Journal of Agricultural Engineering

https://www.agroengineering.org/

Field-based implementation of machine learning forecasting for energy-efficient temperature control in a greenhouse system

Sunyong Park,[1] DaeHyun Kim,[2] Kwang Cheol Oh[3]

[1]Forest Industrial Materials Division, National Institute of Forest Science, Seoul

[2]Department of Biosystems Engineering, Kangwon National University, Chuncheon-si

[2]Korea Research Institute on Climate Change, Chuncheon-si, Republic of Korea

Corresponding author: Kwang Cheol Oh,Korea Research Institute on Climate Change, 11 Subyeongongwon-gil, Chuncheon-si 24239, Republic of Korea. E- mail: okc@kric.re.kr

Publisher's Disclaimer

E-publishing ahead of print is increasingly important for the rapid dissemination of science. The *Early Access* service lets users access peer-reviewed articles well before print/regular issue publication, significantly reducing the time it takes for critical findings to reach the research community.
These articles are searchable and citable by their DOI (Digital Object Identifier).

Our Journal is, therefore, e-publishing PDF files of an early version of manuscripts that undergone a regular peer review and have been accepted for publication, but have not been through the typesetting, pagination and proofreading processes, which may lead to differences between this version and the final one.
The final version of the manuscript will then appear on a regular issue of the journal.

Please cite this article as doi: 10.4081/jae.2026.2199

©The Author(s), 2026
Licensee PAGEPress, Italy

Submitted: 24 March 2026
Accepted: 25 May 2026

***Note**: The publisher is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries should be directed to the corresponding author for the article.*

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Field-based implementation of machine learning forecasting for energy-efficient temperature control in a greenhouse system

Sunyong Park,[1] DaeHyun Kim,[2] Kwang Cheol Oh[3]

[1]Forest Industrial Materials Division, National Institute of Forest Science, Seoul
[2]Department of Biosystems Engineering, Kangwon National University, Chuncheon-si
[2]Korea Research Institute on Climate Change, Chuncheon-si, Republic of Korea

Corresponding author: Kwang Cheol Oh, Korea Research Institute on Climate Change, 11, Subyeongongwon-gil, Chuncheon-si 24239, Republic of Korea. E- mail: okc@kric.re.kr

Contributions: Sunyong Park, conceptualization, data curation, and writing—original draft; **Dae Hyun Kim**, methodology development, model implementation, and formal analysis; **Kwang Cheol Oh**, supervision, research design, validation, project administration, and writing—review and editing. All authors read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare no competing interests.

Funding: This work was supported by the National Research Foundation of Korea [grant number NRF- NR060130 and 2022R1C1C2009821].

Abstract

Greenhouse heating systems require control strategies that reduce unnecessary energy use while maintaining crop-safe temperatures under dynamic winter conditions. This study evaluated the practical application of machine learning (ML)-based forecasting for greenhouse temperature control in a single winter greenhouse equipped with a pellet boiler, heat-storage tank, and tube-rail heating system. Data were collected for 55 d under real operating conditions, including internal environmental variables and external weather variables such as external temperature, humidity, dew point, and solar radiation. Prediction and forecasting models were developed using time-series ML algorithms with gap labeling, and iterative learning was conducted using 7 d of training data and 1 d of validation data. The developed models showed an average r^2 of 0.77 and RMSE of 1.43°C across the evaluated cases. The

forecasting results were then applied to a data-driven predictive control (DDPC) strategy to assess whether heating supply could be stopped earlier than in the conventional on/off control scenario. Compared with the baseline control scenario, the DDPC-based strategy showed an estimated 5-15% reduction in heating supply operation time. This value should be interpreted as an operation-time-based estimate under the tested pellet-boiler greenhouse conditions, rather than as a direct measurement of pellet-fuel consumption. The results indicate that 30- and 60-min forecasting horizons can support practical heating-control decisions, whereas longer horizons may increase the risk of low-temperature stress. Therefore, this study emphasizes the field-based implementation of ML forecasting for real-time greenhouse control and clarifies its operational limitations under actual winter conditions.

Keywords: Smart greenhouse; forecasting; machine learning; data-driven predictive control; energy efficiency

Introduction

The increasing global population and climate change pose significant challenges to food security, requiring substantial improvements in agricultural productivity and energy efficiency. On November 15, 2022, The United Nations announced that the global population had reached 8 billion and was expected to grow to approximately 9.7 billion by 2050, necessitating an increase of more than 70% in food production over current levels (da Rosa Righi *et al.*, 2020; Pardey *et al.*, 2014). This significant challenge demands urgent and innovative solutions to satisfy the increasing food demand while addressing the pressing concerns of climate change and resource constraints. Therefore, there is a demand for automated technologies to execute effective agricultural operations and improve crop productivity and energy consumption. The application of automation technology in conventional field farming is challenging because of environmental variability, high costs, and technical constraints. Therefore, automation technologies are continuously being developed and used in controlled cultivation systems. Greenhouses offer a controlled environment that mitigates the unpredictability of traditional farming, providing a viable solution for enhancing productivity and crop resilience amid climate change.

Greenhouses maintain stable production owing to their limited exposure to the external environment; however, they consume a considerable amount of energy. Energy is required for

cooling during summer and heating during winter, which increases crop production costs and causes environmental pollution. With the rising energy consumption in greenhouses, there is an urgent need for efficient and sustainable energy management practices. Without such measures, the environmental and economic costs of greenhouse agriculture could outweigh its benefits and jeopardize the long-term viability of these systems. Hence, it is necessary to develop energy optimization technologies to operate controlled agricultural greenhouses efficiently. Controlled agriculture demands more energy than conventional open-field farming, making it imperative to implement effective environmental controls to reduce energy usage. An accurate prediction of the internal conditions of a greenhouse is essential for this purpose, and various simulation studies have been conducted in this regard. Developing advanced predictive models is essential not only for optimizing energy usage but also for maintaining a balance between crop productivity and sustainability.

In terms of simulation modeling for prediction and forecasting, Conventional simulation approaches, including analytical models and computational fluid dynamics, are effective for steady-state analysis but have limitations in capturing dynamic environmental changes, necessitating more advanced data-driven approaches. However, with recent advancements in computer technology, three-dimensional simulation techniques have emerged (e.g., computational fluid dynamics), allowing for more accurate predictions of environmental changes based on actual internal configurations and boundary conditions (Cemek *et al.*, 2017; Chen *et al.*, 2015; Guzmán *et al.*, 2019; Kim *et al.*, 2017; Outanoute *et al.*, 2015). These simulation models can be used to analyze phenomena based on set boundary conditions in a steady state, enabling the evaluation of systems without considering time-dependent effects. However, they have limitations in predicting transient states. Consequently, steady-state simulation models are primarily used for initial design reviews or simple assessments of internal environments. Therefore, advanced analytical methods are required for complex dynamic situations and unexpected environmental predictions. Steady- and transient-state conditions use similar but distinct concepts for prediction and forecasting. Prediction focuses on estimating the current state, whereas forecasting aims to estimate future states (Runge and Saloux, 2023; Runge and Zmeureanu, 2019). Prediction models perform analyses based on theoretical models and existing simulations according to boundary conditions and are used for design modifications and system configurations. By contrast, forecasting models based on current data and data characteristics enable real-time control and predictive maintenance. Optimization studies have been conducted based on these differences, and detailed distinctions between prediction and forecasting have been investigated. As mentioned previously, when

developing a machine learning (ML) simulation model using time-series data, differences between the prediction and forecasting models arise depending on the data labeling method. Prediction models are used to predict the internal state at the current time, whereas forecasting models predict future states and can be used as a basis for control.

In terms of optimal control systems for energy reduction, conventional control systems manage each device individually, and operators manually adjust them, leading to excessive energy use and increased overall energy consumption (Kadlec and Gabrys, 2009). To address these issues, intelligent integrated management systems have been developed (e.g., building energy management systems). These systems, through model predictive control (MPC), have been demonstrated to save 15-40% of building energy usage (Choi and Lee, 2023; Lee *et al.*, 2022; Serale *et al.*, 2018; Yao and Shekhar, 2021; Zhang *et al.*, 2018). Moreover, to overcome the limitations of MPC, data-driven predictive control (DDPC) can be used to achieve more effective energy optimization (Deng *et al.*, 2023; López Santos *et al.*, 2023). The most representative time-series analysis model for predictive control is the autoregressive integrated moving average (ARIMA) model. This model is useful for analyzing data that change over time to predict future variations (Dahl *et al.*, 2017; Frausto *et al.*, 2003; Song *et al.*, 2021). The ARIMA model has been extensively used in research to precisely predict internal environments through environmental control and improve agricultural production (Lv *et al.*, 2018; Wu *et al.*, 2023). Recently, the Prophet model, which was enhanced using statistical techniques, has also been employed. Given that model performance varies depending on the data characteristics and analysis methods, efforts to develop an optimal model are ongoing (Wang *et al.*, 2023). Building systems are relatively less affected by the external environment, whereas controlled agriculture is associated with challenges owing to additional variables (e.g., solar radiation, external temperature, external humidity, and crop characteristics). Therefore, it is difficult to apply predeveloped models, and hybrid intelligent control systems have been proposed (Wang *et al.*, 1997). The results of various studies on optimization control have enhanced energy efficiency by enabling precise predictions of the internal greenhouse environment (Jung *et al.*, 2020; Venkateswaran and Cho, 2024). With recent advancements in artificial intelligence (AI), more precise predictive models essential for smart agricultural systems are being developed (Kumari and Toshniwal, 2021). AI-based predictive models can provide more accurate forecasts by considering various environmental variables, thereby optimizing greenhouse heating and cooling systems.

Anticipating environmental changes is essential for reducing energy consumption in controlled agriculture. Forecasting models based on real-time data enable proactive control of

greenhouse environments, improving both energy efficiency and operational performance. However, most previous studies have primarily focused on model accuracy, with limited attention to how forecasting results can be integrated into actual greenhouse heating-control decisions. The research gap addressed in this study is the field-based implementation of forecasting-based control in a real pellet-boiler greenhouse system, rather than the proposal of a new ML architecture. Specifically, this study evaluates whether short-term ML forecasting can support earlier heating-supply decisions under the operational constraints of a pellet boiler, heat-storage tank, and tube-rail heating system. Accordingly, the objective of this study is to clarify the practical applicability and limitations of DDPC for greenhouse heating operation under actual winter conditions. A preprint version of this manuscript has been previously published in SSRN (Oh et al., 2024a).

Materials and Methods

Figure 1 illustrates the development and optimization process for the internal environment temperature prediction and forecasting model proposed in this study. The specific details are as follows.

a. Data mining

Data were collected from a greenhouse system, and characteristic variables were extracted to identify those with the highest correlation coefficients.

b. Data splitting

The selected data were normalized and divided into training and testing datasets.

c. Algorithm selection

The models were developed using time-series algorithms (e.g., recurrent neural networks [RNN], long short-term memory [LSTM], and gated recurrent units [GRU]).

d. Development of prediction models and optimization

Prediction models were developed, and the hyperparameters were optimized through a grid search, aiming for a coefficient of determination of 0.9.

e. Development of forecasting model and evaluation

The application strategies of the forecasting model were analyzed based on the prediction model, and the baseline energy reduction was estimated.

Data processing and analysis

Data collection

This study was conducted in a greenhouse at Kangwon National University (300 m², Hyoja-dong 192-1, Chuncheon-si, Gangwon-do). A heating system using renewable energy was introduced to maintain the internal temperature of the greenhouse during winter (Figure 2).

The heat energy supply device used was a typical domestic pellet boiler (KN-23D, KYUWON Co., Korea) with a maximum heating output of 23,000 kcal/h (Figure 3a). Various issues arise in field systems that supply energy through renewable sources. For pellet boilers, the variable characteristics of solid fuel make it difficult to provide a stable heat source. To address this issue, a heat storage tank (Figure 3b) was added as a buffer to ensure a stable heat supply. The internal environment temperature was controlled with an on/off mechanism set between 17°C and 20°C. The final supply to the greenhouse was conducted using a tube–rail system (Figure 3c). The tube rail, made of galvanized steel, was designed to transfer heat through radiant energy with the heating fluid flowing inside the pipe. In addition, it can be used as a track for smart-farm automatic transport robots and harvesting devices. The maximum heat dissipation capacity of the tube rail was 71 W/m at a hot water temperature of 55°C, and it was designed with a length of 124 m and a diameter of 41 mm.

The heat source supply method is determined by the internal temperature of the environment and the target temperature. It must have a high heating capacity relative to the required heat load, considering the greenhouse insulation efficiency, external temperature, and weather conditions specific to each region. In addition, excessively high or low heat sources in specific areas can cause problems with crop growth. Therefore, various heating systems are required, depending on the location of the smart farm and the characteristics of the crops being cultivated. The internal environmental temperature for prediction was measured at a height of 1 m above the ground using a thermocouple (GTPK-02-17) (Figure 3c). External meteorological data (external temperature, external humidity, solar radiation, and internal temperature) were measured using a weather observation device (JNGW100, JINONG) installed outside the greenhouse. The recorded data were logged at 1-min intervals using a data logger (GL840, GRAPHTEC Co., Japan) (Table 1).

Baseline control scenario and estimation of heating-operation reduction

The baseline scenario was defined as the conventional on/off heating control used in the experimental greenhouse. In this system, the pellet boiler, heat-storage tank, and tube-rail heating system were operated to maintain the internal air temperature within the target range

of 17-20°C. The DDPC scenario was evaluated by using the forecasting model to determine whether heating supply could be stopped earlier before the internal temperature reached the upper set-point.

The reduction effect was evaluated using the difference in heating supply operation time between the conventional control model and the DDPC-based control model. Because direct pellet-fuel consumption was not measured separately for each control scenario, the estimated reduction should be interpreted as a heating-operation-time-based indicator rather than as a direct fuel-consumption measurement.

Extracting characteristic variables

Data characteristic analysis is an essential process for identifying the key variables that affect the internal environment of a greenhouse and for investigating the statistical relationships between the collected data. We used Pearson's correlation coefficient method for the correlation coefficient calculations, which is useful for numerically representing the strength and direction of the linear relationship between variables. Pearson's correlation coefficient can be calculated as follows (Eq. 1):

$$r = \Sigma[(x\_i - \bar{x})(y\_i - \bar{y})] / \sqrt{\{\Sigma(x\_i - \bar{x})^2 \Sigma(y\_i - \bar{y})^2\}} \quad \text{(Eq. 1)}$$

where: y_i, observed values of each variable; $\bar{x}$, $\bar{y}$, mean values of each variable; r, correlation coefficient.

A correlation coefficient, r, value close to +1 indicates a strong positive linear relationship, whereas a value close to −1 indicates a strong negative linear relationship. A value near zero indicates a lack of a linear relationship between the two variables. Using this formula, the analysis results clearly demonstrate the correlation characteristics between the variables. Selecting variables with high correlations can improve the prediction accuracy of the model and maximize the data processing efficiency by eliminating unnecessary variables.

Data preprocessing

During the preprocessing stage, the selected data underwent data reduction, including outlier analysis and dimensionality reduction. During the data cleansing process, the format of the data was standardized, missing values or gaps were handled, and unnecessary values were either removed or imputed to improve data quality. Generally, noisy data are removed using imputation, regression, or clustering. However, for the greenhouse system under analysis, which is strongly influenced by natural phenomena, the basis for determining the outliers is insufficient. Therefore, the dataset was constructed using all collected information, and any

missing values resulting from various errors were handled by removing the corresponding data points (rows) using the list-wise deletion method (Abasilim *et al.*, 2024). Furthermore, using the selected data as is could impact model training owing to the large absolute value differences between the input data. Therefore, the analysis was performed using normalization.

Data preprocessing methods encompass two main techniques: feature normalization and mean centering. Mean centering adjusts data by shifting it around its mean to facilitate learning and reduce certain types of bias effects. However, it is highly vulnerable to the influence of extreme values and outliers (Kim *et al.*, 2017). Conversely, feature normalization involves dividing each feature value by its standard deviation. This approach enhances model performance, ensuring consistency even with varying training data sizes (Aggarwal, 2018) (Eq. 2). Hence, in this study, where the connection between data collections was crucial, analysis was performed using feature normalization (Chang *et al.*, 2019). This step is essential for enhancing the accuracy of the analysis and optimizing the model performance.

$$x'_{ij} = (x_{ij} - x_{min,j}) / (x_{max,j} - x_{min,j}) \quad \text{(Eq. 2)}$$

where: x'_{ij}, standardized index value; $x_{min,j}$, value of the *j*-th dimension of the *i*-th data; $x_{max,j}$, maximum value of the *j*-th feature.

Split data training and testing

To enhance the reliability of an ML model, it is crucial to separate the training data from the validation data. Based on previous research results (Oh *et al.*, 2022), we developed a model using 8-day training data (11,520 data points). For each predictive model, 7 d (10,080 data points) of data were used for training, and 1-day training data (1,440 data points) were used for model verification (testing). Subsequently, a model was developed sequentially using the remaining data.

Algorithm selection

To effectively use normalized data through preprocessing, it is essential to select an analysis method that suits the characteristics of the target system. The internal environment of a greenhouse changes over time and is influenced by real-time external conditions, resulting in the appearance of various characteristics in the data. Therefore, the data collected over time were assumed to maintain a certain level of stationarity, and based on this assumption, a greenhouse internal temperature prediction model was developed using artificial neural networks.

Studies on supervised learning using regression analysis have been widely conducted (Ariga

et al., 2018; Aggarwal, 2018; Huh, 2017; Muller and Guido, 2017). RNNs are fundamentally suitable algorithms for analyzing and predicting temporally continuous data. RNNs can model the temporal dependencies between inputs by maintaining an internal state (memory). This is particularly useful when data characteristics are interdependent in chronological order, such as in time-series data. In addition, because RNNs process data sequences step-by-step and pass the internal state to the next step, they can learn the underlying patterns of sequential data (e.g., trends or seasonality). By learning the complex relationships between data points, RNNs can predict future environmental changes.

Model structuring method

Prediction and forecasting models

Figure 4a illustrates the data labeling method used for the prediction models. The training data included past data used to train the model (e.g., time-series data of external temperature, external humidity, radiation, and internal temperature labeled within the same timeframe). These data were used by the prediction model to estimate the current internal state. Figure 4b illustrates the data-labeling method applied to the forecasting models. In addition to past data, the training dataset includes "forecasting data" labels to predict future states. The "Gap labeling" that occurs here is crucial, as it allows the forecasting model to consider the time interval between the present and the future. Finally, the model was developed using data from the previous 7 d to predict and forecast up to one day ahead.

Hyper-parameters

Existing simulation models exhibit characteristics in which the internal flow changes based on boundary condition variations with each data transmission method (e.g., nodes or meshes), as determined by finite element analysis (FEA) or the finite element method (FEM). Therefore, the final convergence state reached by the actual values owing to changes in the boundary conditions determines the prediction results. The number of iterations varied according to the actual boundary conditions and meshes that drove the simulation.

ML iteration learning is defined as follows and is related to the concept of hyperparameter epochs (Oh *et al*., 2024b) (Eq. 3):

$$\text{Iteration} = N / B \quad \text{(Eq. 3)}$$

An epoch is defined as the state in which learning over the entire dataset is complete (Eq. 4).

$$\text{Epoch} = \text{one complete pass through the full training dataset} \quad \text{(Eq. 4)}$$

The performance of an ML model can vary depending on the number of epochs (Lederrey *et*

al., 2021; Muller and Guido, 2017). To clearly define the learning characteristics, model optimization was performed with the delta value of the early stopping function set to 10^{-4}. Table 2 lists the final model development conditions.

The importance of the parameters and hyperparameters is crucial in the training process of an ML model. The parameters are updated during the training process, whereas hyperparameters (e.g., learning rate and regularization) are used to adjust the training process. Several methods, including grid, manual, random, and Bayesian searches, can be used to tune the hyperparameters. In this study, a grid search method was employed for model development (Aggarwal, 2018). The Adam optimization function and MSE loss function were used for model optimization. The Adam optimizer was set at an initial learning rate of 0.001 to optimize the weights and biases of the model. The MSE loss function was used to minimize the difference between the predicted and actual values (Smith *et al.*, 2017).

Therefore, the ML model can learn the correlations between the input and output data, allowing for the optimization of the algorithm selection and hyperparameters. The model development process in this study was conducted using Python-based deep-learning libraries, TensorFlow, and the Keras interface. To ensure model stability, the Seed() function was used, and appropriate iterations were performed to prevent underfitting and overfitting (Oh *et al.*, 2024b). The ML model developed in this study was applied to predict the internal environment of a greenhouse with high accuracy and precision. This contributes to the effective modeling and prediction of the characteristics of real-time changing data.

Model evaluation

The correlation between the experimental and simulation results was verified using the coefficient of determination (r^2) (Eq. 5), and the optimal model was selected using the root-mean-square error (RMSE) (Eq. 6) (Oh *et al.*, 2019).

$$R^2 = 1 - [\Sigma(Yi - Xi)^2 / \Sigma(Yi - \bar{Y})^2] \quad \text{(Eq. 5)}$$

where: Y_i, i-th measured value; X_i, i-th estimated value; r2, coefficient of correlation

$$RMSE = \sqrt{[(1/n)\,\Sigma(Xi - Yi)^2]} \quad \text{(Eq. 6)}$$

where: d_i, difference between the i-th estimated and i-th measured values; Y_i, i-th measured value; RMSE, root-mean-square error.

Statistical analysis

The Wilcoxon signed-rank test was applied to evaluate whether the performance differences between the models were statistically significant. This non-parametric test was selected because it does not assume normality and is suitable for paired comparisons of model performance metrics.

Results

Extracting characteristic variables

Figure 5 illustrates the correlation analysis results of the various external environmental data collected over 55 d and the internal greenhouse temperatures in this study.

First, there was a strong positive correlation between the internal and external temperatures. This was attributed to the significant heat transfer between the inside and outside of the greenhouse, which was constructed using thin vinyl or glass materials to ensure sufficient solar radiation transmission for crop growth. In addition, the heat source supplied from the storage tank to the greenhouse demonstrated a high correlation with the internal temperature, as it directly controlled the internal temperature. It was initially expected that solar radiation would have the most significant impact on internal temperature changes. However, the relatively low correlation observed was likely attributable to the inclusion of nighttime data when there was no solar radiation.

Internal and external humidity demonstrated a negative correlation with internal temperature, which is attributed to the temperature-reduction characteristics caused by moisture evaporation. Moreover, during the winter experiments, no ventilation was used, except at the entry and exit for internal environment checks, resulting in a low correlation with the wind speed or direction, as these factors do not affect the internal temperature. The rainfall sensor data, which were binary (zero and one), demonstrated a low correlation with the internal temperature.

Finally, the irrigation and drainage amounts related to plant growth demonstrated a correlation coefficient close to zero, which was likely due to the relatively small amounts supplied relative to the greenhouse system scale. However, the medium temperature exhibited a high correlation, which was attributed to the heating effect of solar radiation and rapid heat supply from the nearby tube rail.

Based on the final correlation analysis results and the physical relevance of greenhouse heat gain, the internal temperature, external temperature, humidity, dew point, solar radiation, medium temperature, storage tank inlet and outlet temperatures, and internal humidity data were selected to develop the subsequent ML model. Although the overall Pearson correlation between solar radiation and internal air temperature was lower than initially expected, this was mainly attributed to the inclusion of nighttime data, during which solar radiation was zero. Therefore, solar radiation was retained as an input variable for subsequent model development because of its physical relevance to daytime greenhouse heat gain.

Algorithm selection

The RNN algorithm uses the tangent hyperbolic (tanh) activation function to determine the current input and previous hidden state. This information was passed on to the next hidden unit and finally to the output layer to predict the internal temperature. By repeatedly transmitting the previous hidden state to the next time step in each stage, the model can capture the data flow over time. However, as the data sequence lengthens, the RNN might suffer from a vanishing gradient problem, which can reduce the accuracy of the model. To address this issue, LSTM and GRU algorithms were proposed.

The LSTM algorithm includes input, forget, and output gates that are specifically designed to effectively learn long-term dependencies by updating the cell state (Hochreiter and Schmidhuber, 1997; Wei *et al.*, 2019). It combines the information extracted from the input data and the previous cell state to generate new cell and hidden states, which were processed through a dense layer to predict the internal temperature. This approach helps mitigate the vanishing gradient problem (Chang *et al.*, 2019; François, 2017).

Finally, the GRU is a simplified version of LSTM that uses only two gates: update and reset. These gates use sigmoid and tanh activation functions to calculate the new hidden state. By efficiently combining the previous hidden state and input, the GRU discards unnecessary information while retaining essential information, thus enhancing computational efficiency and reducing computational load (Choi *et al.*, 2012; He *et al.*, 2022). An internal temperature prediction model was developed using these algorithms. The selected models were evaluated not only in terms of predictive accuracy but also considering computational efficiency for real-time control applications.

Prediction and forecasting model results

Based on the results of the correlation analysis of the data collected over 55 d, models were

developed using the selected variables. Each model was developed over an 8 d cycle, with 7 d used for training and the last day for validation. A total of 48 models were developed for each algorithm type, with each model including one prediction model at the 0-min interval and three forecasting models at the 30-, 60-, and 120-min intervals, resulting in 192 individual models for the performance evaluation. In ML models, parameter weights are updated through iterations to improve model performance (Muller and Guido, 2017).

All the models were trained under an early stopping condition, in which training was halted if the change in the loss value was less than 10^{-4}. The number of epochs and performance metrics (r^2, RMSE) for each model were derived. The average values of the development characteristics (epochs, average accuracy, and precision) of all models according to the algorithm and time interval are presented (Table 3).

All the algorithm results demonstrated that as the time interval increased, the r^2 values decreased and the RMSE values increased, indicating a decline in the model performance. The number of epochs also increased, probably because of the growing complexity of the temporal relationships in the data as the time interval widened, necessitating additional computations to bridge the information gap. These computations are essential for accurately modeling the data relationships between different time points and maintaining prediction accuracy. Consequently, more epochs were performed to learn the flow and changes in information over longer intervals.

Overall, the differences in r^2 and RMSE among the models were small across all forecasting intervals, indicating comparable predictive capability.

All *p*-values were greater than 0.05, indicating no statistically significant differences between the models in terms of their predictive performance. Therefore, model selection should consider computational efficiency rather than marginal-performance differences. Accordingly, the GRU algorithm required fewer training epochs, indicating higher computational efficiency, and is thus the most practical choice.

Figure 7 illustrates the results of the internal temperature prediction for a specific date (January 5, 2023) using different algorithms. The results for the 0-min time interval shown in Figure 7a demonstrate a high predictive performance, rendering it effective for real-time internal environment changes. The prediction models closely matched the experimental data, especially during peak and trough periods, indicating a strong ability to capture internal temperature dynamics, despite minor deviations.

As 0-min interval models are designed for real-time analysis, their practical application in control systems is limited owing to their inability to effectively predict future changes. By

contrast, the forecasting models at the 30-, 60-, and 120-min intervals demonstrated a higher potential for application in real greenhouse environments (Figure 6 a-c). Forecasting models are more useful for maintaining internal environment plans and control purposes because of their ability to handle longer time intervals and can be applied to a broader range of applications.

The forecasting model results for the 30-, 60-, and 120-min intervals indicated an increasing prediction error with an increase in the time interval. Notably, between 6 AM and 12 PM, when the temperature increased sharply, all models closely followed the actual temperature changes. However, with an increase in the time interval, discrepancies between the predicted and actual data became evident. In particular, the 30- and 0-min forecasting models predicted a faster temperature increase than the actual data, whereas the 120-min model predicted a decrease. During the temperature drop between 3 p.m. and 6 p.m., the accuracy and precision of the models determined their consistency with the actual data. At night, all the models except the 120-min forecasting model matched the actual data well. This enables earlier adjustment of heating operation, contributing to more efficient energy use in practical greenhouse environments.

From an agricultural engineering perspective, the forecasting horizon should be selected by balancing proactive heating control and crop-temperature safety. The average RMSE observed in the field data is acceptable for screening practical control strategies, but the error becomes more important when the forecast is used to stop heating before the set-point is reached. In particular, longer forecasting horizons can accumulate temporal error and may cause excessive interruption of heat supply during rapid temperature changes or nighttime cooling periods. Therefore, the 120-min horizon should be regarded as a model-failure or high-risk case for practical heating control, whereas the 30- and 60-min horizons provide a more reasonable compromise between operational energy reduction and crop safety.

The RNN demonstrated a notable performance decline at longer time intervals because of its short-term memory limitations, whereas the LSTM and GRU effectively handled long-term dependencies and maintained a more stable performance. In particular, the GRU exhibited performance comparable to that of the LSTM at intermediate time intervals (60 and 120 min) while requiring fewer training epochs. Considering computational efficiency, one model demonstrated advantages in real-time control applications.

Heating operation reduction characteristics

Figures 7 to 9 illustrate the results of the DDPC using various time-lagged prediction models

based on different algorithms for January 5, 2023. The left axis illustrates the internal temperature based on the experimental, predicted, and control values, whereas the right axis demonstrates the heating supply curve from the tank to the greenhouse, normalized between zero and one, and represented by a dotted line. This normalization process ensures consistent results across various environmental conditions because the actual temperature supplied from the pellet boiler to the storage tank varies daily depending on the surrounding conditions.

The prediction model in Figure 7a, which uses the RNN algorithm, exhibits high accuracy and precision, with an r^2 value of 0.9971 and RMSE of 0.1983. However, prediction models cannot be used for control because they do not forecast future environments. Figure 7 b-d show the results of the forecasting models using the gap-labeling technique with time delays. The r^2 value decreased with an increase in the time delay in each graph, and the RMSE increased, indicating a decline in the forecasting model performance. This trend is typical, as longer forecasting times generally reduce model performance, which is a common phenomenon in various future prediction models (Wang *et al.*, 2024). Figures 8 and 9, which present the results obtained using the LSTM and GRU algorithms, respectively, also exhibit similar trends. As the time delay increased, both models demonstrated a decline in accuracy, highlighting the common challenge of sustaining prediction accuracy over extended forecasting horizons (Graves, 2012; Liu *et al.*, 2024).

Each model forecasts the future based on the shifted time interval, allowing for control data usage to determine the time required to reach a control temperature of 20°C. Specifically, in Figure 7 b-d, Figure 8 b-d, and Figure 9 b-d, the forecasts are 30-, 60-, and 120-min ahead, respectively, facilitating a pre-emptive cessation of the heating energy supply. The newly controlled heating energy supply curve is represented by the green dotted line, and the final internal environmental temperature is indicated by the green line. The difference between the forecasted and optimally controlled temperature curves is indicated in yellow.

The practical applicability of the GRU model and the resulting energy-saving characteristics are discussed (Figure 9). The model in, in which the energy supply was halted 30 min earlier, exhibited minimal energy-saving effects (Figure 9b). The 60-min forecasting model shown in Figure 7c delayed the internal temperature rise and reduced the peak temperature by about 2°C. However, the 120-min forecasting model shown in Figure 7d had the lowest accuracy and precision because of the extended forecasting time, leading to a prolonged halt in the heat supply. Although it forecasted the time to reach 20°C (the temperature at which the heating system was switched off), similar to the other models, the predicted time appeared earlier, and the controlled result showed a drop in the internal temperature to 12°C, making it unsuitable

for maintaining a conducive growth environment. Therefore, a maximum forecasting time of 60 min was deemed appropriate for DDPC. The observed reduction in heating operation time reflects a decrease in unnecessary energy usage, indicating the improved operational efficiency of the greenhouse system under real conditions.

Estimated heating-operation reduction

The reduction characteristics of the greenhouse heating operation were evaluated using the difference in heating supply operation time between the conventional control model and the DDPC-based control model. This comparison was conducted under the same measured environmental conditions, and the conventional on/off control scenario was used as the baseline. The heating supply time difference was calculated as follows:

$$EST_D = EST_{CCM} - ESTOCM \quad \text{(Eq. 7)}$$

Where: EST_D is the energy supply time difference, EST_{CCM} is the energy supply time under the conventional control model, and EST_{OCM} is the energy supply time under the DDPC-based control model.

If EST_D is positive, the DDPC-based control model reduces heating supply operation time relative to the conventional baseline. Therefore, EST_D was used as a practical operation-based indicator of potential energy reduction. The operation-time-based reduction ratio was calculated as follows:

$$\text{Heating-operation reduction ratio (\%)} = (ESTD / EST_{CCM}) \times 100 \quad \text{(Eq. 8)}$$

This ratio should be interpreted as an estimated reduction in heating supply operation time rather than a direct measurement of fuel or thermal energy consumption. The daily reduction ratio varied depending on the heating demand, outdoor conditions, and the forecasting horizon. In some cases, the difference converged to zero because the internal temperature did not fall below the lower set-point, and heating was not activated. In other cases, the DDPC result changed the timing of heat-supply interruption and reduced unnecessary operation time.

The results indicate that the 30- and 60-min forecasting horizons are more suitable for practical greenhouse heating control. The 30-min horizon provided conservative control with a relatively small reduction effect, whereas the 60-min horizon provided a more meaningful reduction while maintaining a safer temperature range. In contrast, the 120-min horizon increased the risk of excessive heating interruption and low-temperature stress. Accordingly, the reported 5-15% value is presented as a potential reduction in heating supply operation time under the tested pellet-boiler greenhouse conditions, not as a directly verified fuel-consumption saving.

Conclusions

These findings highlight the practical value of integrating short-term forecasting models into real-time greenhouse heating-control strategies. This study does not aim to introduce a novel ML algorithm; rather, it demonstrates the field-based implementation of forecasting-based control in a real greenhouse equipped with a pellet boiler, heat-storage tank, and tube-rail heating system. The results indicate that 30- and 60-min forecasting horizons can provide a reasonable balance between proactive heating control and crop-temperature safety, whereas longer horizons such as 120 min may increase the risk of excessive heating interruption and low-temperature stress. The proposed DDPC approach showed a potential 5%-15% reduction in heating supply operation time compared with the conventional on/off control scenario. However, this value should be interpreted as an operation-time-based estimate obtained from a single winter pellet-boiler greenhouse, rather than as a direct measurement of pellet-fuel consumption. For practical greenhouse operation, the main takeaway is that forecasting horizons should be selected conservatively to avoid crop damage while reducing unnecessary heating operation. Further validation using direct fuel-consumption measurements, multiple greenhouse types, different crops, and diverse climatic conditions is required before broader generalization.

References

Abasilim C, Friedman LS, Martin MC, Madigan D, Perez J, Morera M, et al., 2024. Risk factors associated with indicators of dehydration among migrant farmworkers. Environ Res 251:118633.

Aggarwal CC, 2018. Neural networks and deep learning: a textbook. Cham, Springer.

Ariga M, Nakayama S, Nishibayasi D, 2018. Machine learning at work. Sebastopol, O'Reilly Media.

Cemek B, Atiş A, Küçüktopçu E, 2017. Evaluation of temperature distribution in different greenhouse models using computational fluid dynamics (CFD). J Agric Sci 32:54-54.

Chang Z, Zhang Y, Chen W, 2019. Electricity price prediction based on hybrid model of adam optimized LSTM neural network and wavelet transform. Energy 187:115804.

Chen J, Xu F, Tan D, Shen Z, Zhang L, Ai Q, 2015. A control method for agricultural greenhouses heating based on computational fluid dynamics and energy prediction model. Appl Energy 141:106-118.

Choi W, Lee S, 2023. Performance evaluation of deep learning architectures for load and temperature forecasting under dataset size constraints and seasonality. Energy Build 288:113027.

Choi YS, Lee HJ, Joung ST, 2012. A design and implementation of web-based green house automation system. J Korea Inst Electron Commun Sci 7:1519-1527.

da Rosa Righi R, Goldschmidt G, Kunst R, Deon C, da Costa CA, 2020. Towards combining

data prediction and internet of things to manage milk production on dairy cows. Comput Electron Agric 169:105156.
Dahl M, Brun A, Andresen GB, 2017. Using ensemble weather predictions in district heating operation and load forecasting. Appl Energy 193:455-465.
Deng Z, Wang X, Jiang Z, Zhou N, Ge H, Dong B, 2023. Evaluation of deploying data-driven predictive controls in buildings on a large scale for greenhouse gas emission reduction. Energy 270:126934.
François C, 2017. Deep learning with Python. Shelter Island, Manning Publications.
Frausto HU, Pieters JG, Deltour JM, 2003. Modelling greenhouse temperature by means of auto regressive models. Biosyst Eng 84:147-157.
Graves A, 2012. Long short-term memory. In: Supervised sequence labelling with recurrent neural networks. Berlin, Springer pp. 37-45.
Guzmán CH, Carrera JL, Durán HA, Berumen J, Ortiz AA, Guirette OA, et al., 2019. Implementation of virtual sensors for monitoring temperature in greenhouses using CFD and control. Sensors (Basel) 19:60.
He Z, Jiang T, Jiang Y, Luo Q, Chen S, Gong K, et al., 2022. Gated recurrent unit models outperform other machine learning models in prediction of minimum temperature in greenhouse based on local weather data. Comput Electron Agric 202:107416.
Hochreiter S, Schmidhuber J, 1997. Long short-term memory. Neural Comput 9:1735-1780.
Huh M-H, 2017. Representing variables in the latent space. Korean J Appl Stat 30:555-566.
Jung D-H, Kim HS, Jhin C, Kim H-J, Park SH, 2020. Time-serial analysis of deep neural network models for prediction of climatic conditions inside a greenhouse. Comput Electron Agric 173:105402.
Kadlec P, Gabrys B, 2009. Soft sensors: where are we and what are the current and future challenges? IFAC Proc 42:572-577.
Kim R-W, Hong S-W, Lee I-B, Kwon K-S, 2017. Evaluation of wind pressure acting on multi-span greenhouses using CFD technique, Part 2: Application of the CFD model. Biosyst Eng 164:257-280.
Kumari P, Toshniwal D, 2021. Long short term memory–convolutional neural network based deep hybrid approach for solar irradiance forecasting. Appl Energy 295:117061.
Kwon H, Oh KC, Choi Y, Chung YG, Kim J, 2021. Development and application of machine learning-based prediction model for distillation column. Int J Intell Syst 36:1970-1997.
Lederrey G, Lurkin V, Hillel T, Bierlaire M, 2021. Estimation of discrete choice models with hybrid stochastic adaptive batch size algorithms. J Choice Modell 38:100226.
Lee D, Ooka R, Matsuda Y, Ikeda S, Choi W, 2022. Experimental analysis of artificial intelligence-based model predictive control for thermal energy storage under different cooling load conditions. Sustain Cities Soc 79:103700.
Liu G, Zhong K, Li H, Chen T, Wang Y, 2024. A state of art review on time series forecasting with machine learning for environmental parameters in agricultural greenhouses. Inf Process Agric 11:143-162.
López Santos M, Díaz García S, García-Santiago X, Ogando-Martínez A, Echevarría Camarero F, Blázquez Gil G, Carrasco Ortega P, 2023. Deep learning and transfer learning techniques applied to short-term load forecasting of data-poor buildings in local energy communities. Energy Build 292:113164.

Lv W, Shen C, Li X, 2018. Energy efficiency of an air conditioning system coupled with a pipe-embedded wall and mechanical ventilation. J Build Eng 15:229-235.

Muller AC, Guido S, 2017. Introduction to machine learning with Python: a guide for data scientists. Sebastopol, O'Reilly Media.

Oh KC, Kim SJ, Park SY, Lee CG, Cho LH, Jeon YK, Kim DH, 2022. Development and verification of smart greenhouse internal temperature prediction model using machine learning algorithm. J Bio-Env Con31:152-162.

Oh KC, Park SY, Kim SJ, Choi YS, Lee CG, Cho LH, Kim DH, 2019. Development and validation of mass reduction model to optimize torrefaction for agricultural byproduct biomass. Renew Energy 139:988-999.

Oh KC, Park SY, Kim SJ, Cho LH, Lee CG, Kim DH, 2024a. Development of a greenhouse environmental forecasting model using machine learning to optimize energy consumption. Available from: https://ssrn.com/abstract=4975488

Oh KC, Kwon H, Park SY, Kim SJ, Kim J, Kim D, 2024b. Hyperparameter optimization of the machine learning model for distillation processes. Int J Intell Syst 2024:5564380.

Outanoute M, Lachhab A, Ed-Dahhak A, Selmani A, Guerbaoui M, Bouchikhi B, 2015. A neural network dynamic model for temperature and relative humidity control under greenhouse. Proc. Third Int. Workshop on RFID and Adaptive Wireless Sensor Networks (RAWSN), Agadir; pp. 6-11.

Pardey PG, Beddow JM, Hurley TM, Beatty TKM, Eidman VR, 2014. A bounds analysis of world food futures: global agriculture through to 2050. Aus J Agri Res Econ 58:571-589.

Runge J, Saloux E, 2023. A comparison of prediction and forecasting artificial intelligence models to estimate the future energy demand in a district heating system. Energy 269:126661.

Runge J, Zmeureanu R, 2019. Forecasting energy use in buildings using artificial neural networks: a review. Energies 12:3254.

Serale G, Fiorentini M, Capozzoli A, Bernardini D, Bemporad A, 2018. Model predictive control (MPC) for enhancing building and HVAC system energy efficiency: problem formulation, applications and opportunities. Energies 11:631.

Smith SL, Kindermans PJ, Ying C, Le QV, 2017. Don't decay the learning rate, increase the batch size. arXiv:1711.00489.

Song J, Zhang L, Xue G, Ma Y, Gao S, Jiang Q, 2021. Predicting hourly heating load in a district heating system based on a hybrid CNN-LSTM model. Energy Build 243:110998.

Venkateswaran D, Cho Y, 2024. Efficient solar power generation forecasting for greenhouses: a hybrid deep learning approach. Alex Eng J 91:222-236.

Wang H, Asefa T, Duncan J, 2024. Event-based evaluation of operational ENSO forecasting models in 2002–2020: implications for seasonal water resources management. J Hydrol 636:131295.

Wang Y, Duan X, Zou R, Zhang F, Li Y, Hu Q, 2023. A novel data-driven deep learning approach for wind turbine power curve modeling. Energy 270:126908.

Wang T, Mehdi QH, Gough NE, Griffiths IJ, 1997. A hybrid intelligent controller based on "consultation of doctors". IFAC Proc 30:457-462.

Wei X, Liu Y, Gao S, Wang X, Yue H, 2019. An RNN-based delay-guaranteed monitoring framework in underwater wireless sensor networks. IEEE Access 7:25959-25971.

Wu Y, Gao Y, Wang C, Chen Q, Ming T, 2023. The energy saving performance of the thermal diode composite wall in different climate regions. Renew Energy 219:119360.

Yao Y, Shekhar DK, 2021. State of the art review on model predictive control (MPC) in heating ventilation and air-conditioning (HVAC) field. Build Environ 200:107952.

Zhang Y, Meng F, Wang R, Zhu W, Zeng XJ, 2018. A stochastic MPC based approach to integrated energy management in microgrids. Sustain Cities Soc 41:349-362.

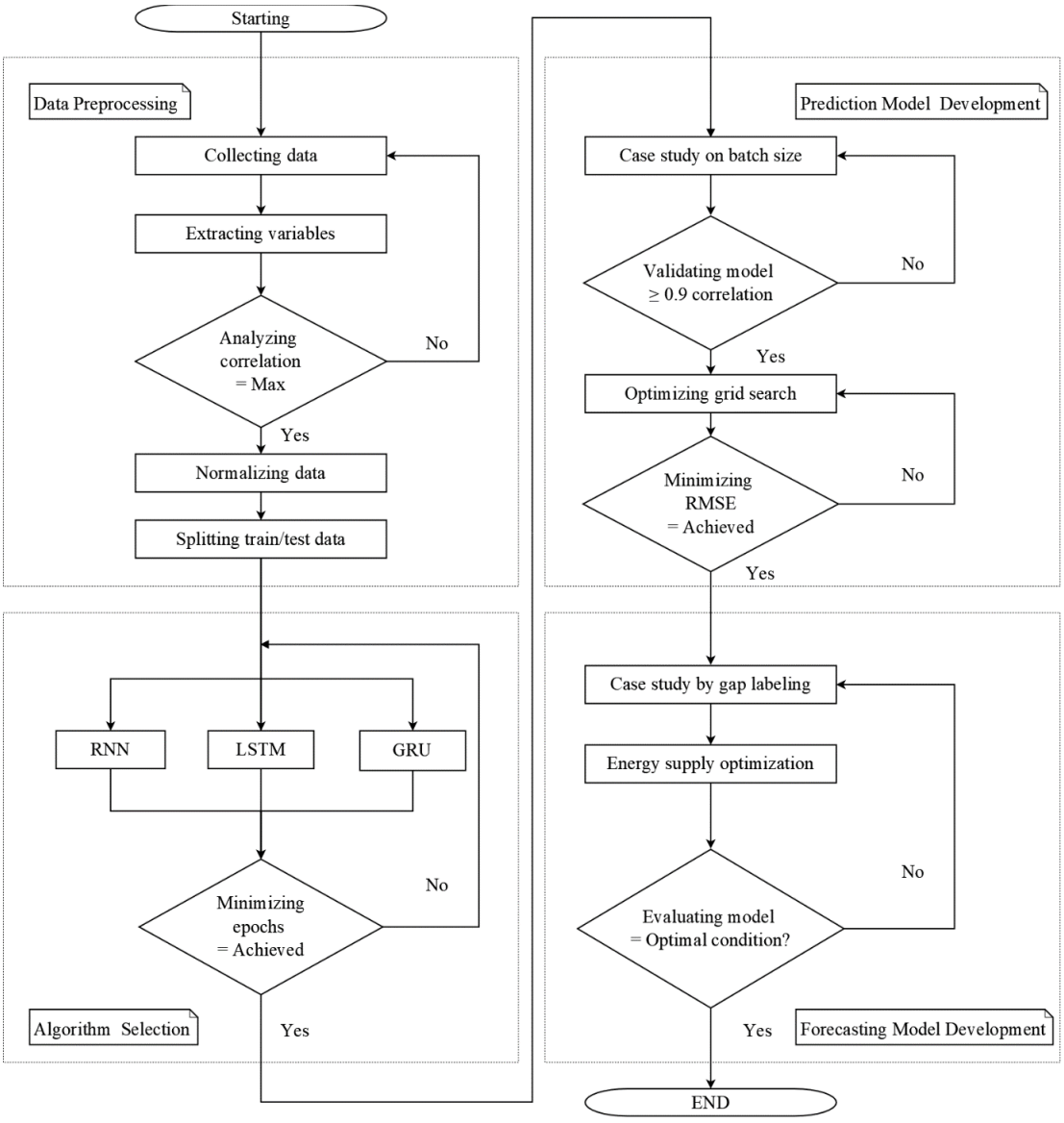


Figure 1. Architecture of machine-learning model development for a greenhouse system.

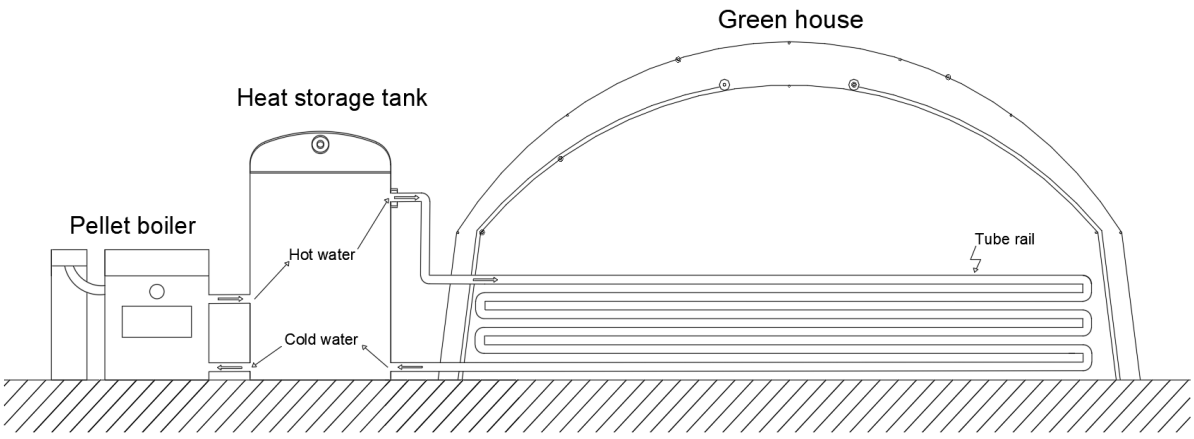
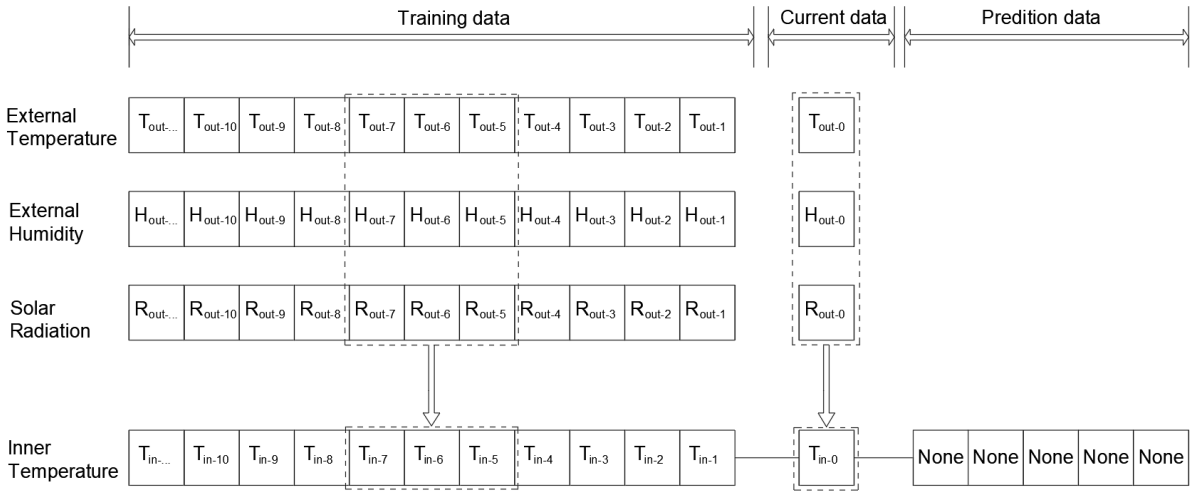


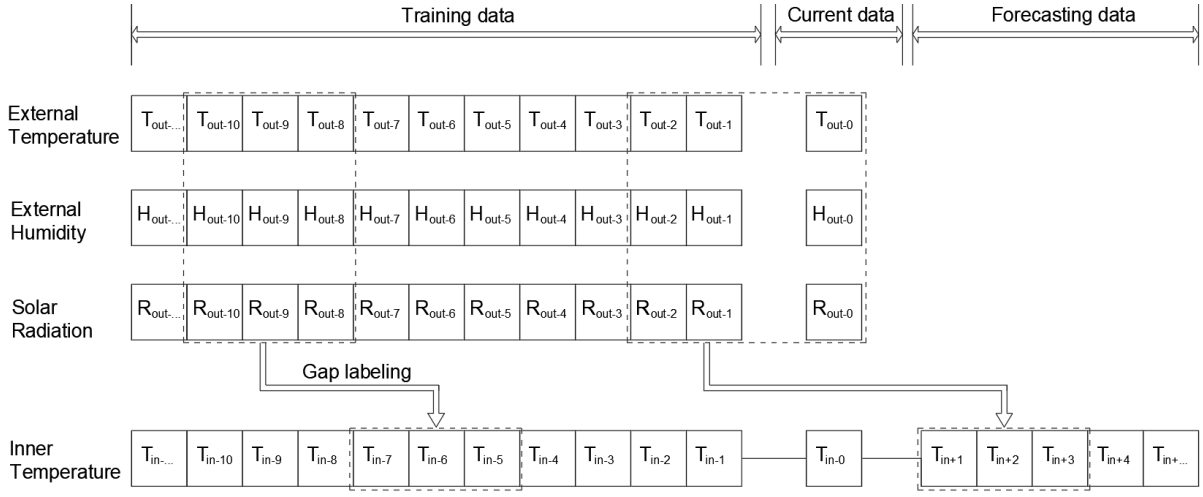
Figure 2. Schematic of the greenhouse energy supply system.

(a) (b) (c)

Figure 3. Pellet boiler **(a)**, heat-storage tank **(b)**, and interior view of a greenhouse **(c)**.



(a) Data-labeling relationship in prediction models



(b) Data-labeling relationship in forecasting models

Figure 4. Labeling methods for data used in prediction and forecasting models.

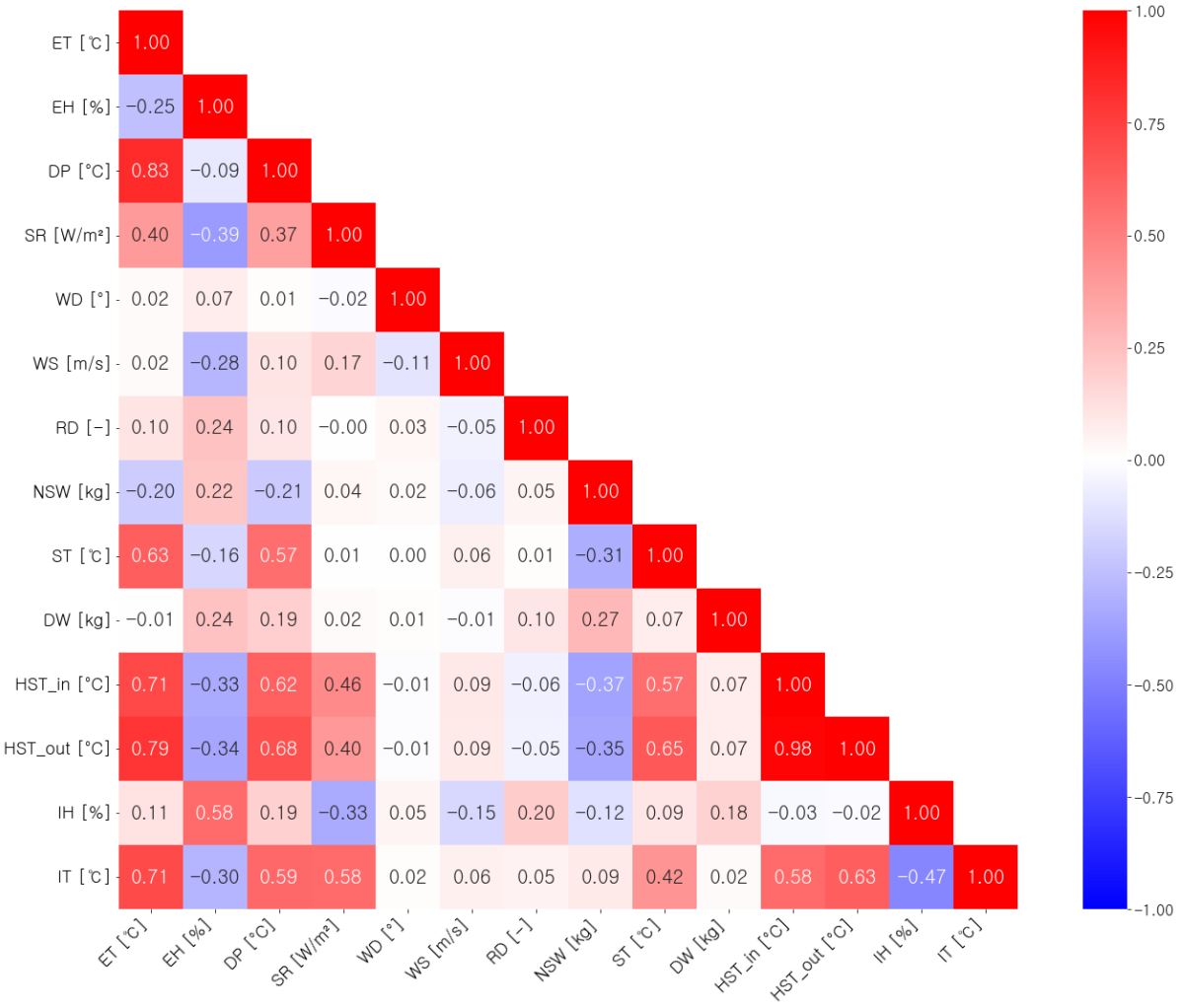


Figure 5. Correlation analysis of environmental variables in a greenhouse.

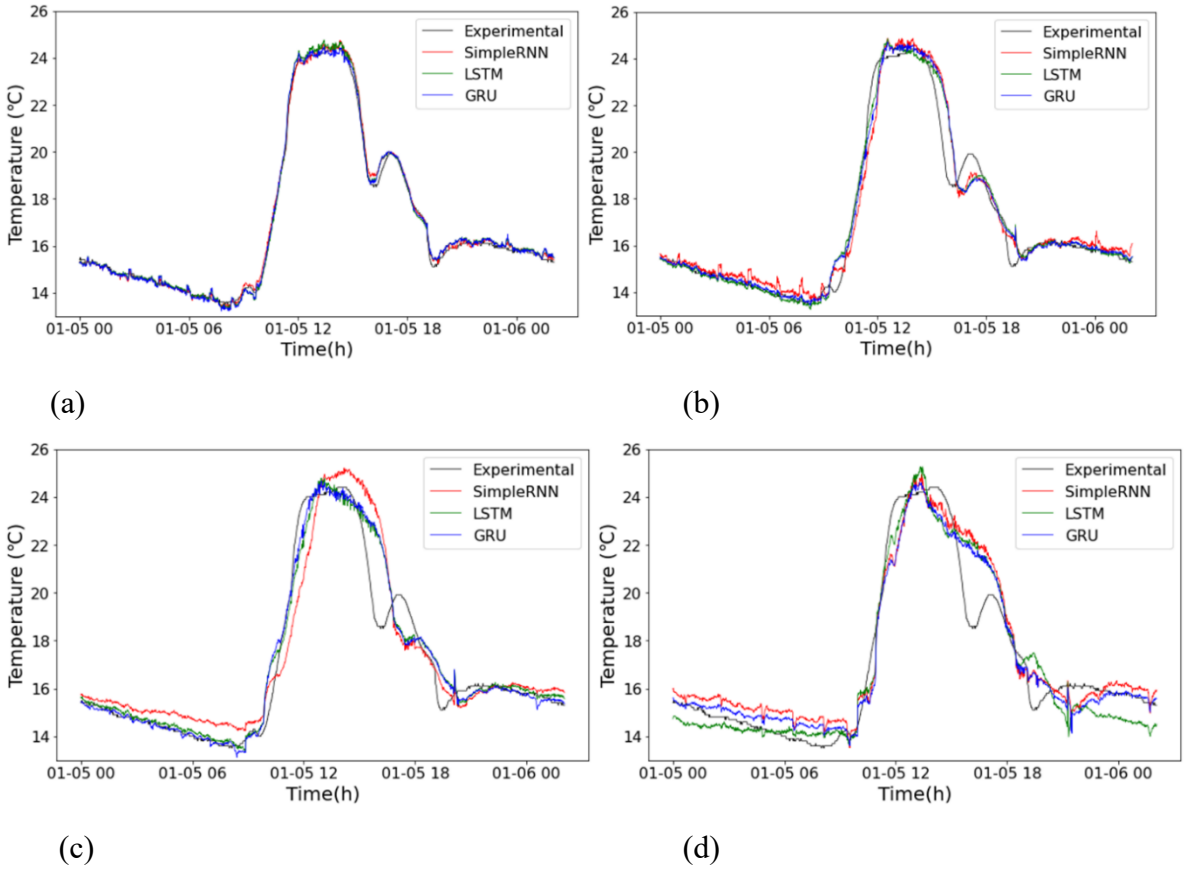
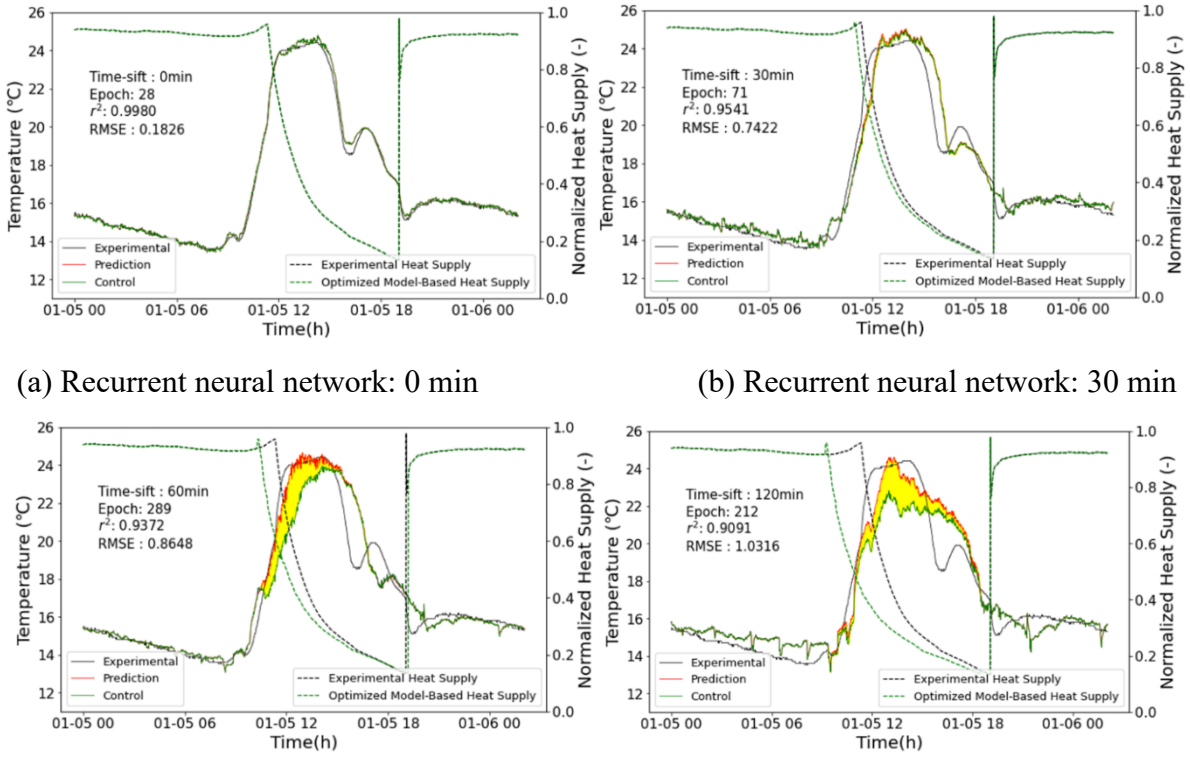


Figure 6. Comparative analysis of the internal temperature using different algorithms at different forecasting intervals (0, 30, 60, and 120 min).



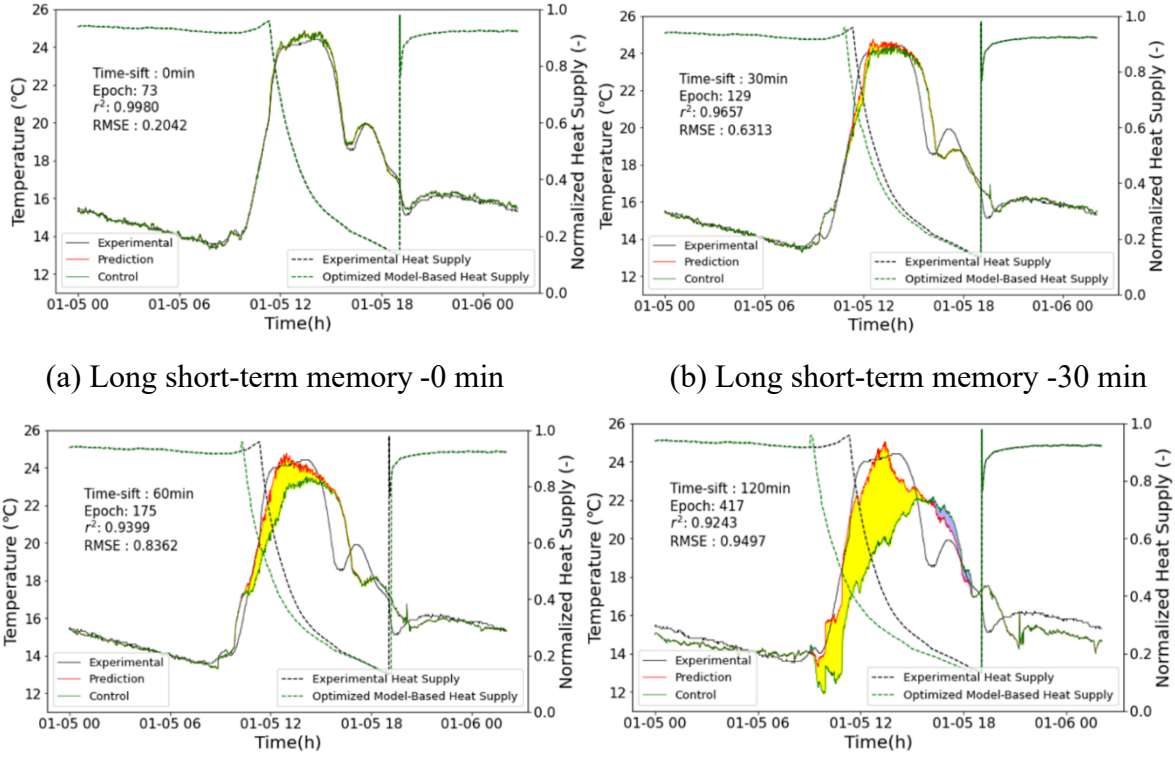
(a) Recurrent neural network: 0 min

(b) Recurrent neural network: 30 min

(c) Recurrent neural network: 60 min

(d) Recurrent neural network: 120 min

Figure 7. DDPC performance and heating-operation reduction using the recurrent neural network algorithm and varying time shifts.



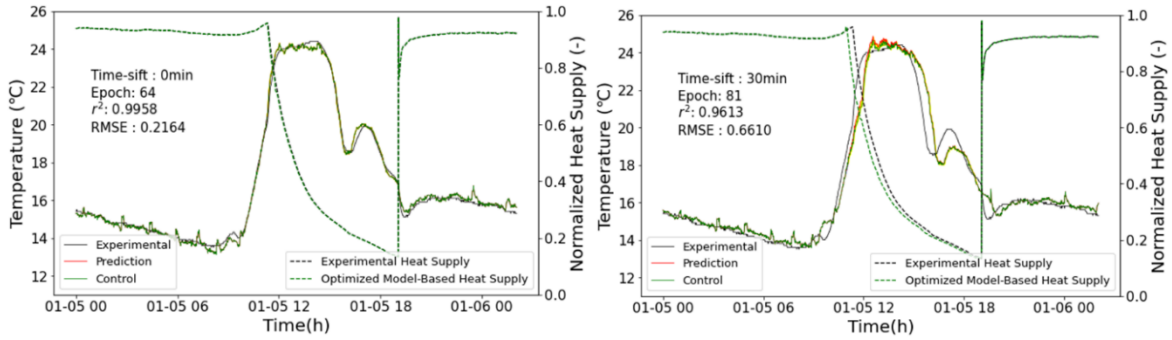
(a) Long short-term memory -0 min

(b) Long short-term memory -30 min

(c) Long short-term memory-60 min

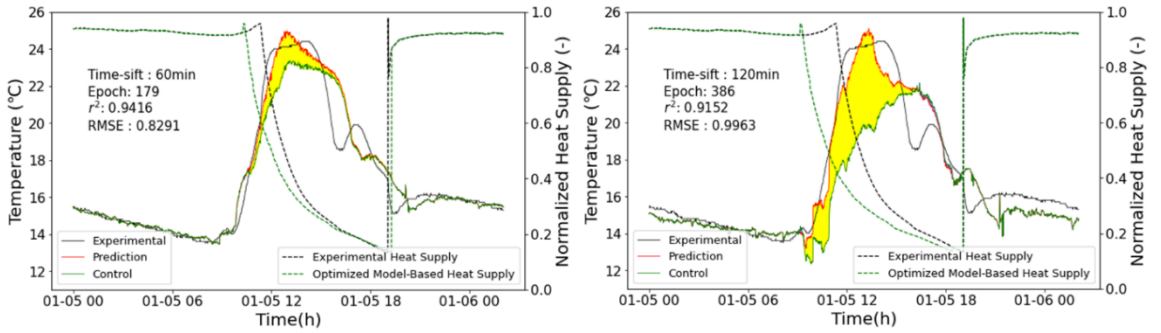
(d) Long short-term memory-120 min

Figure 8. DDPC performance and heating-operation reduction using the LSTM algorithm and varying time shifts.



(a) Gated recurrent units -0 min

(b) Gated recurrent units -30 min



(c) Gated recurrent units -60 min

(d) Gated recurrent units -120 min

Figure 9. DDPC performance and heating-operation reduction using the GRU algorithm and varying time shifts.

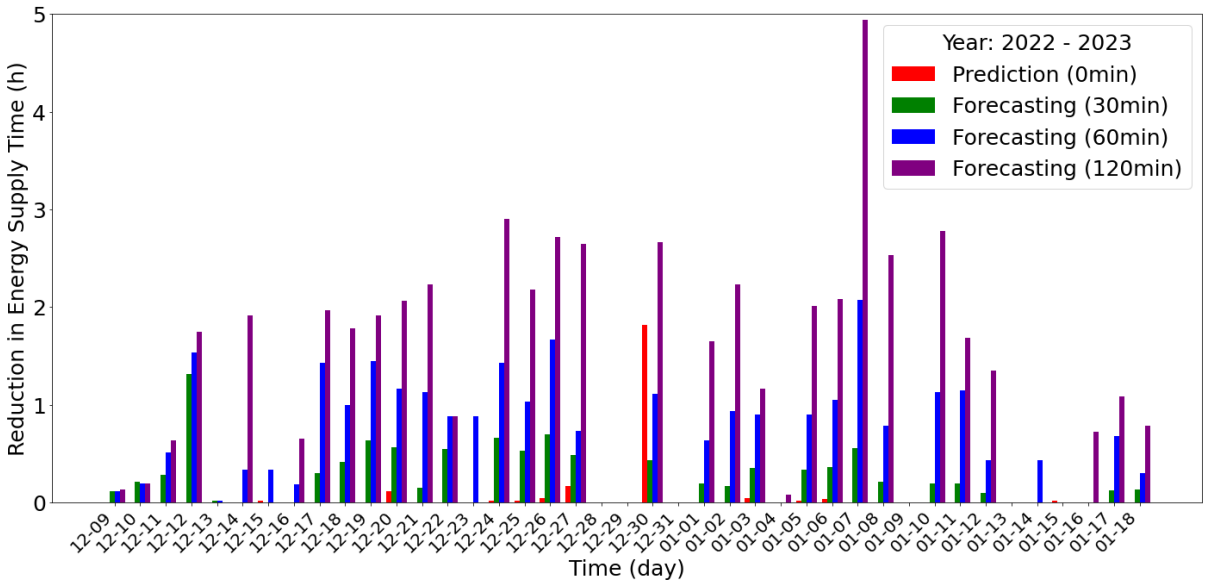


Figure 10. Daily reductions in energy supply operation times.

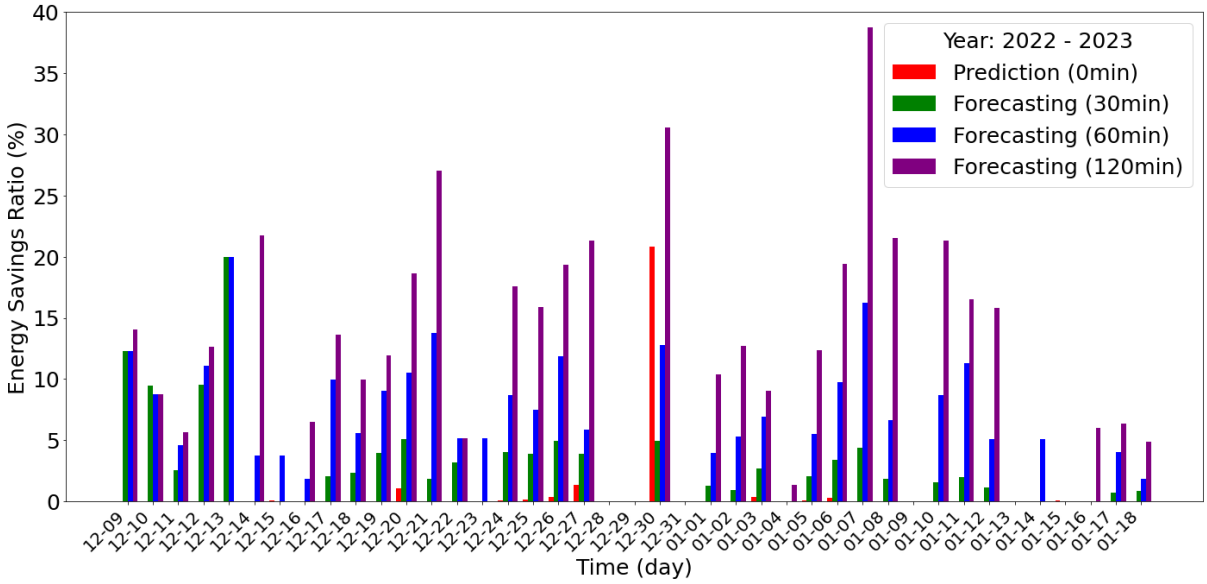


Figure 11. Calculated daily heating-operation reduction ratios.

Table 1. List of data collected.

Measurement data	Variables	Units	Values
External temperature*	ET	°C	Various
External humidity*	EH	%	[0, 100]
Dew point*	DP	°C	Various
Solar radiation*	SR	W/m^2	Various
Wind direction*	WD	-	0–360°
Wind speed*	WS	m/s	Various
Rain detection*	RD	-	[0, 1]
Nutrient solution weight**	NSW	kg/s	Various
Substrate temperature**	ST	°C	Various
Drainage weight**	DW	kg/s	Various
Heat storage tank input temperature**	HST_{in}	°C	Various
Heat storage tank output temperature**	HST_{out}	°C	Various
Internal temperature**	IT	°C	Various
Internal humidity**	IH	%	[0,100]

*External weather station data; **greenhouse internal environment data.

Table 2. Model development conditions (Kwon *et al.*, 2021; Oh *et al.*, 2024b).

Item	Value
Algorithms	RNN, LSTM, GRU
Hidden layer	100
Optimizer	Adam
learning rate	0.001
Loss function	MSE
Batch size	256
Epochs	Early stopping (delta 10^{-4})

GRU, gated recurrent units; LSTM, long short-term memory; RNN, recurrent neural network.

Table 3. Comparative analysis of the average prediction model performance by algorithm.

Time interval (min)	RNN			LSTM			GRU			Model Type
	Epoch	r^2	RMSE	Epoch	r^2	RMSE	Epoch	r^2	RMSE	
0	92.2195	0.9141	0.8664	137.0732	0.9127	0.8787	115.5854	0.9091	0.8860	Prediction
30	93.3658	0.8096	1.2440	168.5854	0.8281	1.1827	150.0732	0.8297	1.1773	Forecasting
60	86.8048	0.7107	1.6583	199.7073	0.7519	1.5252	186.6341	0.7455	1.5379	Forecasting
120	108.7317	0.6072	2.1572	249.9268	0.6602	2.0255	235.5610	0.6545	2.0366	Forecasting

GRU, gated recurrent units; LSTM, long short-term memory; RNN, recurrent neural network.

Table 4. Results of the Wilcoxon signed-rank test for model comparison

Comparison	r^2 (*p*-value)	RMSE (*p*-value)
RNN *vs* LSTM	0.214	0.746
RNN *vs* GRU	0.336	0.777
LSTM *vs* GRU	0.557	0.614

GRU, gated recurrent units; LSTM, long short-term memory; RNN, recurrent neural network.