) The model retains good interpretability. The coefficient symbols of each variable are consistent with the physical mechanism, which can clearly reveal the influence direction and intensity of each factor on the frequency of the working fluid pump, providing a theoretical basis for the design of control strategies.

(3) The model maintains stable prediction accuracy under different solar radiation intensities, demonstrating good robustness and having practical engineering application potential.

(4) The characteristic correlation analysis shows that the frequency of the working fluid pump is significantly correlated with variables such as the flow rate of molten salt, absorbed thermal power, and outlet temperature, but has a weak correlation with variables such as the opening degree of the inlet regulating valve and the frequency of the cold tank pump, and can be excluded, simplifying the model input.

The inverse prediction model developed in this paper (predicting pump frequency from state variables such as outlet temperature) provides an effective tool for sensor redundancy design and condition monitoring of solar thermal collector systems. It is particularly suitable for outlet temperature setpoint tracking and pump frequency sensor fault diagnosis. Subsequent work will focus on the dynamic expansion of the model, multi-condition verification, and the integrated application with advanced control algorithms to further enhance the system's automation level and operational efficiency.

Acknowledgment

The authors are grateful to the National Natural Science Foundation of China Grant No. 51775010 for financial support.

Data availability

The raw/processed data required to reproduce these findings cannot be shared at this time due to technical or time limitations.

Funding

This research was funded by the National Natural Science Foundation of China Grant No. 51775010.

Author Contributions

All authors contributed to the study's conception and design. Hongxia Yan Conceptualization, Software, Writing–original draft; Jialan Sun: Supervision, Methodology, Visualization, Writing-review & editing. All authors read and approved the final manuscript.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

Ethical Approval The work is original and has not been published previously or is not under consideration for publication elsewhere.

Consent to Participate All authors have read and agreed to the submitted version of the manuscript.

References

[1] Bai J, Pan L, Hao J, et al. (2025) Key technology research progress of photovoltaic solar thermal collectors: Overview. Journal of Renewable and Sustainable Energy 17: 052702.

[2] Kurtoğlu M, Eroğlu F (2026) Current trends and challenges in solar PV-integrated battery energy storage technology: Key components, methods, and future prospects. Applied Energy 409: 127461.

[3] Fernández-García, A., Zarza, E., Valenzuela, L., et al. (2020). Parabolic-trough solar collectors and their applications. Renewable and Sustainable Energy Reviews, 14(7), 1695–1721.

[4] Fadzlin WA, Hasanuzzaman M, Rahman SA, et al. (2025) Solar thermal energy storage: global challenges, innovations, and future directions for renewable energy systems. Applied Thermal Engineering 280: 128346.

[5] Tang, J. F., Dong, J., Liang, L., et al. (2023). Research on temperature control of parabolic trough solar collector field based on stepwise predictive control. Thermal Power Generation, 52(9), 181–189.

[6] Camacho, E. F., Berenguel, M., Rubio, F. R., et al. (2022). Control of solar energy systems (pp. 89–124). London: Springer.

[7] Celik Toker, S., & Kizilkan, O. (2024). Development of dynamic thermal modeling for evacuated U-tube solar collectors. Arabian Journal for Science and Engineering, 49(2), 2345–2360.

[8] Hai, D. T. H., & Phu, N. M. (2023). A critical review of all mathematical models developed for solar air heater analysis. Journal of Advanced Research in Fluid Mechanics and Thermal Sciences, 105(1), 1–14.

[9] Wang, L., Zhang, Y., Liu, H., et al. (2025). Research on multi-model switching control of linear Fresnel heat collecting subsystem. Sustainability, 17(17), 7780.

[10] Wang, H., Li, M., & Zhao, J. (2022). Experimental study on factors affecting the efficiency of parabolic trough solar collector. Acta Energiae Solaris Sinica, 43(4), 301–308.

[11] Zhang, H., Li, Z. G., & Wang, Q. (2021). Experimental study and numerical simulation of molten salt freezing blockage. CIESC Journal, 72(5), 2456–2465.

[12] Liu, B., Wu, Y. T., & Ma, C. F. (2022). Numerical simulation of freezing blockage process in molten salt pipeline. Journal of Engineering Thermophysics, 43(2), 456–463.

[13] Zhou, X. M., Zhang, H., & Li, J. (2023). Anti-freezing strategy of molten salt in parabolic trough solar power plant. Acta Energiae Solaris Sinica, 44(1), 156–163.

[14] Chen, H. P., Yu, X., Zhou, T. L., et al. (2021). Prediction of outlet temperature of parabolic trough solar collector field based on PSO-LSSVM. Thermal Power Generation, 50(6), 127–133.

[15] Meng, F., Li, X., Wang, Y., et al. (2025). Application of selective ensemble learning soft sensor modeling based on GPR in the solar thermal power collection system. Energy, 289, 129876.

[16] Yan, L., Lei, D. Q., Li, X., et al. (2023). Research on outlet temperature prediction of parabolic trough solar collector field based on hybrid neural network. Acta Energiae Solaris Sinica, 44(5), 412–420.

[17] Camacho, E. F., Gallego, A. J., & Sánchez, A. J. (2022). Model predictive control for solar thermal power plants. Control Engineering Practice, 120, 105012.

[18] Stuetzle, T., Blair, N., Mitchell, J., et al. (2021). Automatic control of a 30 MW SEGS VI parabolic trough plant. Solar Energy, 215, 456–468.

[19] Gallego, A. J., & Camacho, E. F. (2023). Adaptive state-space model predictive control of a parabolic-trough field. Solar Energy, 251, 234–247.

[20] Zhang, Q., Li, R., & Wang, G. (2023). Prediction method for abnormal parameters of steam turbine based on random forest and LSTM. Journal of Engineering for Thermal Energy and Power, 38(1), 89–97.