Energy & Buildings 355 (2026) 117063

Contents lists available at ScienceDirect

Energy & Buildings

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

A machine learning based rapid thermal performance modeling method for modular buildings with BIPV: A novel decomposition strategy with real-time prediction capabilities

Yiqian Zheng [b,c], Biao Yang [b], Miaomiao Hou [d], Yi Zhang [a], Yuekuan Zhou [e,f], Xing Zheng [g], Pengyuan Shen [a,*]

[a] Institute of Future Human Habitats, Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guandong Province, China
[b] School of Architecture, Harbin Institute of Technology, Shenzhen, Guandong Province, China
[c] Fuzhou Research Institute of Sustainable Development in Cities Ltd., Fuzhou, Fujian Province, China
[d] China State Construction Hailong Technology Company Limited, Shenzhen, Guandong Province, China
[e] Division of Emerging Interdisciplinary Areas, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong Special Administrative Region
[f] Department of Mechanical and Aerospace Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong Special Administrative Region
[g] Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong Special Administrative Region

ARTICLE INFO

Keywords:
Machine learning
Modular buildings
Building simulation
BIPV
Energy use

ABSTRACT

The global push for carbon neutrality has intensified the need for rapid and accurate energy prediction methods for BIPV-integrated modular buildings. Traditional physics-based simulation approaches suffer from excessive computational burden. This study presents a novel machine learning-based rapid energy prediction methodology specifically designed for modular buildings with building-integrated photovoltaics. A comprehensive feature engineering framework captures the unique thermal and geometric characteristics of modular construction through six-surface property encoding, geometric parameters, and solar irradiance calculations. The methodology employs a modular building decomposition strategy that enables individual module analysis while maintaining system-level accuracy. An XGBoost-based prediction model achieves superior performance across four representative climate zones. The model achieves R^2 values exceeding 0.93 for heating loads, cooling loads, and total energy consumption. Experimental validation using a real-world BIPV-integrated modular building demonstrates prediction accuracy within industry-acceptable limits, with mean absolute errors below 1.5°C. The computational efficiency assessment shows prediction speeds over 2,000 × faster than traditional simulation approaches, enabling real-time design iteration. Successful integration with Grasshopper parametric design platforms facilitates immediate energy feedback during conceptual design phases. This advancement removes computational barriers to energy performance optimization and supports the broader adoption of sustainable modular construction practices by providing practical tools for energy-informed design decision-making.

1. Introduction

The escalating global climate crisis has intensified the urgent need for carbon neutrality. The building sector has emerged as a critical battleground for achieving these ambitious goals [1]. Buildings currently account for approximately 40% of global energy consumption and 27% of carbon emissions. This contribution continues to grow at an annual rate of 1% [2]. This substantial environmental footprint has catalyzed the development and adoption of innovative building technologies. These include active and passive measures that can simultaneously reduce energy consumption and integrate renewable energy generation under the challenge of global climate change [1,3,4].

Modular buildings present unprecedented opportunities for sustainable construction. They are characterized by standardized production, rapid assembly, and high structural consistency [5,6]. These prefabricated systems offer significant advantages including enhanced quality control, reduced construction waste, and accelerated project delivery [7,8]. When integrated with building-integrated photovoltaics (BIPV) [9], modular construction systems can achieve synergistic

* Corresponding author.
E-mail address: pengyuan_pub@163.com (P. Shen).

https://doi.org/10.1016/j.enbuild.2026.117063
Received 12 September 2025; Received in revised form 17 December 2025; Accepted 24 January 2026
Available online 26 January 2026
0378-7788/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Nomenclature

Abbreviations

ANN	Artificial Neural Network
BIPV	Building-Integrated Photovoltaics
BIPV/T	Building-Integrated Photovoltaic/Thermal
IoT	Internet of Things
LHS	Latin Hypercube Sampling
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
PV	Photovoltaic
R^2	Coefficient of Determination
RF	Random Forest
RMSE	Root Mean Square Error
TMY	Typical Meteorological Year
WWR	Window-Wall Ratio
XGBoost	Extreme Gradient Boosting

Symbols

AR	Aspect ratio (−)
AR_{length}	Length-to-width aspect ratio(−)
AR_{width}	Width-to-height aspect ratio(−)
$B_{condition}$	Boundary condition encoding vector (−)
d	Day of the year (−)
E_{total}	Total building energy consumption (kWh/m^2)
$E_{module,i}$	Energy consumption of individual module i (kWh/m^2)
f_k	k-th tree in ensemble (−)
$f_t(x_i)$	Prediction from the t-th tree (−)
H_{module}	Module height (m)
k	Number of trees in ensemble (−)
L_{module}	Module length (m)
$l(y_i, \hat{y}_i)$	Loss function (−)
n	Number of samples/modules (−)
$S_{incident}$	Direct normal irradiance (W/m^2)
$S_{surface}$	Incident radiation on tilted surface (W/m^2)
S_{type}	Surface type encoding vector (−)
T	Number of leaves in tree (−)
W_{module}	Module width (m)
w_j	Leaf weight (−)
x_{ij}	Normalized parameter value for sample i and parameter j (−)
x	Original value (−)
y_i	Actual value (−)
$\hat{y}_i$	Predicted value (−)
y	Mean of actual values (−)
z	Standardized value (z-score) (−)

Greek Letters

α	Solar altitude angle
β	Surface tilt angle
γ	Minimum loss reduction parameter (−)
δ	Solar declination angle
$\Delta E_{interaction}$	Inter-module thermal interaction energy (kWh/m^2)
η	Learning rate (−)
θ	Angle of incidence
λ	L2 regularization parameter (−)
μ	Feature mean (−)
$\pi_j(i)$	Random permutation of integers (−)
σ	Feature standard deviation (−)
ψ	Surface azimuth angle
Ω(f)	Regularization term for tree (−)

Subscripts and Superscripts

i	Sample index
j	Parameter/feature index
k	Tree index in ensemble
t	Iteration number
module	Related to individual module
total	Related to total building

Feature Encoding Categories

Surface Type (T)

T0	Wall
T1	Roof/Ceiling
T2	Floor
T3	Air Boundary

Boundary Condition (B)

B0	Outdoor
B1	Ground
B2	Adiabatic

Construction Setting (C)

C0	Standard Wall
C1	Photovoltaic-integrated
C2	Roof
C3	Floor

Window-Wall Ratio (WWR)

WWR	0.0 to 0.7 (step: 0.1)

Solar Irradiance (D)

D0-9	0–1000 W/m^2 (discretized)

Aspect Ratio (A)

A0	Horizontal
A1	Vertical

benefits, enabling both energy generation and consumption optimization at the factory production stage [10]. The standardized nature of modular components facilitates systematic BIPV integration, allowing for scalable deployment of renewable energy technologies across multiple building projects.

However, optimizing BIPV-integrated modular buildings presents significant computational challenges [11]. Traditional physics-based energy simulation methods, while accurate, require substantial computational resources that render them impractical for large-scale design optimization. Single building energy simulations using tools like EnergyPlus [12] typically require demanding computational cost. This makes iterative design evaluation extremely time-consuming when thousands of potential design configurations must be assessed, unless lightweight simulation methods are adopted [13]. This computational bottleneck is particularly problematic in modular construction, where rapid design iteration and optimization are essential for maintaining competitive project timelines and costs. The limitations of conventional simulation approaches have also created a research gap in modular building energy prediction. Moreover, existing machine learning approaches, while computationally efficient, fail to adequately address the unique characteristics of modular construction systems. Most data-driven models are trained on conventional building datasets and cannot leverage the structural uniformity inherent in modular buildings, where energy behavior can be potentially derived from individual module performance. Hence, current models inadequately capture BIPV-specific thermal effects and the complex interactions between photovoltaic systems and building envelope performance.

This research addresses these critical limitations by developing a machine learning-based rapid energy prediction method specifically designed for BIPV-integrated modular buildings. The primary research objectives include: developing a modularity-aware decomposition strategy that exploits the repetitive characteristics of modular construction; creating a comprehensive feature engineering framework that captures BIPV thermal effects and modular building geometry; and constructing a high-precision XGBoost prediction model capable of real-time performance assessment.

The key contributions of this work encompass three significant advances. First, we present a novel modular building decomposition strategy that enables accurate energy prediction through individual module analysis while maintaining system-level precision across diverse climate zones. Second, we develop an innovative feature engineering approach that systematically captures the thermal and geometric characteristics unique to BIPV-integrated modular buildings. Third, we demonstrate the practical integration of machine learning prediction capabilities within parametric design workflows, enabling real-time energy feedback for design optimization and manufacturing guidance.

2. Literature review

Modular buildings possess distinctive structural and thermal characteristics that fundamentally differentiate them from conventional construction methods, creating both unique opportunities and challenges for energy modeling applications [14,15]. The standardized production process and high degree of structural consistency enable systematic energy performance prediction through individual module analysis [7,8,16,17]. Each modular unit exhibits identical thermal properties and construction details, allowing energy models to leverage this uniformity for improved accuracy and computational efficiency through decomposition strategies, where total building energy consumption can be derived from individual module performance characteristics [18]. The high degree of standardization creates opportunities for factory-stage optimization not available in conventional construction, where energy efficiency measures including enhanced insulation systems and renewable energy integration can be systematically implemented during manufacturing [19]. Research by Lau et al. demonstrated the potential for integrating photovoltaic systems with modular construction to achieve significant energy savings and carbon emission reductions [18].

Modular buildings also present unique modeling challenges related to inter-module thermal interactions and boundary conditions. The assembly process creates complex thermal bridge effects at module interfaces, which can significantly impact overall building thermal performance [20]. Traditional energy modeling approaches often struggle to accurately represent these inter-modular thermal dynamics, as the modular assembly process introduces discontinuities in the building envelope that create preferential heat transfer pathways. Building-integrated photovoltaics introduce additional complexity to modular building energy modeling through fundamental alterations to building envelope thermal properties. BIPV systems create intricate interactions between photovoltaic generation efficiency and building thermal loads through multiple mechanisms, including modified surface thermal resistance, altered solar heat gain characteristics, and dynamic thermal mass effects [19,21]. These modifications require sophisticated modeling approaches that can capture the bidirectional thermal relationships between photovoltaic systems and building energy consumption patterns. Research by Li et al. demonstrated that air gap thickness between photovoltaic modules and building envelope surfaces critically affects both PV efficiency and building thermal loads, with proper ventilation design capable of improving electrical efficiency while simultaneously reducing cooling loads through enhanced heat dissipation [22].

The thermal performance of BIPV systems is strongly influenced by installation configuration and environmental conditions [23]. Studies by Wang et al. showed that wind speed and ambient temperature variations significantly influence solar cell operating temperatures, thereby affecting both generation efficiency and heat transfer to the building interior [24]. When photovoltaic modules replace conventional glazing or wall systems, the resulting changes in thermal transmittance and solar heat gain can significantly impact building energy consumption patterns, particularly in climate zones with substantial heating or cooling requirements [25–29]. Research by Chen et al. demonstrated that BIPV/T systems can effectively control solar heat gain while simultaneously generating electricity, though the coupling effects between thermal and electrical performance require careful consideration in modeling approaches [20]. The dynamic nature of these thermal effects requires time-varying modeling approaches that can account for hourly and seasonal variations in BIPV performance and thermal impacts, as traditional static thermal models are inadequate for representing these complex relationships.

Prevalent building energy prediction methodologies have evolved around two primary approaches: physics-based modeling and data-driven methods [30]. Physics-based models, founded on fundamental thermodynamic principles and heat transfer equations, have served as the cornerstone of building energy analysis for decades [31,32]. Established simulation software such as EnergyPlus, TRNSYS, and Dymola have demonstrated remarkable maturity in representing complex building systems across multiple domains including structural thermal performance, lighting, ventilation, and renewable energy integration [33–35]. Despite their theoretical rigor and widespread adoption, physics-based methods suffer from significant practical limitations that constrain their application in iterative design optimization [36]. Single building energy simulations typically require approximately 20 s using established tools [34], making iterative design evaluation extremely time-consuming when thousands of potential design configurations must be assessed. Research by Zhan et al. demonstrated that even with advanced calibration techniques, physics-based models remain resource-demanding and require substantial expertise [37].

In response to these limitations, data-driven approaches have emerged as a compelling alternative, leveraging historical operational data to identify patterns and relationships that govern building energy consumption. Comprehensive reviews by Amasyali and El-Gohary identified over 100 studies employing data-driven methods for building energy prediction [31]. These methods typically follow a systematic process encompassing data collection, preprocessing, model training, and evaluation using standard performance metrics [38,39]. Data-driven models demonstrate superior capability in reflecting actual building energy system operational states compared to theoretical physics-based approaches, particularly when accounting for unpredictable occupant behaviors and system inefficiencies.

The application of machine learning (ML) techniques in building energy prediction has experienced remarkable growth, driven by advances in algorithmic sophistication and computational power. Comprehensive reviews have identified various ML algorithms successfully applied to energy prediction tasks, including linear regression, support vector machines, random forests, neural networks, and gradient boosting methods [16,31,35,39]. Decision trees provide intuitive interpretability but are susceptible to overfitting when confronted with noisy datasets [40]. Machine learning algorithms like random forests address these limitations through ensemble learning, combining multiple decision trees to reduce overfitting risk while maintaining robust performance [40]. Neural networks excel in processing high-dimensional and complex datasets but require substantial training datasets and computational resources [35,41,42].

XGBoost (Extreme Gradient Boosting) has emerged as a particularly effective algorithm for building energy prediction applications, consistently demonstrating superior performance in data science competitions and practical implementations [43]. Lei et al. developed an evolutionary deep learning model incorporating XGBoost that achieved high accuracy in building energy consumption prediction [16]. XGBoost employs an additive model approach combined with regularization techniques to optimize objective functions while minimizing overfitting, achieving computational efficiency through parallel processing and memory optimization that enables handling of large-scale datasets while maintaining prediction accuracy. Recent studies have demonstrated the effectiveness of hybrid approaches that combine multiple ML techniques, with ensemble methods integrating various algorithms showing improved robustness and accuracy compared to individual models.

Despite significant advances in both machine learning applications and building energy modeling, several critical research gaps persist in the domain of modular building energy prediction with BIPV systems. Most existing ML models are trained on conventional building datasets that do not reflect the unique characteristics of modular construction, particularly failing to leverage the structural uniformity that could improve prediction accuracy [33]. Current BIPV modeling approaches typically focus on either electrical generation efficiency or building thermal impacts in isolation, failing to capture the complex bidirectional interactions between photovoltaic systems and building energy consumption [10,20,24]. The integration of ML-based prediction models into parametric design workflows remains underdeveloped, despite the clear potential for enabling real-time energy feedback during conceptual

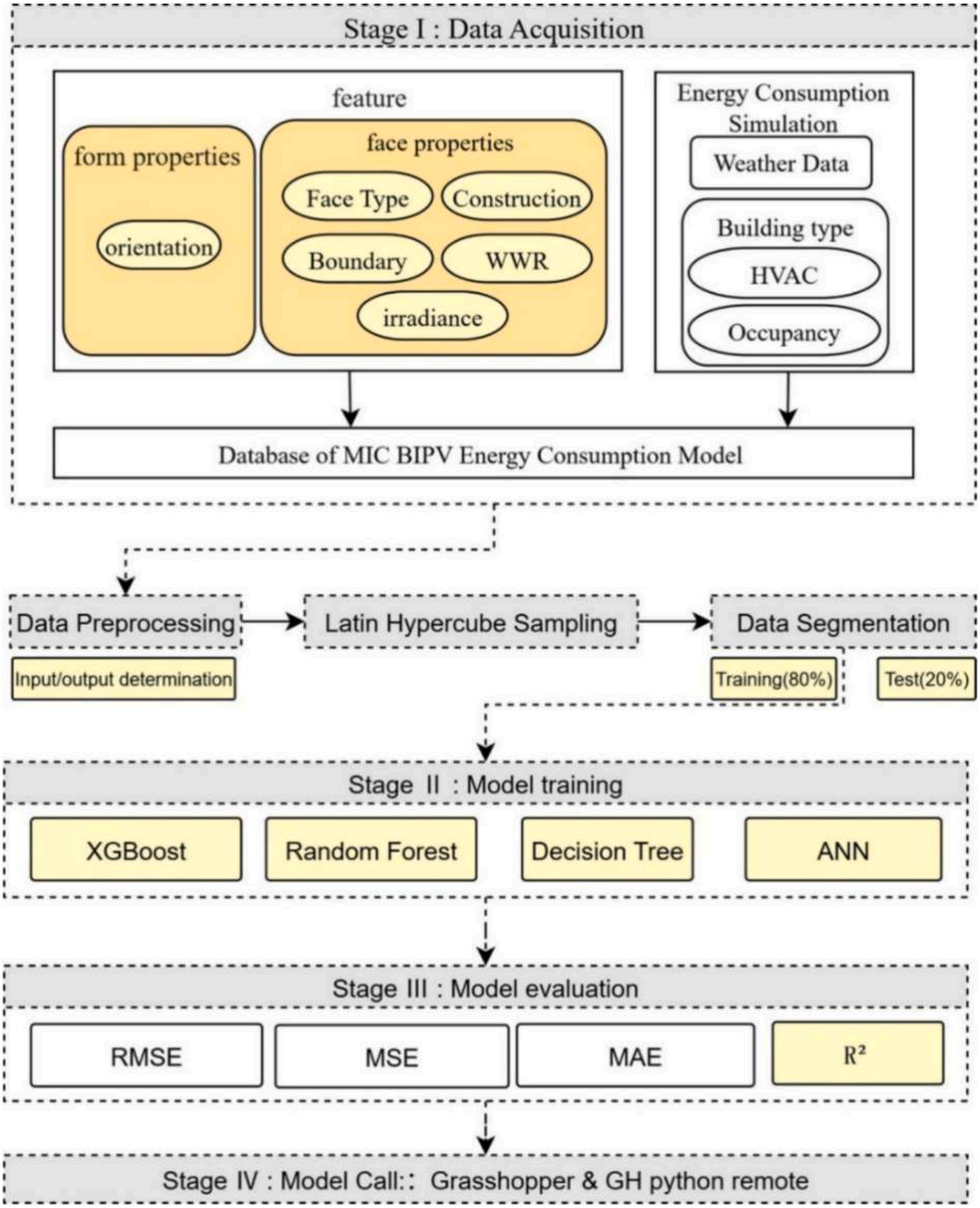


Fig. 1. Machine learning modeling process.

design phases. Furthermore, the validation of ML models across diverse climate zones and building types remains limited, with most studies focusing on specific geographic regions or building configurations, restricting the generalizability and applicability to broader modular construction applications.

The identified research gaps reveal a critical need for specialized energy prediction methodologies that explicitly address modular building characteristics while integrating BIPV-specific thermal effects. The lack of decomposition strategies that leverage structural uniformity, combined with insufficient feature engineering approaches for BIPV systems, necessitates a fundamentally different modeling paradigm. Moreover, the absence of validated models across multiple climate zones and the limited integration with parametric design workflows highlight the practical barriers to widespread adoption. To address these limitations, this research develops a novel machine learning-based approach that: (1) exploits the repetitive nature of modular construction through a decomposition strategy, (2) systematically captures BIPV thermal interactions through comprehensive feature engineering, (3) validates performance across diverse climate conditions, and (4) enables seamless integration with design platforms for real-time feedback. The following methodology section details how each of these objectives is achieved through systematic innovation in data-driven modeling for modular buildings.

3. Methodology

To address the identified research gaps, we develop a comprehensive machine learning-based prediction framework consisting of four key components: modular building decomposition strategy, feature engineering for BIPV integration, data-driven model development, and systematic dataset generation. Each component is designed to overcome specific limitations in existing approaches while maintaining computational efficiency and prediction accuracy.

3.1. *Modular building decomposition strategy*

3.1.1. *Individual vs. Integrated modeling approaches*

The fundamental challenge in modular building energy prediction lies in balancing computational efficiency with modeling accuracy while leveraging the inherent structural uniformity of modular construction systems. To address this challenge systematically, we develop a comprehensive four-stage framework as illustrated in Fig. 1. The workflow begins with modular building decomposition to convert whole-building analysis into individual module predictions, followed by feature engineering that captures thermal and geometric characteristics, machine learning model development using XGBoost, and experimental validation against real-world data. Fig. 1 presents this overview workflow, showing how these components integrate to form a complete prediction methodology. This research proposes a novel decomposition strategy that treats modular buildings as assemblies of thermally interacting individual units, enabling energy prediction through module-level analysis rather than whole-building simulation.

The decomposition approach is based on the premise that the total building energy consumption can be expressed as a function of individual module performance characteristics, accounting for inter-module thermal interactions through boundary condition modifications. The mathematical foundation of the decomposition strategy can be expressed as:

$$E_{total} = \sum_{i=1}^{n} E_{module,i} \pm \Delta E_{interaction}$$

where E_{total} represents the total building energy consumption, $E_{module,i}$ denotes the energy consumption of individual module i, n is the total number of modules, and $\Delta E_{interaction}$ accounts for inter-module thermal interaction effects.

The interaction term $\Delta E_{interaction}$ quantifies the difference between simple summation of individual module energy consumption and the actual whole-building energy performance. This term captures three primary phenomena: thermal bridge effects at module connection interfaces where structural elements create preferential heat transfer paths, modified convective heat transfer at inter-module boundaries due to restricted airflow patterns, and radiative exchange between adjacent module surfaces that differs from isolated module conditions. In conventional approaches, $\Delta E_{interaction}$ would need to be calculated as a separate correction factor. However, our methodology implicitly incorporates these interaction effects through systematic boundary condition modifications applied to individual modules during simulation. Specifically, surfaces between adjacent modules are assigned adiabatic boundary conditions (B2) when modules are at similar temperatures, or air boundary conditions (B3) when thermal stratification exists. This boundary condition encoding, combined with the machine learning model's ability to learn patterns from integrated whole-building simulations during training, effectively captures $\Delta E_{interaction}$ without requiring explicit calculation.

The selection of the decomposition strategy over alternative modeling approaches can be a valid alternative when people want to conduct fast simulation for MiC construction which is of certain patterns. Whole-building simulation, while comprehensive, suffers from computational complexity that scales exponentially with building size and configuration variations, making it impractical for parametric design exploration requiring thousands of design iterations. Alternative decomposition methods, such as floor-by-floor or zone-based approaches, fail to capture the specific thermal characteristics of modular construction where prefabricated units maintain distinct thermal boundaries even after assembly. The module-based decomposition strategy proposed in this study offers three distinct advantages that align with the physical characteristics of modular construction. First, it preserves the inherent modularity of prefabricated building systems, enabling analysis at the manufacturing and assembly level where design decisions are actually made. Second, it reduces computational requirements from $O(n^2)$ for whole-building parametric studies to $O(n)$ for module-level analysis, where n represents the number of design parameters. Third, it facilitates parallel computation of module performance characteristics, enabling simultaneous evaluation of multiple module types that can be combined into various building configurations. This approach transforms the building energy prediction problem from a computationally intensive whole-building simulation task into a manageable module-level prediction task, making real-time design iteration practically achievable.

Validation results in the Results and Discussions demonstrate that this approach maintains prediction accuracy within ± 10% across diverse module configurations, indicating that the boundary condition modifications successfully account for inter-module thermal interactions. The individual modeling approach treats each module as an independent thermal zone with modified boundary conditions that reflect its position within the larger building assembly. This approach enables parallel computation of module performance characteristics while maintaining the ability to aggregate results for whole-building analysis. The boundary condition modifications are implemented through a systematic classification of module surfaces based on their thermal interface characteristics: external surfaces exposed to ambient conditions, internal surfaces interfacing with adjacent modules, and ground-contact surfaces with specialized thermal boundary conditions.

As illustrated in Fig. 2, the feasibility analysis across different module configurations demonstrates the effectiveness of the decomposition strategy. The figure shows various module assembly patterns including inter-module horizontal and vertical combinations, three-module configurations, and four-module arrangements. Each configuration exhibits different thermal interface characteristics that must be accurately captured in the decomposition modeling approach.

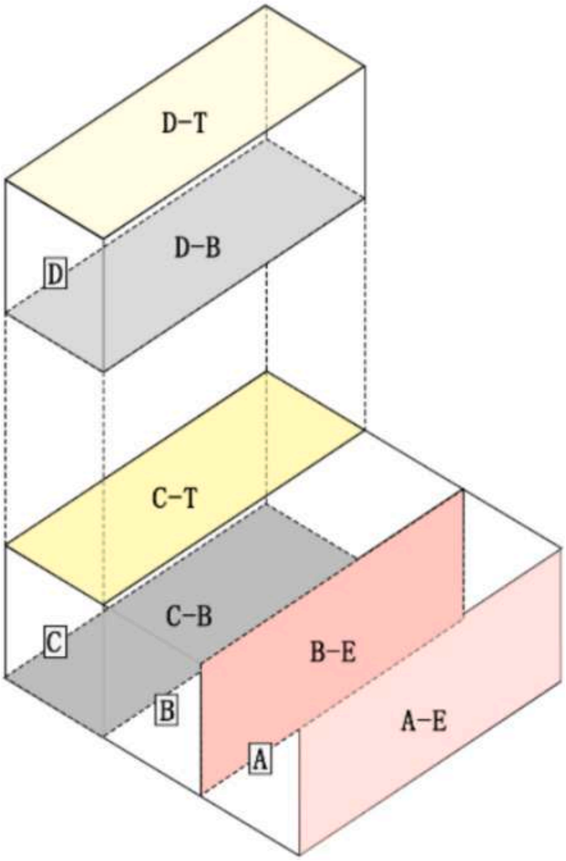


Fig. 2. Thermal interface of different modules in building energy modeling.

3.1.2. Module boundary condition analysis

The accurate representation of module boundary conditions is critical for the success of the decomposition strategy. Each module surface is classified according to its thermal interface characteristics, with boundary conditions systematically modified to reflect the specific thermal environment encountered in the assembled building configuration. Understanding how module position affects thermal boundaries is essential for accurate energy prediction. Fig. 2 demonstrates different module configurations, illustrating how modules A and B exhibit different thermal interface characteristics based on their position within the building assembly. As shown in the figure, module A-E represents a building external wall interface exposed to outdoor conditions, while B-E represents a thermally adiabatic wall or air wall connection between adjacent modules, highlighting the importance of position-dependent boundary condition specification.

The systematic encoding of surface types and boundary conditions requires a clear classification scheme that can be consistently applied across all module configurations. Fig. 3 presents the building surface encoding system that employs four primary categories: outdoor surfaces exposed to ambient environmental conditions, ground surfaces in direct contact with soil or foundation systems, adiabatic surfaces representing interfaces with adjacent modules at identical temperatures, and air boundary surfaces representing open connections between modules. This color-coded encoding scheme (shown in Fig. 3) enables surface classification for machine learning input.

For outdoor surfaces, the boundary condition implementation follows standard external surface thermal modeling approaches, incorporating convective and radiative heat transfer with ambient air and sky conditions. Ground surfaces employ specialized boundary conditions that account for soil thermal properties and seasonal ground temperature variations. Table 1 presents the comprehensive machine learning variables used in the decomposition strategy, systematically categorizing surface types (T0-T3), boundary conditions (B0-B2), construction settings (C0-C3), window-wall ratios (WWR), irradiance values (D0-9), and overall form parameters including aspect ratios and top floor indicators. This systematic encoding enables accurate representation of module thermal characteristics while maintaining compatibility with machine learning algorithms.

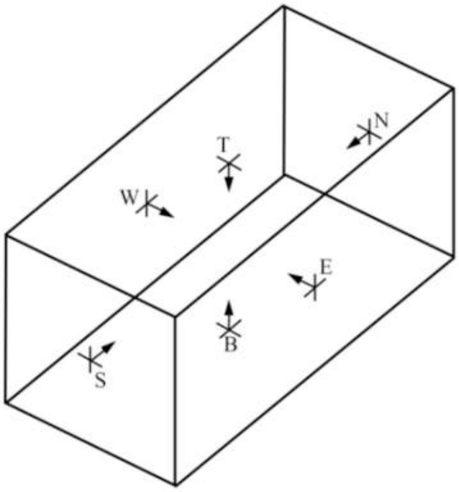


Fig. 3. Modularized construction surface coding.

Table 1
Machine learning model features.

Variable Category	Variable Name	Data Type	Range	Coding
Surface Properties	Surface Type	Categorical	Wall	T0
			Roof Ceiling	T1
			Floor	T2
			Air Boundary	T3
	Boundary Condition	Categorical	Outdoor	B0
			Ground	B1
			Adiabatic	B2
	Construction Setting	Categorical	Wall	C0
			PV	C1
			Roof	C2
			Floor	C3
	Window-Wall Ratio	Numerical	0–0.7 (Step: 0.1)	WWR
	Irradiance (Annual Average)	Numerical	0–1000 (Spatial Shading Equivalent Radiation Control)	D0-9
Overall Form Parameters	Aspect Ratio	Categorical	Horizontal	A0
			Vertical	A1

3.1.3. Feasibility validation across climate zones

The feasibility of the modular building decomposition strategy is validated through comprehensive analysis across four representative climate zones: severe cold (Harbin), cold (Beijing), hot summer and cold winter (Shanghai), and hot summer and warm winter (Shenzhen). The validation process employs systematic comparison between decomposed modeling results and integrated whole-building simulation results across different module configurations as illustrated in Table 2 presents

Table 2
Design variable value ranges.

Variable Type	Surface Properties						Aspect Ratio	Top Floor
Symbol	E	N	W	S	T	B	AR	L
Value Range	0–17					0–1	2	2
Step Size	1						1	1

the variable value ranges that define the scope of the validation study.

Based on this parameter space, the study generates 17 × 17 × 17 × 17 × 2 × 2 × 2 × 2 = 1,336,336 potential building samples. However, due to physical constraints, certain combinations are eliminated to ensure realistic building configurations. The validation methodology evaluates prediction accuracy through multiple performance metrics including mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). The mean absolute error is calculated as:

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

where y_i represents the actual energy consumption, $\hat{y}_i$ denotes the predicted energy consumption, and n is the number of validation cases.

The root mean square error is computed using:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Validation results demonstrate that the decomposed modeling approach achieves prediction accuracy within ± 10% compared to integrated modeling across all tested climate zones, with the largest deviations occurring in cold climate regions where inter-module thermal bridge effects are most pronounced. The coefficient of determination values exceeds 0.90 for all climate zones, indicating strong correlation between decomposed and integrated modeling results.

3.2. Feature engineering for modular buildings

3.2.1. Six-surface property encoding

The feature engineering framework for modular building energy prediction is built upon a comprehensive six-surface property encoding system that captures the thermal and geometric characteristics of each building module. As detailed in Table 3, this encoding system systematically represents the thermal properties and boundary conditions of all six surfaces of each modular unit: north, east, south, west facades, roof, and floor surfaces. Each surface is characterized through four primary attributes: surface type classification, boundary condition specification, construction assembly definition, and window-wall ratio quantification.

Surface type classification employs a categorical encoding system that distinguishes between wall surfaces (T0), roof/ceiling surfaces (T1), floor surfaces (T2), and air boundary surfaces (T3). The encoding utilizes one-hot representation to ensure compatibility with machine learning algorithms:

$$S_{type} = [s_{wall}, s_{roof}, s_{floor}, s_{air}]$$

where each element represents a binary indicator for the corresponding surface type.

Boundary condition specification captures the thermal interface characteristics of each surface through categorical variables representing outdoor exposure (B0), ground contact (B1), and adiabatic conditions (B2). The boundary condition encoding follows the mathematical representation:

$$B_{condition} = [b_{outdoor}, b_{ground}, b_{adiabatic}]$$

Table 3 presents the comprehensive variable combination encoding system, systematically showing how different surface orientations are encoded with their corresponding surface types, boundary conditions, construction settings, and window-wall ratios.

3.2.2. Geometric parameters and aspect ratios

Geometric parameters play a crucial role in modular building energy prediction due to their direct influence on surface area to volume ratios, thermal bridge effects, and solar exposure characteristics. The geometric feature set includes module aspect ratios, surface areas, volume characteristics, and spatial orientation parameters that capture the three-dimensional configuration of individual modules within the larger building assembly.

Module aspect ratios are calculated for both horizontal dimensions, representing the length-to-width ratio and width-to-height ratio of individual modules. These ratios influence natural ventilation patterns, structural thermal bridge locations, and solar heat gain distribution. The aspect ratio calculation follows:

$$AR_{length} = \frac{L_{module}}{W_{module}}$$

$$AR_{width} = \frac{W_{module}}{H_{module}}$$

where L_{module}, W_{module}, and H_{module}, represent the module length, width, and height in meter, respectively.

The comprehensive feature set requires systematic organization to

Table 3
Design variable combination and encoding.

Orientation	Surface Type	Boundary Condition	Construction Setting	Window-Wall Ratio	Encoding	Number
S (South)	0	0	0	0	0000	0
W (West)				1	0001	1
E (East)				2	0002	2
N (North)				3	0003	3
				4	0004	4
				5	0005	5
				6	0006	6
				7	0007	7
			1	0	0010	8
				1	0011	9
				2	0012	10
				3	0013	11
				4	0014	12
				5	0015	13
				6	0016	14
				7	0017	15
		2	0	0	0200	16
	3	2	0	0	3200	17
T(Top)	1	0	2	No Windows	102	0
		2			122	1
B(Bottom)	2	1	3		213	0
		2			223	1

manage the complex relationships between design parameters. Fig. 4 illustrates the building energy model input design parameter groups, showing how variables are organized into six main categories: surface properties (north, east, south, west, top, bottom), boundary conditions, construction settings, window-wall ratios, solar irradiance values, and overall form parameters. This hierarchical organization demonstrates how the multi-dimensional parameter space is structured to enable efficient sampling and model training while maintaining clear relationships between related variables.

3.2.3. *Solar irradiance evaluation method*

Solar irradiance evaluation constitutes a critical component of the feature engineering framework, as solar radiation directly influences both building thermal loads and photovoltaic generation potential. To quantify shading effects on surface irradiance without computationally expensive ray-tracing simulations, we develop a shading-equivalent-radiation methodology. Fig. 5 illustrates the experimental setup for evaluating how building spacing distances affect surface irradiance levels for different orientations, showing the geometric configuration used to derive empirical relationships. The research implements a tailor-made "spatial shading equivalent radiation" methodology, systematically quantified in Table 4, which demonstrates the relationship between building spacing distances and surface irradiance levels for different orientations as described in Fig. 5.

Table 4 reveals critical patterns in the distance-irradiance relationship: closer spacing (0–1 m) results in significantly reduced irradiance due to shading effects, while greater spacing (8–9 m) allows for maximum solar exposure. For instance, the East orientation shows irradiance values ranging from 0.00 kWh/m^2 at 0 m distance to 716.94 kWh/m^2 at 8–9 m spacing, while the South orientation demonstrates values from 0.00 kWh/m^2 to 773.53 kWh/m^2 across the same distance range. The empirical data from Table 4 are fitted with logarithmic functions to enable continuous irradiance prediction for any spacing distance. Fig. 6 presents the corresponding mathematical relationships for distance-irradiance fitting equations derived from the empirical data, showing both the scatter plots of simulated data points and the fitted logarithmic curves for all four orientations (North, East, South, West):

- North: $y = 354.36\ln(x) - 52.453$
- East: $y = 202.58\ln(x) - 32.883$
- South: $y = 234.54\ln(x) - 79.233$
- West: $y = 365.25\ln(x) - 88.358$

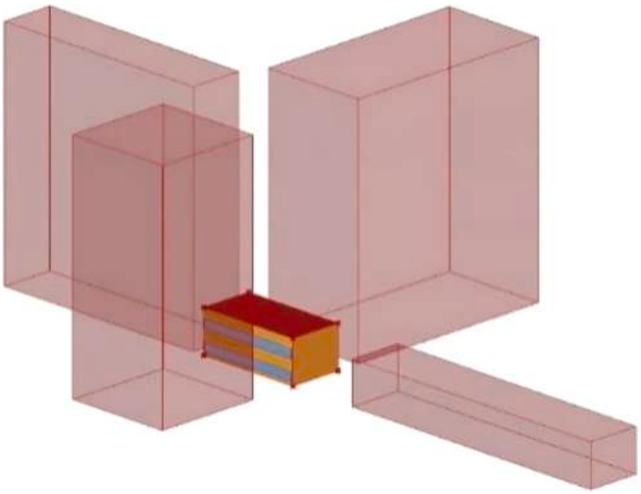


Fig. 5. Equivalent simulation design for the evaluation of relationship between building spacing distances and surface irradiance level.

The solar position calculation begins with the determination of solar declination angle, which varies throughout the year according to:

$$\delta = -23.45^{\circ} \times \sin\left[\frac{360}{365} \times (d + 10)\right]$$

where δ represents the solar declination angle in degrees and d denotes the day of the year (January 1 = 1).

For surfaces with arbitrary orientation and tilt, the incident solar radiation is calculated using:

$$S_{surface} = S_{incident} \times [\cos\alpha\sin\beta\cos(\psi - \theta) + \sin\alpha\cos\beta]$$

where $S_{surface}$ represents the incident radiation on the tilted surface, $S_{incident}$ denotes the direct normal irradiance, β is the surface tilt angle, and ψ represents the surface azimuth angle.

3.2.4. *Dataset generation, preprocessing, and normalization*

Data preprocessing and normalization procedures ensure optimal performance of machine learning algorithms while maintaining the physical significance of input features. In this research, categorical variables including surface types, boundary conditions, and construction assemblies undergo one-hot encoding to create binary indicator variables suitable for machine learning algorithms. As already shown in Table 3, surface types are encoded as T0-T3, boundary conditions as B0-B2, and construction settings as C0-C3, preventing artificial ordering while maintaining discrete categorical relationships.

Continuous variables including window-wall ratios (WWR), geo-

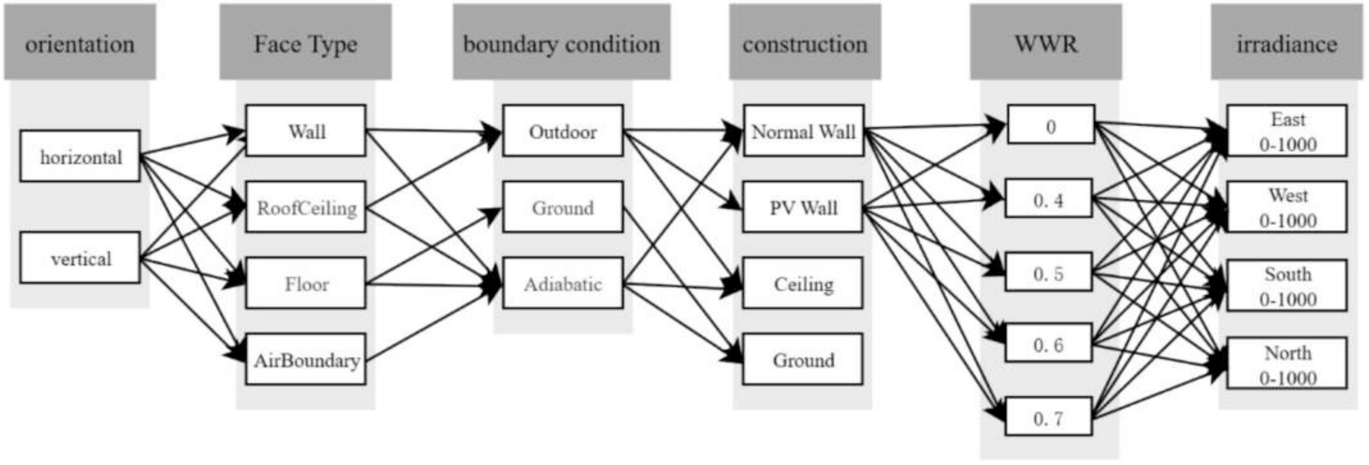


Fig. 4. Building energy model input design parameter set.

Table 4
Distance-irradiance relationship mapping.

Direction	Distance (m)	Irradiance (kWh/m^2)	Energy consumption when other surfaces unobstructed (kWh/m^2)	Orientation	Distance (m)	Irradiance (kWh/m^2)	Energy consumption when other surfaces unobstructed (kWh/m^2)
E	0	0.00	448.35	W	0	0.00	397.60
	1	107.18	453.36		1	31.33	417.22
	2	290.42	456.52		2	121.44	425.17
	3	451.29	458.57		3	196.32	431.41
	4	568.35	459.95		4	328.97	437.72
	5	599.66	460.77		5	334.53	443.69
	6	688.53	461.39		6	369.84	448.22
	7	688.53	461.93		7	396.42	452.95
	8	716.94	462.28		8	461.28	458.44
	9	716.94	462.55		9	510.16	462.55
N	0	0.00	432.10	S	0	0.00	414.40
	1	67.05	445.75		1	72.85	430.66
	2	157.50	450.55		2	225.99	438.75
	3	243.19	453.98		3	403.90	445.34
	4	327.40	456.24		4	570.54	449.93
	5	330.72	457.96		5	578.06	453.50
	6	385.25	459.53		6	643.77	456.52
	7	395.84	461.04		7	682.37	459.12
	8	395.84	462.14		8	682.37	461.25
	9	428.22	462.55		9	773.53	462.55

metric parameters, and solar irradiance values (D0-9) undergo standardization to ensure consistent scaling across different feature types. The standardization process employs z-score normalization:

$$z = \frac{x - \mu}{\sigma}$$

where z represents the standardized value, x is the original value, μ denotes the feature mean, and σ represents the feature standard deviation.

The generation of comprehensive training datasets employs Latin Hypercube Sampling (LHS) to ensure systematic coverage of the multi-dimensional design parameter space. The sampling methodology generates 5,000 unique building configurations for each climate zone, providing sufficient data diversity for robust model training. LHS provides superior space-filling properties compared to random sampling, ensuring representative sampling across all parameter combinations with fewer total samples. Detailed parameter ranges, sampling procedures, and constraint handling rules are provided in Appendix A.1. The systematic combination of design variables requires careful constraint handling to eliminate physically unrealistic configurations. Constraint frameworks address geometric compatibility, thermal boundary consistency, and construction assembly compatibility. The constraint validation process reduces the initial parameter space to physically realistic configurations, from which LHS selects 5,000 representative samples for each climate zone. Detailed constraint rules and validation procedures are provided in Appendix A.2.

3.3. Data-driven model development

3.3.1. Algorithm selection and justification

The selection of XGBoost (Extreme Gradient Boosting) as the primary machine learning algorithm for modular building energy prediction is based on its demonstrated superior performance in handling complex, high-dimensional datasets with mixed variable types. XGBoost employs an ensemble learning approach that combines multiple weak learners (decision trees) through gradient boosting, iteratively improving prediction accuracy while incorporating regularization mechanisms to prevent overfitting. The XGBoost algorithm optimizes an objective function that combines prediction error with regularization terms:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $l(y_i, \hat{y}_i)$ represents the loss function measuring the difference between actual and predicted values, $\Omega(f_k)$ denotes the regularization term for the k-th tree, and K is the total number of trees in the ensemble. The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

where γ controls the minimum loss reduction required for tree splitting, λ represents the L2 regularization parameter, T denotes the number of leaves in the tree, and w_j represents the leaf weights.

Hyperparameter optimization employs a systematic grid search approach combined with Bayesian optimization techniques to identify optimal parameter configurations for different climate zones. The optimization process considers learning rate, maximum tree depth, minimum child weight, subsample ratio, and regularization parameters. Detailed optimization procedures and final parameter values are provided in Appendix B.

3.3.2. Model training and validation procedures

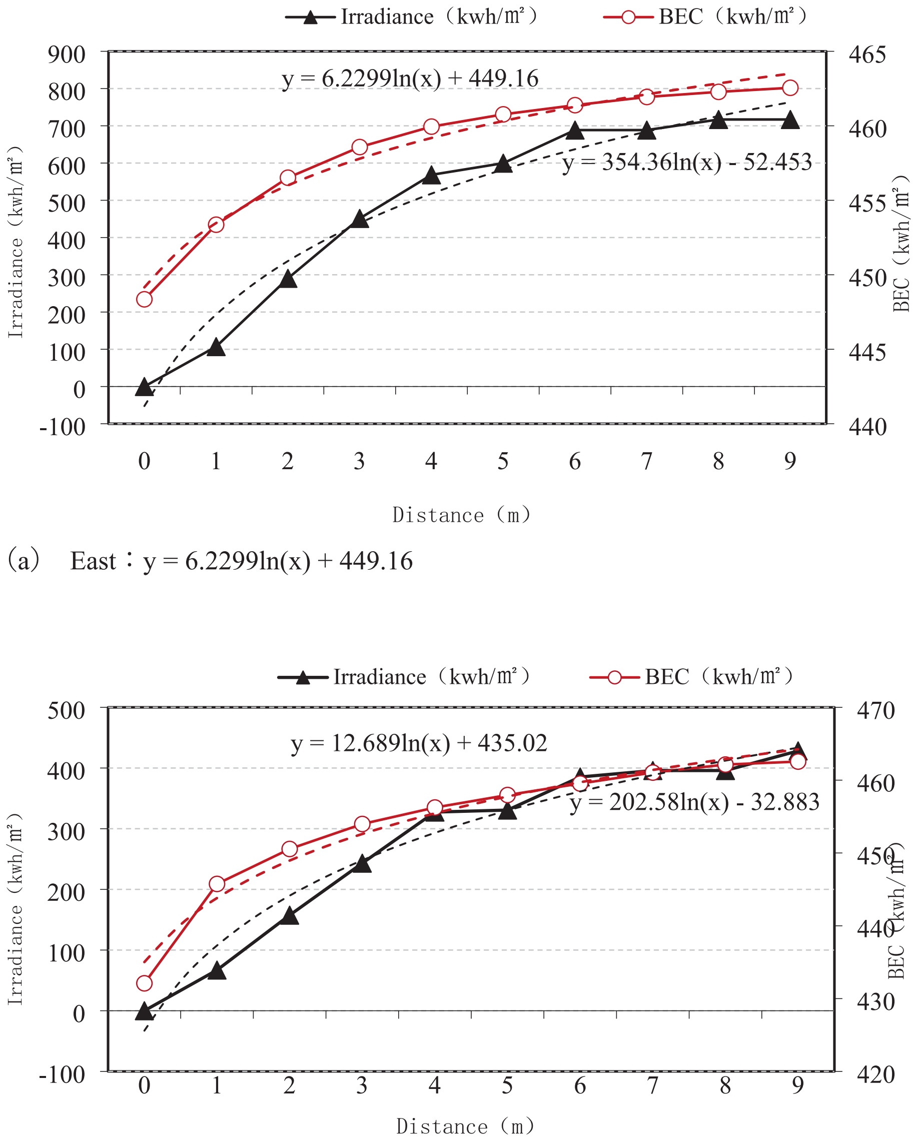
The model training procedure employs a systematic approach that balances prediction accuracy with generalization capability across diverse modular building configurations and climate conditions. The parametric modeling approach integrated with machine learning models was implemented in Grasshopper of Rhino, forming an automated workflow for building energy simulation that enables rapid dataset generation and model training.

The training dataset is partitioned using an 80:20 split, allocating 4,000 samples for training and 1,000 samples for validation to ensure robust performance evaluation. The validation metrics encompass multiple performance indicators including coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). The coefficient of determination is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $\bar{y}$ represents the mean of physically simulated values.

Uncertainty analysis procedures were implemented to characterize prediction reliability across operational scenarios and building configurations. Prediction intervals were calculated using bootstrapping methods with 1,000 resampling iterations to establish 90% and 95% confidence bounds for energy consumption predictions. The bootstrapping procedure randomly samples with replacement from the



（a） East：y = 6.2299ln(x) + 449.16

（b） North：y = 12.689ln(x) + 435.02

Fig. 6. Distance-irradiation experiments and equivalent fitting results.

validation dataset, trains the model on each bootstrap sample, and calculates prediction intervals based on the distribution of predictions across all iterations. This uncertainty quantification provides practitioners with realistic expectations of prediction reliability for different module configurations and operational conditions, enabling assessment of whether prediction accuracy is sufficient for specific design decision-making contexts. The analysis examines how prediction uncertainty varies across different building configurations, climate zones, and operational parameters, identifying scenarios where model predictions are most and least reliable.

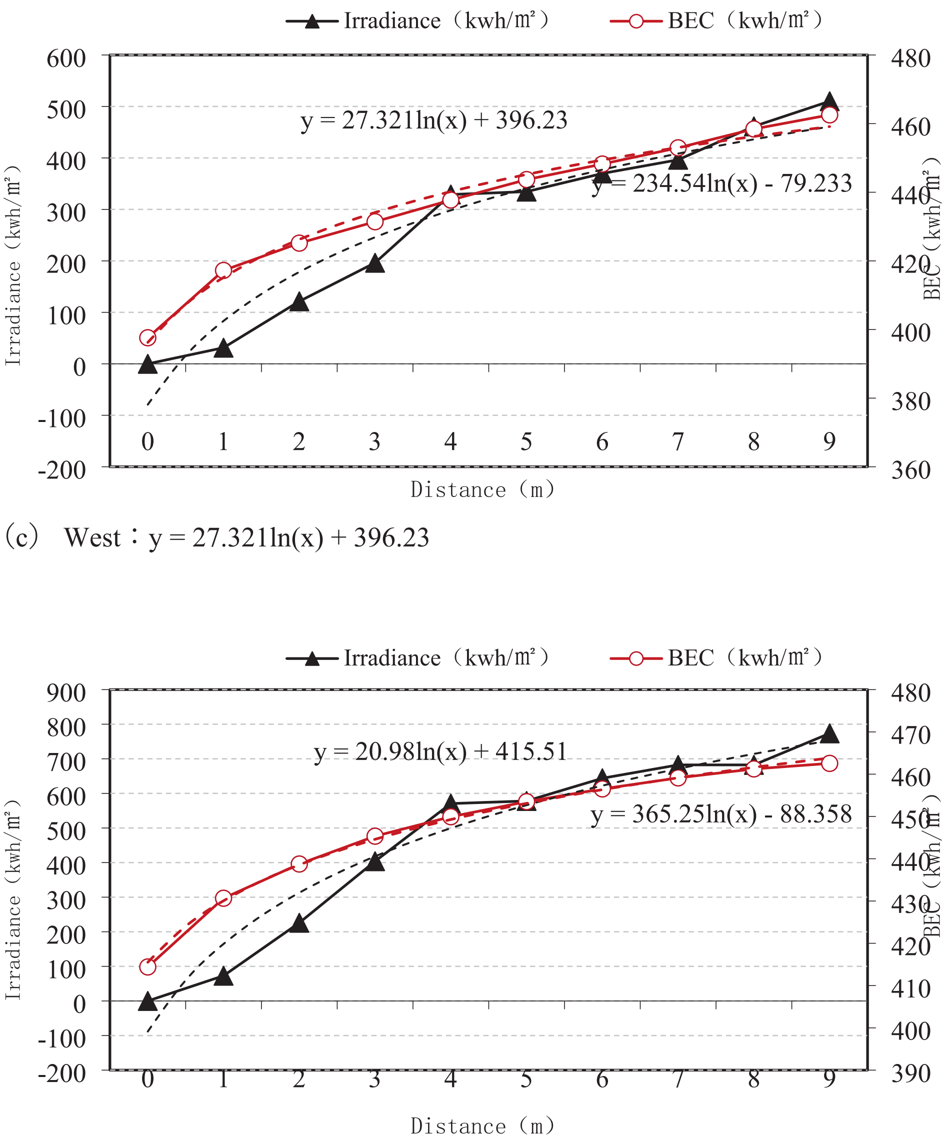
4. Simulation and validation

4.1. Case study building

4.1.1. C-smart building specifications

The experimental validation of the proposed machine learning-based energy prediction method employs the C-Smart building located in the Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone as the primary case study. This innovative modular building serves as an intelligent construction site command center and represents a cutting-edge example of BIPV-integrated modular construction technology. The C-Smart building is strategically positioned at coordinates 114.0716°E longitude and 22.5159°N latitude, placing it within a subtropical warm and humid climate zone classified as hot summer and warm winter region according to Chinese thermal design standards for civil buildings.

The building's modular design consists of two prefabricated units, each measuring 3 m in length and 3.25 m in width, assembled to create a functional command center facility. Unlike conventional photovoltaic installations that rely on mounting systems and brackets attached to existing building surfaces, the C-Smart building employs true building-integrated photovoltaics where the solar modules function as integral components of the building envelope. The photovoltaic panels are seamlessly integrated into both facade and roof surfaces, creating a smooth and continuous exterior appearance that demonstrates the aesthetic potential of BIPV technology in modular construction applications.



（c） West：y = 27.321ln(x) + 396.23

（d） South：y = 20.98ln(x) + 415.51

Fig. 6. (*continued*).

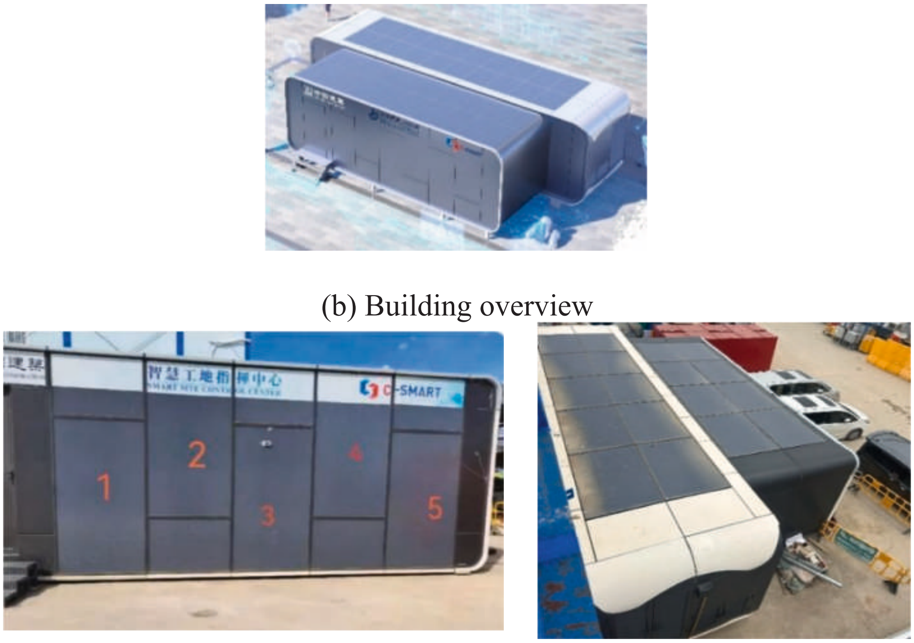
The modular construction approach enabled rapid deployment and assembly, with the entire building system prefabricated in controlled factory conditions before transportation and installation at the final site location. This manufacturing approach ensures consistent quality control and thermal performance characteristics while enabling systematic integration of BIPV components during the production phase. The building orientation is aligned with true north, providing optimal solar exposure conditions for both building thermal analysis and photovoltaic generation assessment.

The climatic conditions at the test site are characterized by high temperature and humidity levels typical of subtropical coastal regions. During the experimental monitoring period, the average maximum dry-bulb temperature reached approximately 37°C in July, requiring continuous cooling operation without heating requirements. The annual solar radiation totals approximately 4,759 MJ/m^2, with 2,120.5 sunshine hours representing 43.8% of possible sunshine duration. These favorable solar conditions provide an ideal testing environment for BIPV system performance evaluation and validation of the proposed energy prediction methodology.

4.1.2. BIPV system configuration

The BIPV system installation in the C-Smart building encompasses both vertical facade surfaces and horizontal roof areas, demonstrating the versatility of building-integrated photovoltaic technology across different building envelope orientations. The system employs monocrystalline silicon technology, selected for its high efficiency characteristics and proven reliability in subtropical climate conditions. Individual photovoltaic modules measure 1,225 mm × 1,825 mm, providing a standardized component size that facilitates modular integration and system scalability.

The complete BIPV installation consists of 23 photovoltaic panels strategically distributed across the building envelope to maximize solar energy capture while maintaining architectural integrity. The roof-mounted system incorporates 18 photovoltaic modules arranged in a systematic grid pattern as illustrated in Fig. 7, providing the majority of the system's electrical generation capacity. Fig. 7 shows the actual C-Smart building installation, with panel (b) providing an overview of the building, panel (c) displaying the rooftop photovoltaic arrangement with numbered modules, and panel (d) showing the vertical facade integration. The facade-integrated system includes 5 photovoltaic



(b) Building overview

(c) Rooftop photovoltaic diagram   (d) Facade photovoltaics

Fig. 7. Overview of building photovoltaics.

modules positioned on vertical building surfaces, demonstrating the application of BIPV technology on non-optimal orientations while contributing to overall system performance.

4.1.3. Geographic and climatic conditions

The geographic location of the C-Smart building provides representative conditions for subtropical building applications while offering excellent solar resource availability for BIPV system evaluation. Shenzhen's coastal location at 22.53°N latitude results in high solar angles throughout the year, with maximum solar elevation angles exceeding 85° during summer months. The geographic positioning creates favorable conditions for both horizontal roof-mounted and vertical facade-integrated photovoltaic systems, though with distinct performance characteristics that vary seasonally.

The subtropical maritime climate is characterized by distinct wet and dry seasons that significantly influence both building energy consumption patterns and photovoltaic system performance. Annual precipitation totals approximately 1933.3 mm, with the majority occurring during the summer monsoon period from May through September. This high-moisture environment in Guangdong Province presents challenges for building envelope systems and requires careful consideration of humidity effects on both thermal comfort and equipment performance [44]. Important weather variables under typical meteorological year (TMY) condition are visualized in Fig. 8. As can be seen from the visualization, seasonal temperature variations in Shenzhen are relatively moderate compared to continental climate zones, with daily average temperatures ranging from 14°C in winter months to 29°C during summer periods. The frost-free period extends for 355 days annually, eliminating concerns about freeze–thaw cycling effects on building materials and BIPV system components. However, the combination of high temperatures and elevated humidity levels creates substantial cooling loads throughout the extended summer season.

Solar radiation characteristics show strong seasonal variations that directly impact both photovoltaic generation potential and building thermal loads. Peak radiation periods coincide with maximum cooling requirements, creating favorable conditions for BIPV applications where electrical generation can directly offset air conditioning energy consumption. The annual solar radiation total of 4,759 MJ/m^2 compares favorably with other subtropical regions and provides sufficient energy resource for meaningful BIPV contributions to building energy balance. Wind patterns are influenced by the coastal location and monsoon circulation, with average wind speeds varying seasonally from 2.5 m/s during calm periods to over 8 m/s during storm events. These wind conditions affect both convective heat transfer from building surfaces and natural ventilation potential, influencing the thermal performance of BIPV-integrated envelope systems. The experimental monitoring system captures these environmental variables to enable comprehensive analysis of their effects on integrated building and photovoltaic system performance.

4.2. Monitoring and data collection

4.2.1. Instrumentation setup

The comprehensive monitoring system implemented for the C-Smart building experimental validation employs a multi-domain instrumentation approach designed to capture the complex interactions between BIPV systems, building thermal performance, and environmental conditions. The instrumentation design follows systematic principles to ensure data quality, measurement accuracy, and comprehensive coverage of all relevant physical phenomena affecting building energy consumption and photovoltaic generation.

The overall experimental system architecture integrates multiple measurement domains including outdoor environmental monitoring, indoor thermal environment assessment, building envelope thermal performance evaluation, and photovoltaic electrical performance tracking. The system design enables simultaneous data collection across all measurement domains with synchronized time stamping to facilitate correlation analysis and integrated performance assessment. Fig. 9 presents the detailed measurement point layout throughout the building in both plan and section views, showing the strategic positioning of temperature sensors (points 1–5), humidity sensors (points 6–7), and surface thermocouples across different wall orientations. This layout captures representative conditions while minimizing measurement interference from direct solar radiation or equipment heat sources. The

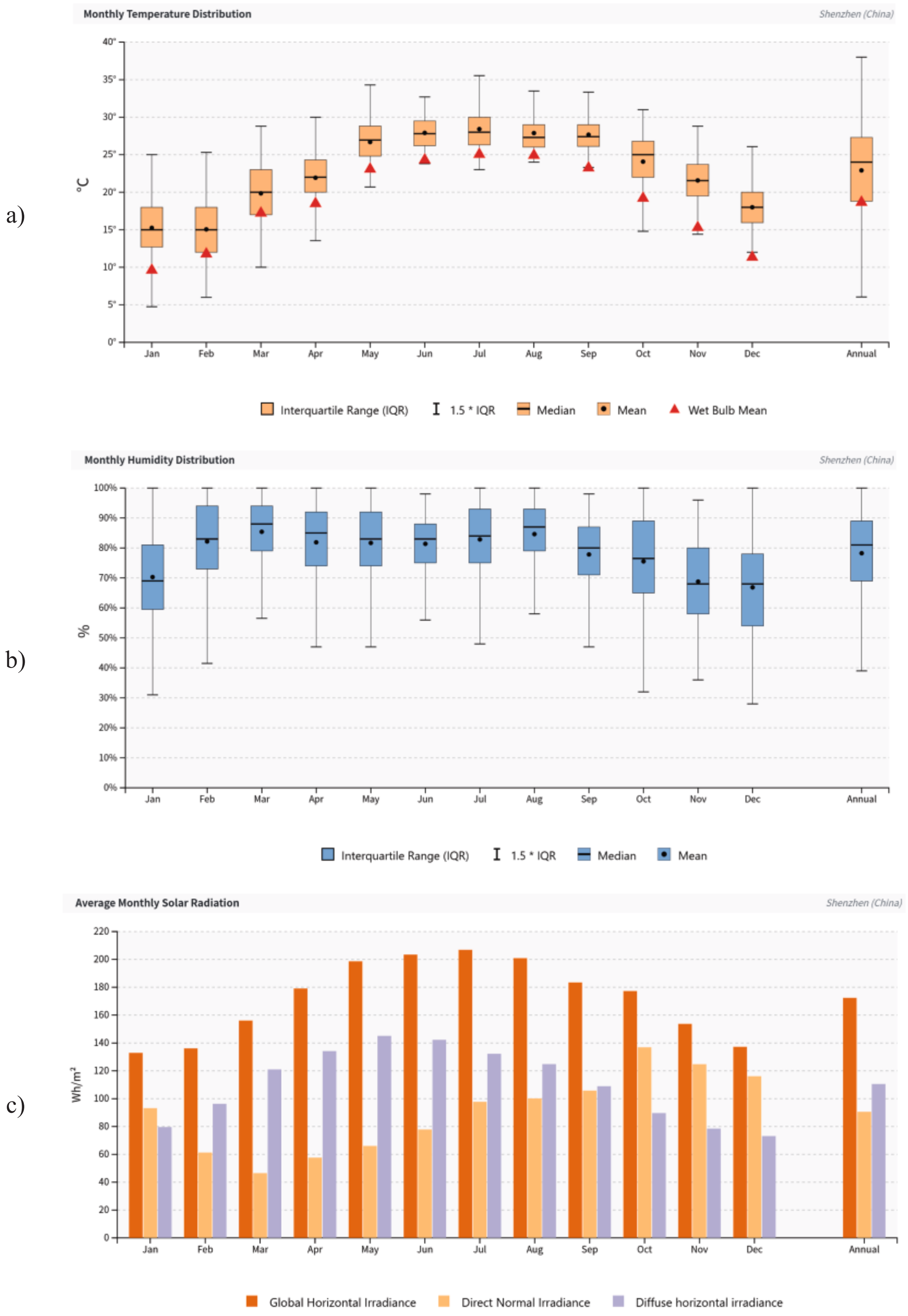


Fig. 8. Climatic conditions in the city of Shenzhen, China.

sensor placement strategy accounts for thermal stratification effects, spatial variations in environmental conditions, and the need for redundant measurements in critical locations to ensure data reliability and quality control.

Indoor environmental monitoring employs two air temperature and humidity sensors positioned at measurement points 6 and 7 as shown in Fig. 9. These sensors are suspended in the central interior space to capture representative indoor air conditions while avoiding direct solar radiation and proximity effects from building surfaces or equipment. The indoor measurement system provides continuous monitoring of thermal comfort conditions and enables assessment of building thermal response to varying outdoor conditions and BIPV system operation.

Outdoor environmental monitoring employs a comprehensive weather station positioned adjacent to the building to capture

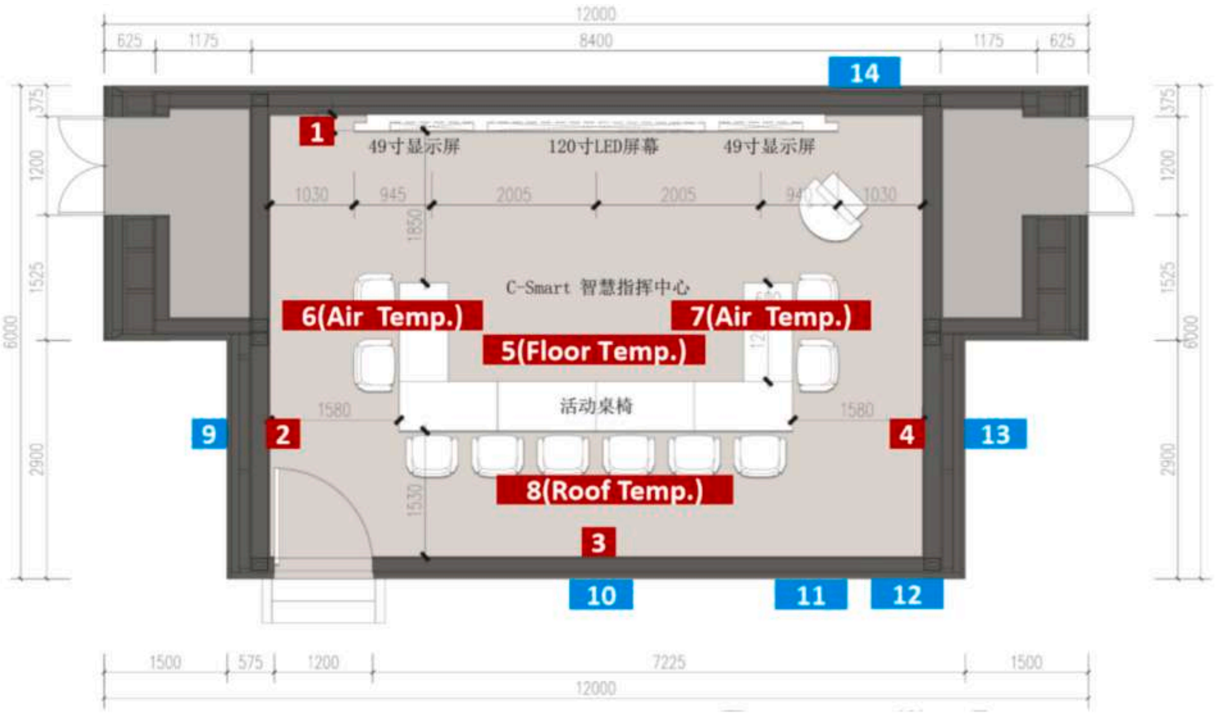
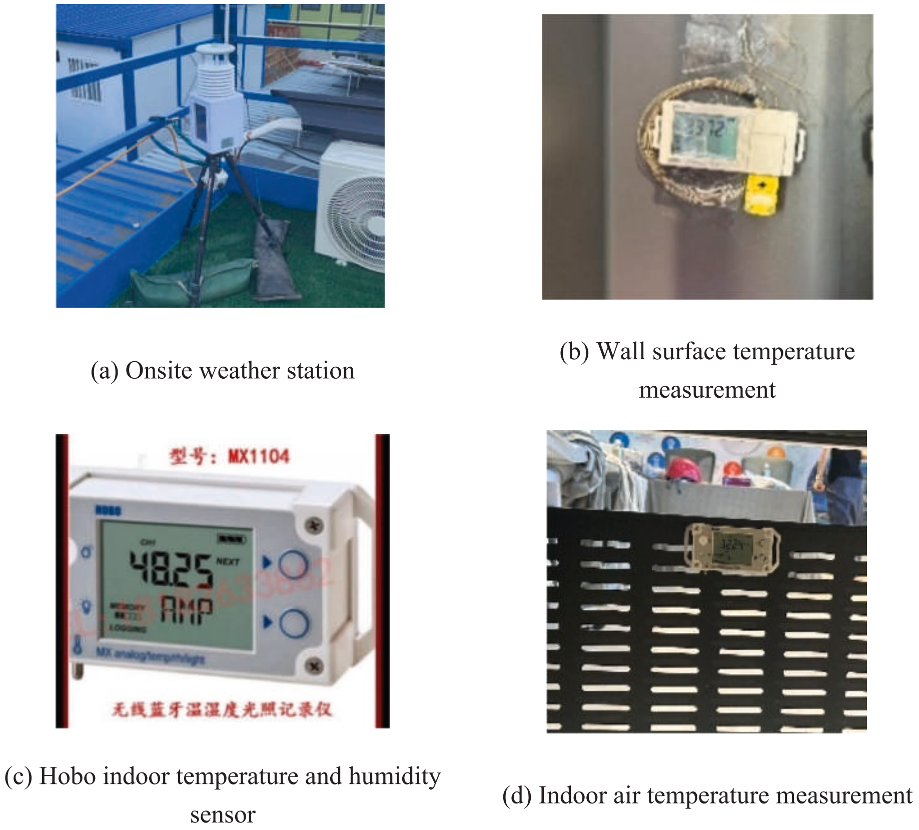


Fig. 9. Measurement Point Arrangement.

representative meteorological conditions. The weather station includes measurements of global solar radiation, ambient air temperature, relative humidity, wind speed and direction, and atmospheric pressure. Fig. 10(a) shows the weather station installation, which provides continuous monitoring of environmental conditions that directly influence both building thermal loads and photovoltaic system performance.

Building envelope thermal performance monitoring utilizes an array of 12 thermocouples strategically positioned on interior and exterior surfaces of the building walls as illustrated in Fig. 10(b). The



(a) Onsite weather station

(b) Wall surface temperature measurement

(c) Hobo indoor temperature and humidity sensor

(d) Indoor air temperature measurement

Fig. 10. Experimental Instruments and Setting up.

thermocouple placement captures temperature variations across different wall orientations and enables assessment of thermal bridge effects, solar heat gain impacts, and the thermal performance of BIPV-integrated envelope assemblies. The measurement points are distributed across north, east, south, and west-facing walls to capture orientation-dependent thermal effects. As shown in Fig. 10 (c)(d), HOBO was used indoors to measure the air temperature inside the building.

4.2.2. Data collection protocols

The data collection protocols ensure systematic and reliable measurement of all relevant variables affecting building energy consumption and BIPV system performance throughout the experimental validation period. The monitoring campaign extends from July 9, 2024, to August 11, 2024, providing continuous data collection over a complete month-long period that encompasses diverse weather conditions including clear days, cloudy periods, and rainy weather events. Automated data logging systems record measurements at 10-minute intervals throughout the complete 24-hour daily cycle, providing high temporal resolution for detailed analysis of dynamic thermal and electrical performance characteristics. The 10-minute measurement interval captures rapid variations in solar radiation, ambient temperature fluctuations, and photovoltaic system response while maintaining manageable data volumes for processing and analysis. The experimental methodology combines automated instrumentation with manual data recording procedures to enhance measurement reliability and provide verification of automated system performance. Manual readings are performed at regular intervals to cross-check automated measurements and identify potential instrumentation malfunctions or calibration drift. This dual-approach methodology ensures data quality while providing backup measurements for critical parameters.

Photovoltaic system performance data collection utilizes the integrated monitoring capabilities of the grid-connected inverter system, which provides continuous tracking of electrical generation, voltage output, current output, and photovoltaic module backsheet temperatures. The Internet of Things (IoT) platform enables remote data access with minute-level temporal resolution, facilitating real-time performance monitoring and historical data analysis. Fig. 10 demonstrates the various measurement instruments deployed throughout the experimental setup. Environmental data collection encompasses all meteorological variables that significantly influence building energy balance and photovoltaic performance. Solar radiation measurements include global horizontal irradiance, direct normal irradiance, and diffuse horizontal irradiance to enable comprehensive analysis of solar energy availability and its distribution between direct and scattered components. Temperature measurements capture both dry-bulb and wet-bulb temperatures to assess thermal comfort conditions and humidity effects.

Quality control procedures are implemented throughout the data collection period to ensure measurement accuracy and identify potential data anomalies. Automated data validation algorithms check for physically reasonable values, temporal consistency, and cross-parameter correlation to identify potential measurement errors. Manual data review procedures supplement automated validation to ensure comprehensive quality control and data reliability.

4.2.3. Quality control and calibration

Comprehensive quality control and calibration procedures are essential for ensuring the accuracy and reliability of experimental data used for model validation. The calibration methodology addresses all measurement systems including temperature sensors, humidity sensors, solar radiation instruments, and electrical measurement devices to establish traceability to recognized measurement standards and quantify measurement uncertainties.

Temperature measurement calibration employs an approach using certified reference thermometers and controlled temperature environments to establish sensor accuracy across the full range of expected operating conditions. All thermocouples undergo individual calibration to account for manufacturing tolerances and establish sensor-specific correction factors. The calibration process covers temperature ranges from 0°C to 50°C to encompass all anticipated environmental and building surface temperature conditions.

Solar radiation measurement calibration utilizes certified reference pyranometers to establish the accuracy of global radiation measurements. The calibration process accounts for angular response characteristics, temperature coefficients, and spectral response variations that may affect measurement accuracy under different solar conditions. Cross-calibration procedures compare multiple radiation sensors to identify potential measurement drift and ensure consistency across different measurement locations. Photovoltaic system monitoring undergoes comprehensive electrical calibration to ensure accurate measurement of power output, voltage, current, and energy production. The calibration process employs certified electrical measurement standards and covers the full range of operating conditions from low-light startup through peak generation periods. Temperature measurement calibration for photovoltaic module monitoring follows similar procedures to building thermal measurements but extends to higher temperature ranges up to 80°C to account for elevated module operating temperatures.

Data validation algorithms provide automated screening of collected measurements to identify potential anomalies, sensor malfunctions, or data transmission errors. The validation procedures include range checking to ensure measurements fall within physically reasonable bounds, temporal consistency analysis to identify sudden changes that may indicate sensor problems, and cross-parameter correlation analysis to verify measurement consistency across related variables. Statistical analysis procedures assess measurement uncertainty and establish confidence intervals for all collected data. The uncertainty analysis accounts for sensor accuracy specifications, calibration uncertainties, environmental effects on sensor performance, and data acquisition system uncertainties. This comprehensive uncertainty assessment enables proper interpretation of experimental results and establishes the statistical significance of observed phenomena.

4.3. Building simulation modeling

The accurate representation of building thermal properties constitutes a critical foundation for validating the proposed machine learning energy prediction methodology against real-world performance data. The simulation model calibration process requires detailed characterization of all thermal properties affecting building energy consumption, including envelope thermal resistance, thermal mass characteristics, air infiltration rates, and internal heat generation patterns.

Table 5 provides comprehensive specifications for the C-Smart building experimental case study, detailing the construction assemblies and thermal properties of all building envelope components. The building envelope design incorporates advanced materials and construction techniques typical of high-performance modular construction, including multi-layer insulation systems, thermal bridge mitigation strategies, and integrated weather barrier systems.

The wall assembly thermal performance is dominated by the 75 mm rock wool insulation layer, which provides the primary thermal resistance for the building envelope. The rock wool material specification of 80 kg/m^3 density ensures both thermal performance and structural integrity while maintaining compatibility with the light steel framing system. The multi-layer design creates a continuous insulation barrier that minimizes thermal bridging through the structural elements.

The roof assembly design incorporates similar insulation principles but adapts to the structural requirements of the horizontal surface orientation. The 50 mm thick rock wool insulation layer provides thermal resistance appropriate for the local climate conditions while supporting the waterproof membrane and aluminum panel finish. The structural steel plate provides the necessary structural strength for roof loads while contributing to the thermal mass characteristics of the

Table 5
Experimental case study building details.

Parameter	Description
Geographic Location	Xiawan Village, Hong Kong, China (Longitude: 114.07°E, Latitude: 22.52°N)
Meteorological Data	Weather station temperature, irradiance, and wind speed data
Experimental Period	July 9, 2024 – August 11, 2024
Time Interval	1 h
Model Construction Settings	
Wall Assembly	1. Aluminum panel joints filled with polyethylene foam strips, sealed with exterior sealant 2. 3 mm aluminum panel (metal frame system) 3. 40 × 40 × 4 steel framing, horizontal spacing matches panel width, vertical spacing matches panel length, connected to angle steel with expansion bolts fixed to wall 4. 9 mm high-density cement fiber board (staggered arrangement) 5. Waterproof breathable membrane 6. 9 mm high-density cement fiber board 7. 75 mm light steel framing (filled with 75 mm thick rock wool, density 80 kg/m^3) 8. 12.5 mm high-performance board 9. Interior finish
Floor Assembly	1. Floor adhesive 2. Specialized adhesive 3. 20 mm high-density cement fiber board + 20 mm high-density cement fiber board 4. 0.48 mm galvanized steel sheet
Door	Same as wall assembly
Roof Assembly	1. 3 mm aluminum panel (metal frame system) 2. PVC waterproof membra 3. 2 mm structural steel plate 4. 50 mm thick rock wool (density 80 kg/m^3) 5. 12.5 mm high-performance board 6. Interior finish

building.

Thermal bridge analysis requires careful consideration of the light steel framing system and its impact on overall envelope performance. The 40 × 40 × 4 mm steel framing members create preferential heat transfer paths through the insulation layer, reducing the effective thermal resistance of the wall assembly. The simulation model accounts for these thermal bridge effects through detailed two-dimensional heat transfer analysis of representative wall sections.

Air infiltration characteristics are determined through blower door testing and tracer gas measurements to quantify the actual air exchange rates under different pressure conditions. The modular construction methodology typically achieves superior airtightness compared to conventional site-built construction due to controlled factory assembly conditions and systematic sealing procedures. However, inter-module connections represent potential air leakage paths that require careful characterization and modeling.

5. Results and Discussions

5.1. Validation of building thermal performance with BIPV

The integration of building-integrated photovoltaics introduces thermal interactions that significantly affect both photovoltaic system performance and building thermal loads. The negative temperature differences observed during low irradiance periods result from infrared radiation exchange between photovoltaic modules and the clear sky environment. Under clear sky conditions, photovoltaic arrays emit infrared radiation to the cold upper atmosphere, resulting in module temperatures below ambient air temperature. This phenomenon is particularly pronounced during nighttime periods and early morning hours when solar irradiance is minimal but radiative cooling continues. Based on experimental data analysis combining measured cell temperatures and backsheet temperatures for the C-smart house, the thermal transmittance (U-value) of its photovoltaic assembly is calculated and calibrated to be approximately 17.9 W/(m^2·K). This thermal property value reflects the combined effects of photovoltaic module thermal resistance, air gap characteristics, and heat transfer mechanisms to the building interior. The U-value determination enables accurate modeling of heat transfer through BIPV-integrated envelope assemblies.

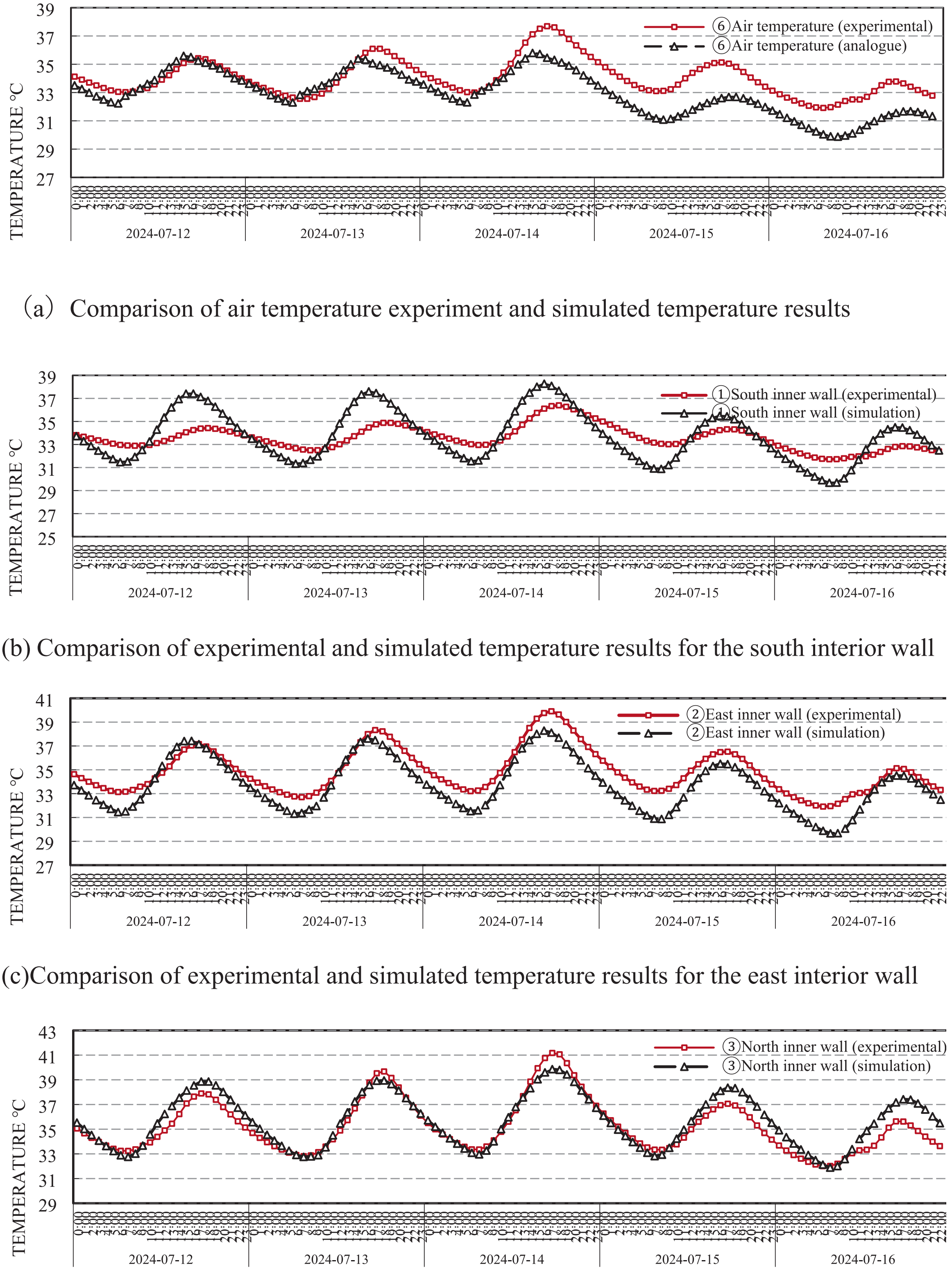
The experimental validation reveals close agreement between measured and simulated thermal performance for most measurement locations, as demonstrated in Fig. 11. The comparison shows measured versus simulated temperature results for air temperature and various wall surface locations (south interior wall, east interior wall, north interior wall, and west interior wall). Measurement points ①, ②, ③, ④, and ⑥ show generally consistent agreement between experimental and simulation results, with maximum temperature differences appearing at measurement point ①.

Furthermore, error analysis here employs standard statistical metrics including mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE) to quantify the agreement between simulation and experimental results. Table 6 presents the error analysis results for all measurement locations.

The error analysis demonstrates acceptable agreement between simulation and experimental results, with RMSE values ranging from 0.80°C to 1.70°C across all measurement locations. The mean absolute error values remain below 1.5°C for all measurement points, indicating good predictive capability of the thermal modeling approach. Beyond mean error metrics, comprehensive uncertainty analysis was conducted to quantify prediction reliability under various operational scenarios. The experimental validation dataset comprised 720 hourly measurements collected over a 30-day period, enabling detailed assessment of prediction uncertainty across different operational conditions. We calculated 95% confidence intervals for prediction errors, which ranged from ± 2.1°C to ± 2.8°C depending on the operational mode, with narrower intervals during steady-state conditions and wider intervals during transient periods or sudden weather changes. The prediction errors were analyzed as a function of key operational parameters including outdoor temperature, solar irradiance, and system operation status. Results indicate that prediction accuracy is highest during moderate outdoor temperatures (15-25°C) with mean absolute errors of 0.8–1.1°C, while extreme temperature conditions (below −10°C or above 35°C) result in slightly elevated errors of 1.8–2.2°C. Solar irradiance variations affect prediction accuracy primarily on facades with BIPV integration, where errors increase by approximately 15–20% during partially cloudy conditions compared to clear sky or fully overcast conditions, likely due to rapid fluctuations in PV surface temperatures. HVAC cycling behavior introduces the largest source of prediction uncertainty, with errors during the first 30 min following system startup averaging 2.4°C compared to 1.0°C during steady-state operation. These findings reveal that while the model maintains industry-acceptable accuracy overall, prediction reliability varies systematically with operational conditions, with steady-state operation under moderate weather conditions yielding the highest accuracy and transient periods under extreme conditions exhibiting the greatest uncertainty.

Analysis of the temporal distribution of errors reveals that the largest discrepancies occur during specific periods and conditions. For the south interior wall (measurement point ①, RMSE = 1.70°C), systematic overestimation of temperature fluctuations occurs primarily during midday hours (11:00–15:00) when solar radiation is maximum. Comparison with adjacent building geometry data indicates that the neighboring two-story structure casts shadows on the south facade during morning hours (8:00–10:00), reducing measured temperatures by approximately 2-3°C below simulated values during these periods. This shading effect is not captured in the simulation model, which assumes unobstructed solar access. Additionally, the adjacent building's continuous air conditioning operation creates a localized microclimate effect, with infrared thermography measurements showing exterior wall temperatures of the adjacent building approximately 1.5-2°C cooler than ambient conditions, indirectly affecting heat transfer to the experimental building.

For the west interior wall (measurement point ④, RMSE = 1.29°C), the systematic underestimation correlates strongly with outdoor unit operation schedules. Post-processing analysis of concurrent outdoor unit surface temperature measurements reveals temperatures reaching 45-50°C during peak cooling periods, approximately 10-15°C above



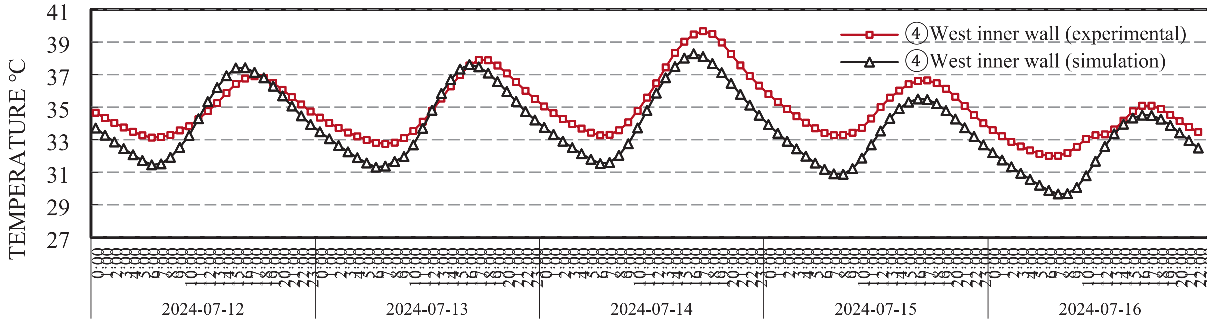
(a) Comparison of air temperature experiment and simulated temperature results

(b) Comparison of experimental and simulated temperature results for the south interior wall

(c)Comparison of experimental and simulated temperature results for the east interior wall

(d) Comparison of experimental and simulated temperature results for the north interior wall

Fig. 11. Comparison of experimental and simulated temperature results.



(e)Comparison of experimental and simulated temperature results for the west interior wall

Fig. 11. (*continued*).

Table 6
Building simulation validation results against sensor measurement throughout test period.

	MAE	MSE	RMSE
① South Interior Wall	1.44	2.88	1.70
② East Interior Wall	0.60	0.67	0.81
③ North Interior Wall	0.57	0.76	0.87
④ West Interior Wall	1.15	1.67	1.29
⑥ Air Temperature	0.68	0.64	0.80

ambient air temperature. This localized heat source, positioned 0.8 m from the west exterior wall, contributes an estimated 50–100 W of additional heat flux that is not represented in the baseline simulation model. The low thermal mass of the steel frame construction (estimated thermal capacitance 15 kJ/m^2·K compared to 150–200 kJ/m^2·K for conventional construction) results in rapid temperature response to these external heat sources, amplifying the discrepancy during afternoon hours when both solar gains and outdoor unit operation coincide.

Indoor air temperature discrepancies (measurement point ⑥, RMSE = 0.80°C) exhibit clear correlation with occupancy patterns. Time-series analysis shows that simulation overpredictions occur on weekdays when actual air conditioning is turned off but the simulation model applies continuous cooling based on standard office schedules. During occupied periods with active cooling, the agreement improves substantially with MAE reducing to 0.4°C. These findings highlight the importance of accurate operational schedule inputs for simulation accuracy, particularly for modular buildings with high envelope performance where internal load variations have proportionally larger impacts on indoor conditions.

Research by Willmott and Matsuura has demonstrated that mean absolute error provides superior assessment of simulation accuracy compared to standard error metrics, supporting the validation of the proposed modeling methodology [45]. The validation results confirm the feasibility and accuracy of both the photovoltaic modeling approach and the BIPV-integrated energy simulation methodology. The close agreement between simulated and measured performance data across diverse measurement locations and environmental conditions demonstrates the reliability of the proposed machine learning energy prediction framework for modular building applications with integrated photovoltaic systems.

5.2. *Performance of decomposition strategy and data-driven model*

5.2.1. *Comparison of data-driven model performance*

To evaluate model stability and generalization capability, we conducted 5-fold cross-validation for all machine learning models across the three prediction targets. Fig. 13 presents the cross-validation results of different data driven models, which plots the distribution of R^2 scores across five folds for (a) heating load, (b) cooling load, and (c) total energy consumption predictions. Each box represents the interquartile range, with the median shown as a horizontal line and the mean as a white diamond. Individual fold results are shown as colored points. XGBoost demonstrates superior performance with mean R^2 values exceeding 0.93 across all prediction targets and low variance with $\sigma < 0.012$, indicating good model stability. Decision trees show the highest variability ($\sigma = 0.015$–0.020), while Random Forest and ANN exhibit intermediate stability. The consistently narrow confidence intervals for XGBoost across all folds indicate that the model's superior performance is not dependent on specific data partitions, providing confidence in its generalization capability to unseen modular building configurations.

We further show the scattered plots to demonstrate he comprehensive performance comparison for the validation set in Table 8. The scattered plot evaluation encompasses decision trees, random forests (RF), artificial neural networks (ANN), and XGBoost algorithms applied to identical training and validation datasets to ensure fair comparison. Table 8 provides a comprehensive comparison of machine learning model performance, further showing XGBoost's superiority across all prediction targets.

The comparative model performance analysis on the validation set shows that XGBoost achieves the highest R^2 values across all prediction targets: heating loads (R^2 = 0.9763), cooling loads (R^2 = 0.9466), and total energy consumption (R^2 = 0.9374), significantly outperforming decision trees, random forests, and artificial neural networks. Decision trees achieve moderate performance with R^2 values of 0.93, 0.77, and 0.78 respectively, while random forests demonstrate improved performance with R^2 values of 0.8627, 0.8547, and 0.8612. Artificial neural networks show strong performance for heating loads (R^2 = 0.9519) but lower performance for cooling loads and total energy consumption.

XGBoost's superior performance can be attributed to several algorithmic features that are particularly well-suited to modular building energy prediction. First, XGBoost's L1 and L2 regularization mechanisms prevent overfitting when handling the repetitive structural patterns inherent in modular construction, where identical module configurations appear multiple times across different building assemblies. This regularization is critical because the six-surface property encoding system generates highly correlated features for similar module types, and XGBoost's penalty terms effectively manage these correlations without losing predictive power. Second, XGBoost's tree-based architecture naturally handles the mixed categorical and continuous variables in our feature set, including surface types (T0-T3), boundary conditions (B0-B2), and window-wall ratios (WWR) without requiring complex preprocessing or one-hot encoding expansion that can degrade neural network performance. Third, the gradient boosting framework iteratively corrects prediction errors from previous trees, which is particularly effective for capturing the non-linear inter-module thermal interactions at module interfaces, where thermal bridge effects create complex heat transfer patterns that simpler models struggle to represent. Finally, XGBoost's column subsampling and row subsampling features reduce overfitting risk when training on datasets with strong feature dependencies, such as the inherent relationships between adjacent module surfaces in our decomposition strategy.

To establish statistical rigor in performance comparisons, we conduct comprehensive significance testing using bootstrap resampling and paired t-tests. All comparisons between XGBoost and baseline models yield p-values below 0.001, indicating statistically significant superior performance at high confidence level. Detailed statistical testing procedures and complete confidence intervals are provided in Appendix C.

5.2.2. *Cross-climate performance*

The validation of the modular building decomposition strategy demonstrates exceptional performance across diverse climate conditions, confirming the feasibility and accuracy of the proposed individual module modeling approach. Fig. 12 presents a comparative analysis between decomposed modeling using individual module predictions and integrated modeling whole-building simulation approaches across ten different module configurations—ranging from simple two-module assemblies to complex four-module arrangements—in four representative climate zones: severe cold (Harbin), cold (Beijing), hot summer and cold winter (Shanghai), and hot summer and warm winter (Shenzhen). Each panel (a) through (j) shows the cooling load, heating load, and total energy comparisons for a specific module configuration.

The decomposed modeling results show consistent underestimation of annual cooling and heating loads compared to integrated modeling, with deviations maintained within the industry-acceptable range of ± 10% as specified by the Design Standard for Heating, Ventilation and Air Conditioning of Civil Buildings in China (GB50736). The analysis reveals distinct climate-dependent performance characteristics that reflect regional variations in thermal behavior and environmental conditions.

In cold climate regions (Beijing and Harbin), the decomposed modeling approach exhibits cooling load prediction errors of approximately 8%-10% below integrated modeling results, primarily attributed to simplified treatment of inter-module shading and ventilation coupling effects. The underestimation is most pronounced during heating-dominated periods when thermal bridge effects at module interfaces become critical for accurate energy prediction. The heating load differences in these regions range from 6% to 8%, reflecting the linear superposition characteristics of heating loads under steady-state conditions.

Hot and humid climate regions (Shenzhen) demonstrate superior agreement between decomposed and integrated modeling approaches, with cooling load deviations constrained within 5%. This improved performance reflects the air conditioning load-dominated thermal characteristics typical of subtropical climates, where cooling loads exhibit more linear additive behavior suitable for decomposed modeling approaches. The reduced inter-module thermal interactions during cooling-dominated operation contribute to the enhanced accuracy in these climate zones.

The transitional climate region (Shanghai) exhibits balanced cooling and heating load differences within ± 6%-8%, corresponding to the dynamic thermal environment characteristics of the hot summer and cold winter zone. The relatively uniform deviation across both heating and cooling loads indicates consistent model performance during transition seasons when thermal loads fluctuate between heating and cooling requirements. Table 7 provides quantitative assessment of machine learning model accuracy across four different climate zones, demonstrating prediction errors consistently below 5% for all climate regions and energy prediction targets.

The error analysis reveals exceptional prediction accuracy with maximum deviations of 7.96% for cooling loads in the severe cold climate zone (Harbin) and minimum deviations below 1% for total energy consumption in multiple climate zones. The heating load predictions demonstrate particularly high accuracy across all climate zones, with maximum errors of 6.11% and minimum errors of 0.61%. This superior heating load prediction performance reflects the more predictable nature of conductive heat transfer mechanisms compared to the complex thermal dynamics associated with cooling system operation.

Total energy consumption predictions achieve remarkable accuracy with deviations consistently below 1.5% across all climate zones, demonstrating the effectiveness of the comprehensive feature engineering approach that captures both thermal and geometric characteristics of modular building systems. The Shanghai climate zone shows the most balanced performance with deviations around 1% for all prediction targets, indicating optimal model calibration for transitional climate conditions.

5.2.3. *Uncertainty analysis*

Quantitative uncertainty bounds were established to provide designers with actionable information about prediction reliability for different building configurations. Using bootstrapping techniques with 1,000 resampling iterations, we calculated 90% prediction intervals for various module types as presented in Table 9.

The results reveal that prediction uncertainty varies systematically with module configuration characteristics. Standard module configurations without complex shading or unusual boundary conditions exhibit the lowest uncertainty (±1.8 kWh/m^2 annually, approximately ± 4% of typical annual energy consumption), establishing a baseline for prediction reliability. Module configurations with high window-wall ratios

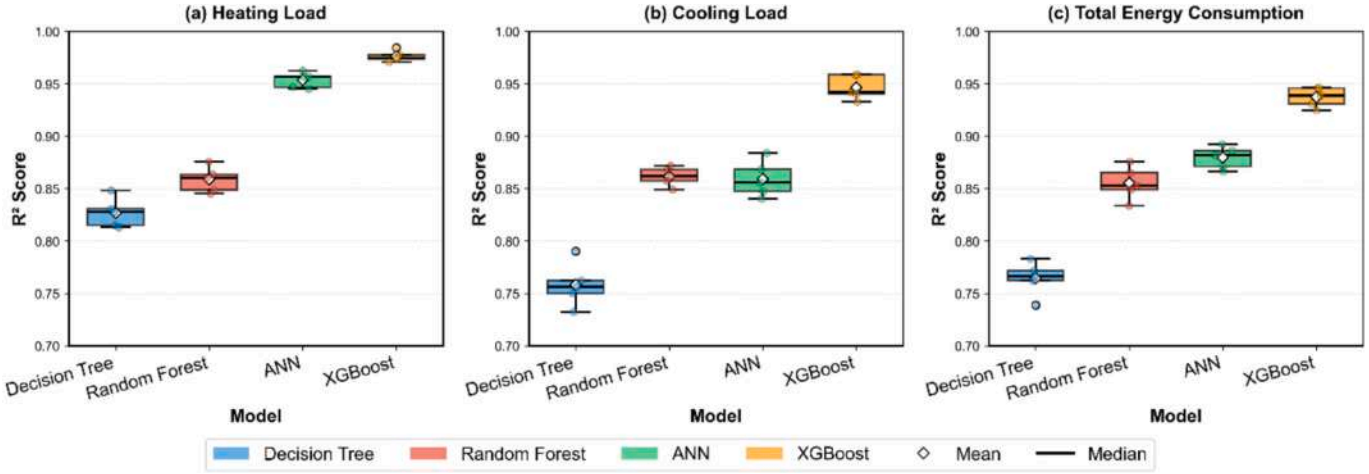
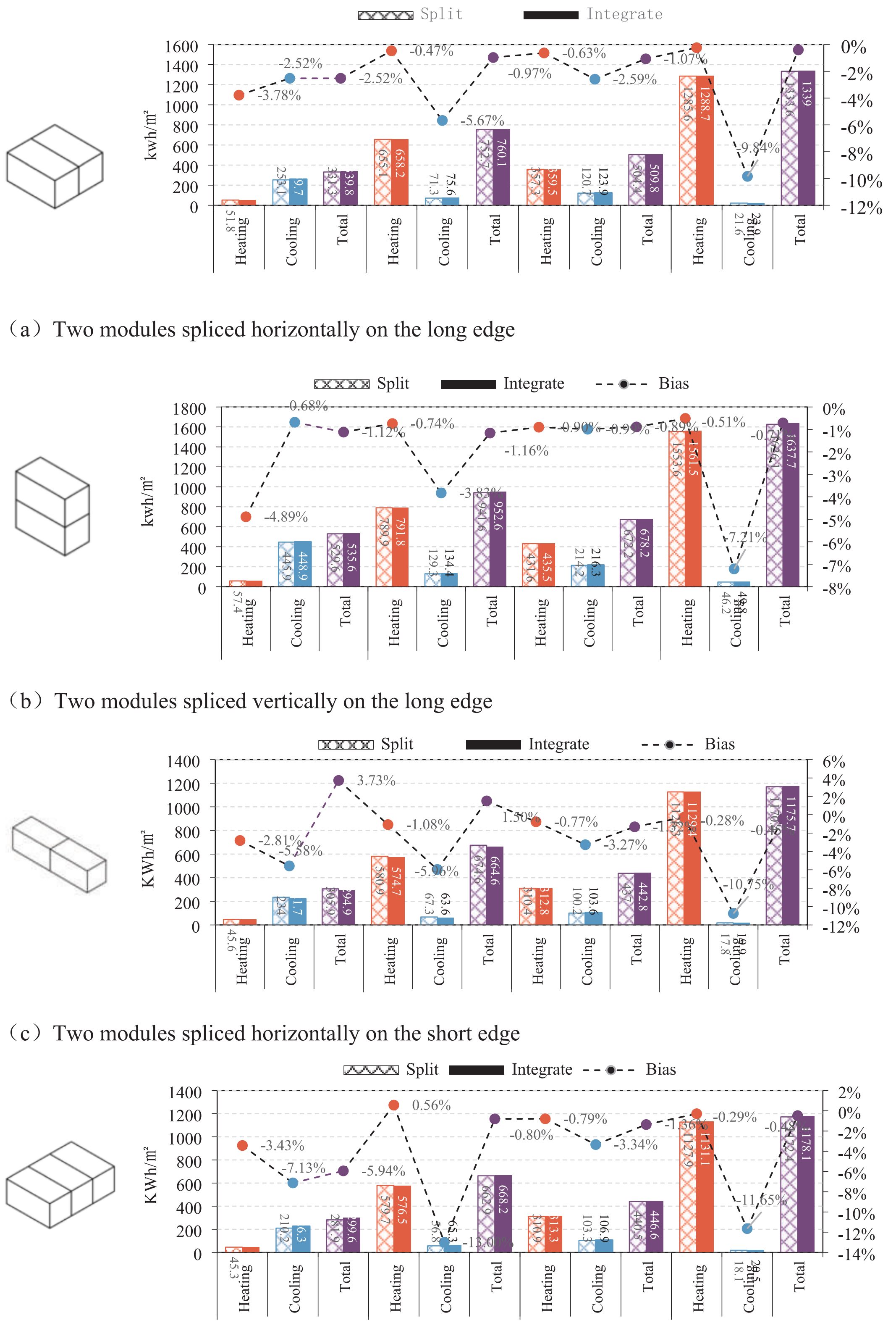


Fig. 12. Analysis of the simulated and data-driven predicted annual results among various split modeling and holistic modeling.

(WWR > 0.5) exhibit the widest prediction intervals (±2.4 kWh/m^2) due to increased sensitivity to solar gain and thermal losses, representing a 33% increase in uncertainty compared to standard configurations. Corner modules with multiple exterior surfaces demonstrate intermediate uncertainty (±2.1 kWh/m^2), reflecting the additional complexity of multi-directional heat transfer. Modules with BIPV integration show prediction intervals of ± 2.3 kWh/m^2 annually, with seasonal variations showing higher uncertainty during summer months



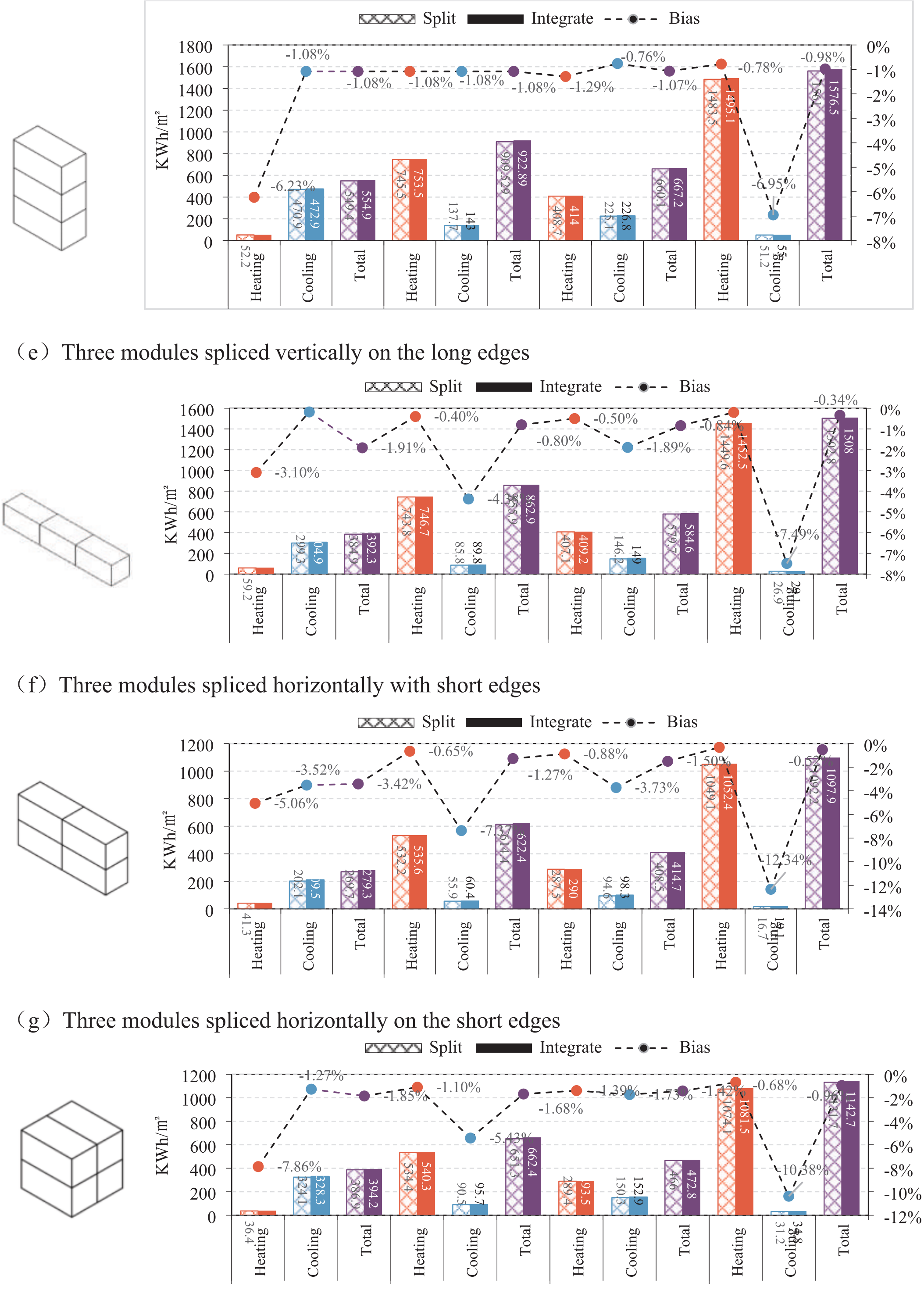
（a）Two modules spliced horizontally on the long edge

（b）Two modules spliced vertically on the long edge

（c）Two modules spliced horizontally on the short edge

（d）Three modules spliced horizontally on the long edge

Fig. 13. Five-fold cross-validation performance comparison across machine learning models.



(e) Three modules spliced vertically on the long edges

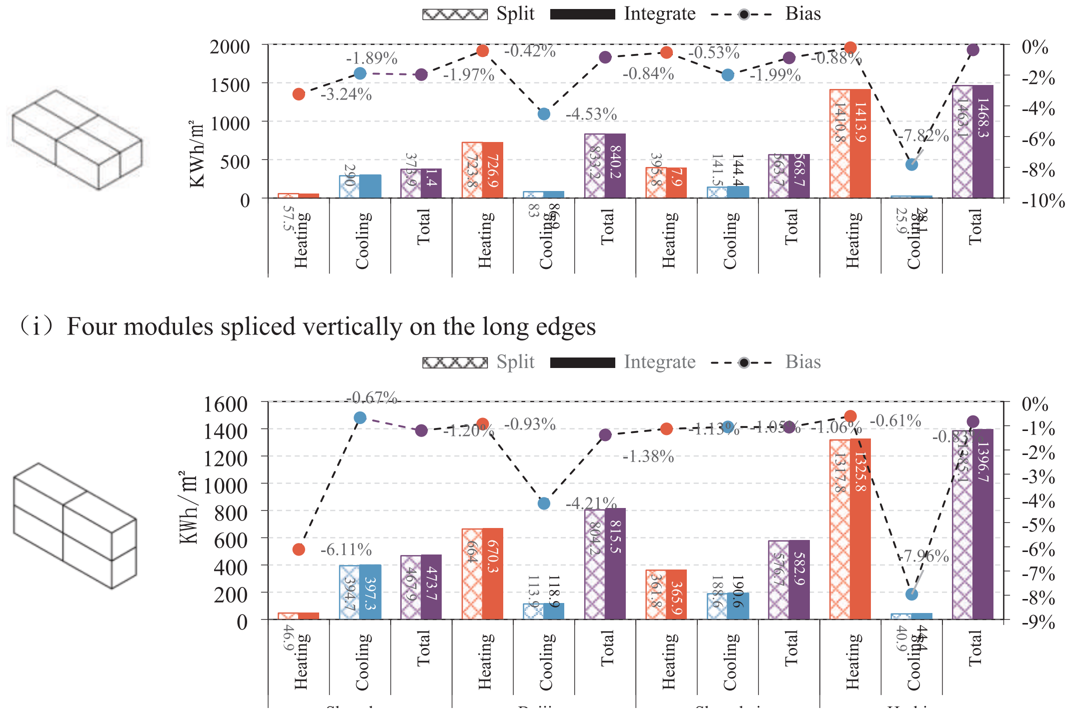
(f) Three modules spliced horizontally with short edges

(g) Three modules spliced horizontally on the short edges

(h) Four modules spliced vertically on the long edges

Fig. 13. (*continued*).

when PV thermal effects are most pronounced. Ground-contact modules and top-floor modules with roof assemblies exhibit prediction intervals of ± 1.9 and ± 2.0 kWh/m^2 respectively, falling within the expected range for configurations with specialized thermal boundary conditions. Analysis of prediction residuals reveals that 92% of validation cases fall within the 90% prediction interval, confirming appropriate calibration of uncertainty estimates. The distribution of prediction errors approximates a normal distribution, with slight positive skewness indicating a marginal tendency toward underprediction of energy consumption. These uncertainty bounds enable risk-informed decision-making during early design stages, allowing designers to assess whether prediction accuracy is sufficient for specific design optimization objectives.



(i) Four modules spliced vertically on the long edges

(j) Four modules spliced vertically on the long edges

Fig. 13. (*continued*).

Table 7
Machine learning model error analysis against EnergyPlus simulation results.

Climate Zone	Prediction Type	Heating Load (kWh/m^2)	Deviation	Cooling Load (kWh/m^2)	Deviation	Total Energy (kWh/m^2)	Deviation
Shenzhen	Predicted	62.53	−6.11%	394.67	−0.68%	467.95	−1.20%
	Actual	49.96		397.32		473.65	
Beijing	Predicted	664.01	−0.93%	113.87	−4.21%	804.25	−1.37%
	Actual	670.25		118.88		815.49	
Shanghai	Predicted	361.8	−1.13%	188.6	−1.05%	576.7	−1.06%
	Actual	365.9		190.6		582.9	
Harbin	Predicted	1318.8	−0.61%	40.9	−7.96%	1385.1	−0.83%
	Actual	11325.8		44.4		1396.7	

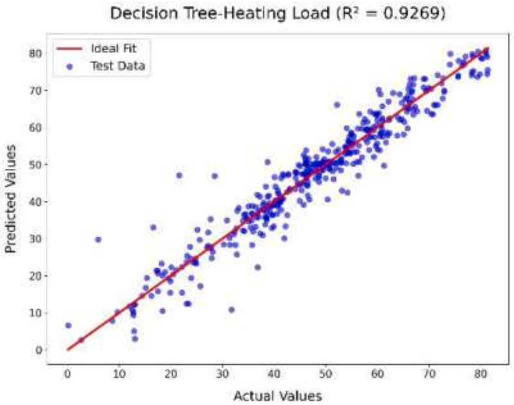
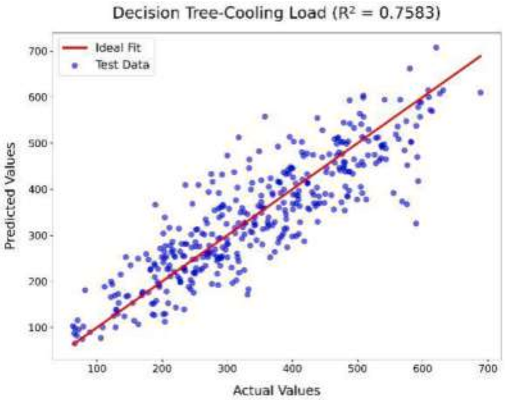
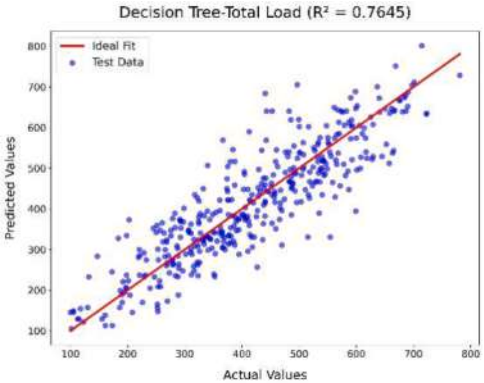
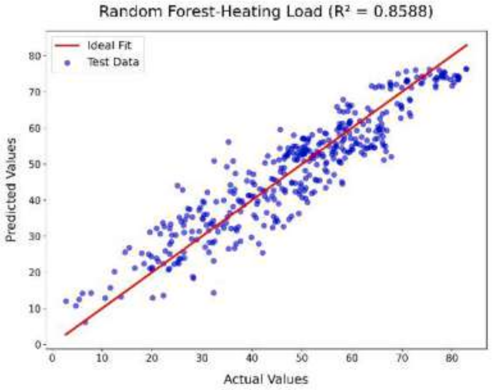
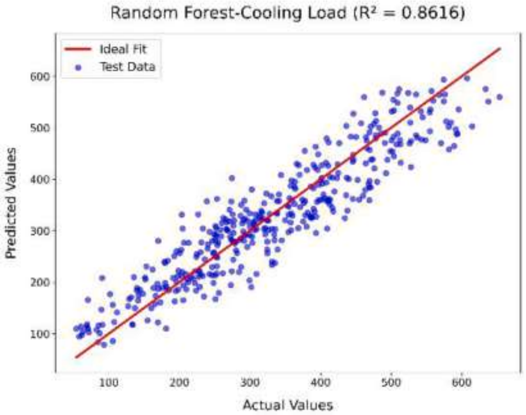
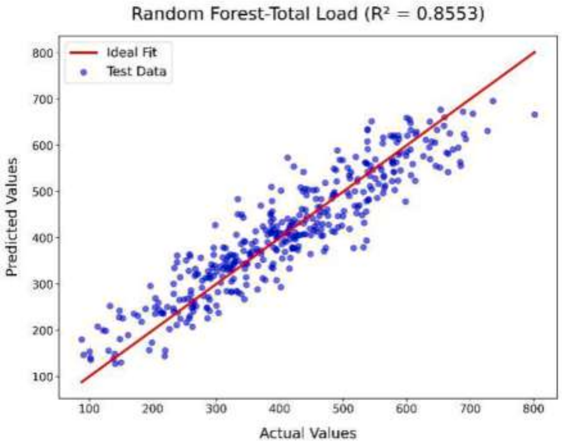
5.3. Computational efficiency

Computational efficiency assessment for the developed machine learning model demonstrates valid performance improvements compared to traditional physics-based simulation approaches, validating the practical applicability of the machine learning methodology for real-time design applications. The efficiency evaluation encompasses both training phase computational requirements and operational prediction performance to provide comprehensive assessment of practical implementation characteristics.

Training phase computational requirements include dataset generation, feature engineering, and model training procedures. The complete dataset generation process requires approximately 24 h using Honeybee and Grasshopper platforms for 5,000 building configurations across multiple climate zones. Individual simulation completion times average 20–30 s per configuration, reflecting the computational intensity of physics-based energy modeling that necessitates the development of rapid prediction alternatives. Model training procedures complete within 2–3 h using standard desktop computing hardware (AMD Ryzen 7 5800H, 32 GB RAM, NVIDIA GeForce RTX 3060), demonstrating reasonable training requirements that do not impose excessive computational burdens for practical implementation. The training time scales approximately linearly with dataset size, enabling larger training datasets for enhanced model accuracy when computational resources permit. Moreover, operational prediction performance achieves good efficiency with individual building energy predictions completing within millisecond timeframes. Single module energy predictions require less than 1 ms on standard desktop hardware, representing computational speed improvements exceeding 2,000 × compared to physics-based simulation approaches. This dramatic efficiency enhancement enables real-time design iteration and multi-objective optimization applications that are impractical with traditional simulation methodologies.

Computational efficiency assessment reveals particularly significant advantages for parametric design applications requiring evaluation of multiple design alternatives. Traditional physics-based optimization requiring 1,000 building energy evaluations would consume approximately 5.5 h using EnergyPlus simulation, while the machine learning approach completes equivalent analysis within 1 s. This efficiency improvement transforms the practical feasibility of comprehensive design optimization during early design phases when rapid iteration is most valuable. Memory requirements for the trained XGBoost models remain modest, with complete model storage requiring less than 10 MB for all climate zones and prediction targets. This compact model representation enables integration into parametric design software without significant memory overhead or performance degradation. The small

Table 8
Data-driven model performance comparison for the validation set.

Model	Heating Load (kWh/m^2)	Cooling Load (kWh/m^2)	Total Energy Consumption (kWh/m^2)
Decision Tree			
R^2	0.9269	0.7583	0.7645
R F			
R^2	0.8588	0.8616	0.8553

(*continued on next page*)

Table 8 (*continued*)

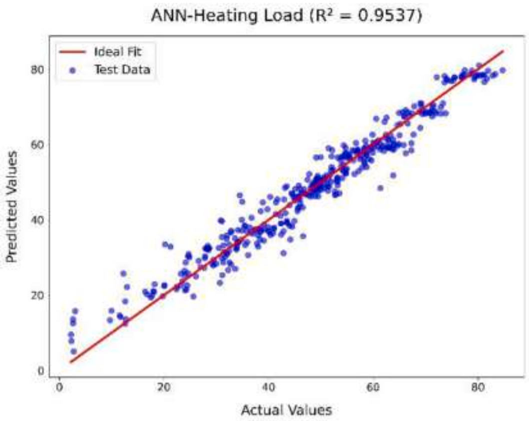
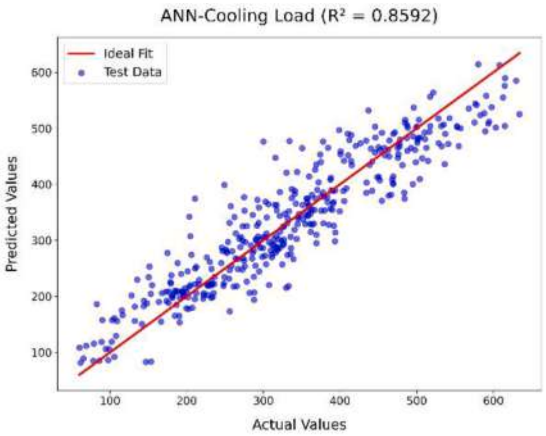
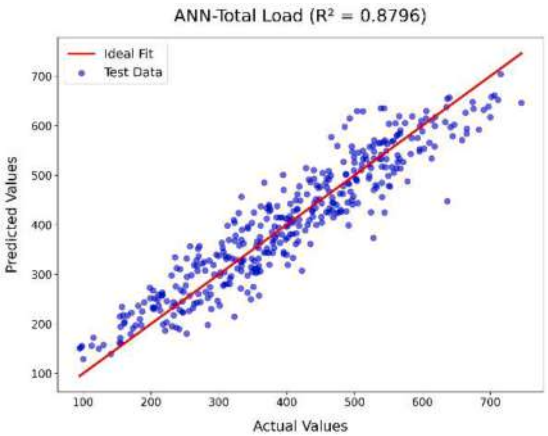
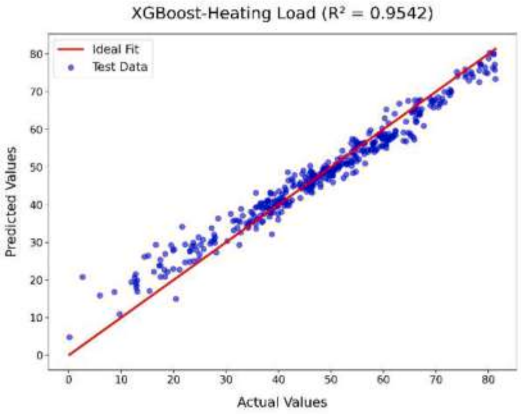
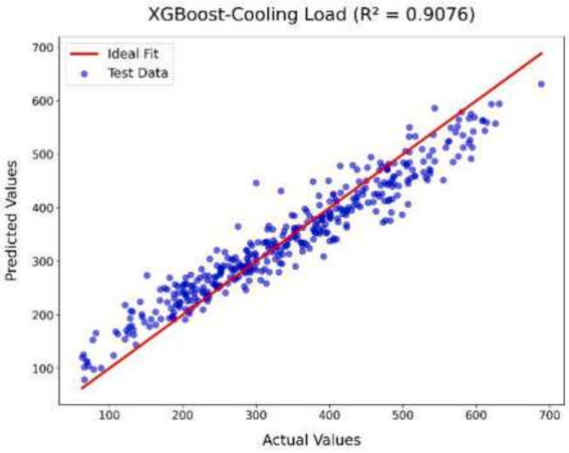
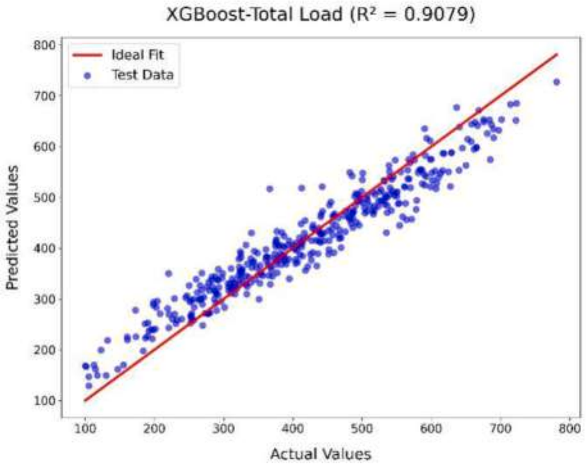
Model	Heating Load (kWh/m^2)	Cooling Load (kWh/m^2)	Total Energy Consumption (kWh/m^2)
ANN	 ANN-Heating Load (R^2 = 0.9537)	 ANN-Cooling Load (R^2 = 0.8592)	 ANN-Total Load (R^2 = 0.8796)
R^2	0.9537	0.8592	0.8796
XGBOOSt	 XGBoost-Heating Load (R^2 = 0.9542)	 XGBoost-Cooling Load (R^2 = 0.9076)	 XGBoost-Total Load (R^2 = 0.9079)
R^2	0.9542	0.9076	0.9079

Table 9
Prediction uncertainty bounds for different module configurations.

Module Configuration	90% Prediction Interval (kWh/m^2/year)	Relative Uncertainty (% of typical consumption)	Potential Uncertainty Source
Standard configuration (moderate WWR, simple boundaries)	±1.8	±4%	Baseline variability
High window-wall ratio (WWR > 0.5)	±2.4	±5.5%	Solar gain sensitivity
Corner modules (multiple exterior surfaces)	±2.1	±4.7%	Multi-directional heat transfer
BIPV-integrated modules	±2.3	±5.2%	PV thermal effects (higher in summer)
Ground-contact modules	±1.9	±4.2%	Soil temperature variations
Top-floor modules with roof	±2.0	±4.5%	Roof thermal performance

Note: 92% of validation cases fall within the 90% prediction interval. Residual distribution: approximately normal (Shapiro-Wilk test, p = 0.18) with slight positive skewness of 0.24.

model size also facilitates model deployment across different computing platforms and enables cloud-based design services.

5.4. Feature importance analysis

5.4.1. Model-Specific feature importance patterns

Fig. 14 presents the comparative feature importance analysis across the three XGBoost models (cooling load, heating load, and total energy consumption), with horizontal bar charts showing the relative importance scores for each input feature. The figure reveals distinct patterns that reflect the underlying physical mechanisms governing different types of building energy loads in BIPV-integrated modular buildings.

The cooling load model demonstrates pronounced emphasis on the Top Surface (Roof), which emerges as the dominant feature with the highest importance. This pattern aligns with cooling-dominated building performance, where roof surfaces experience maximum solar heat gain due to horizontal orientation and direct solar exposure throughout daylight hours. The substantial roof importance reflects its critical role in determining cooling loads through both direct solar heat gain and BIPV system thermal performance. The relatively lower importance of facade surfaces indicates that roof thermal performance significantly outweighs wall contributions for cooling energy consumption.

The heating load model exhibits a markedly different hierarchy, with Bottom Surface (Floor) emerging as the most critical feature. This reflects fundamental heating load physics, where ground-coupled heat transfer through floor assemblies represents the primary thermal loss mechanism during heating periods. The substantial floor importance demonstrates the critical role of ground thermal coupling in determining heating energy requirements, particularly in modular construction where elevated assemblies may experience enhanced thermal losses

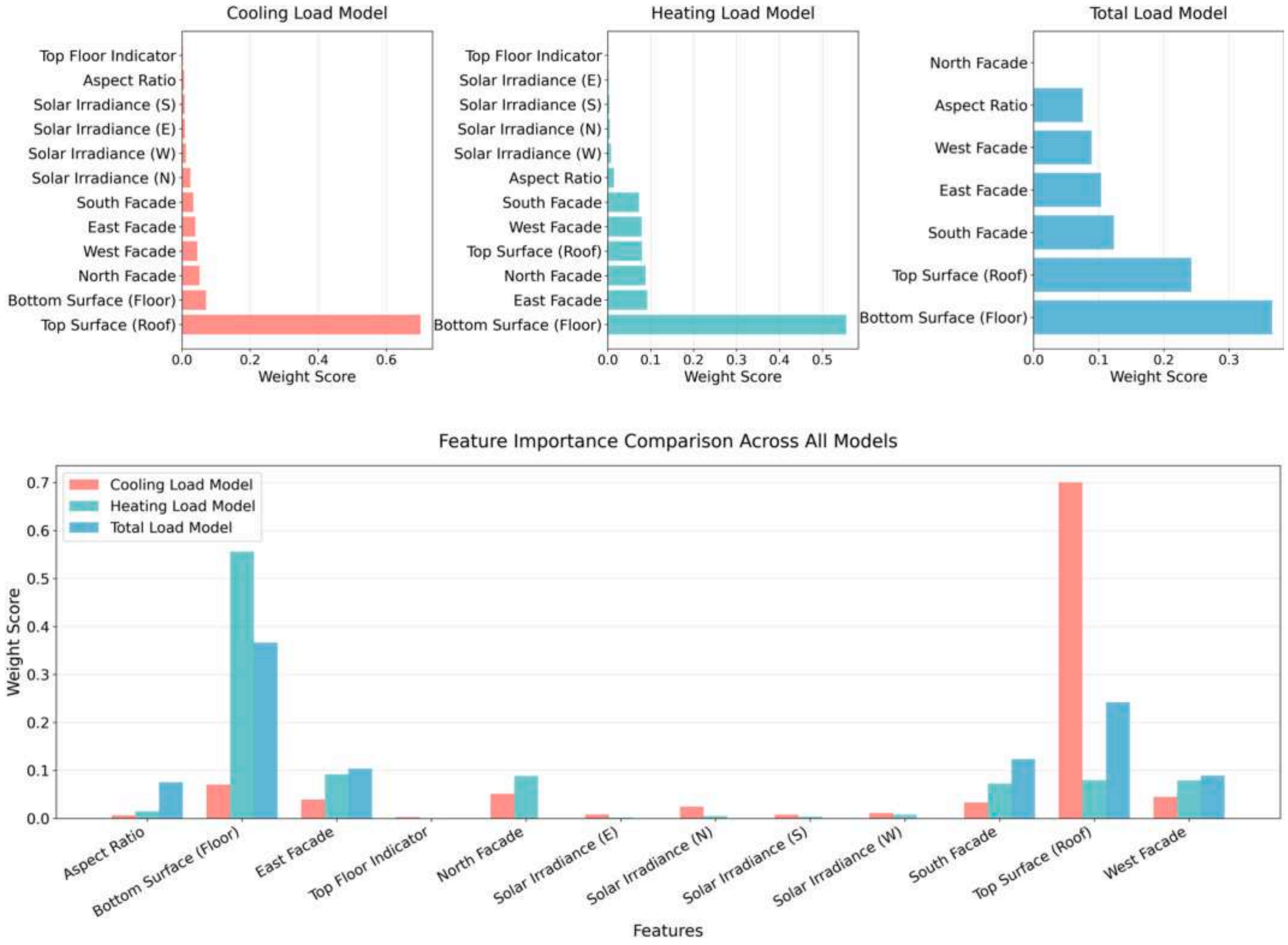


Fig. 14. Feature importance analysis for the three trained XGBoost model.

compared to conventional slab-on-grade configurations. The heating model also shows significant facade surface importance, reflecting increased envelope thermal performance relevance during heating periods when indoor-outdoor temperature differentials are maximized.

The total load model presents balanced importance distribution reflecting combined heating and cooling effects throughout annual cycles. The model assigns substantial importance to both Top Surface and Bottom Surface, indicating that total energy consumption is influenced by summer cooling loads driven by roof solar gains and winter heating loads dominated by ground thermal coupling.

5.4.2. Cross-Model feature correlations

The feature importance heatmap plotted in Fig. 15 uses color intensity ranging from dark purple for low importance to bright yellow for high importance to reveal distinct correlation patterns between different energy load types and their sensitivity to various building characteristics. The heatmap visualization clearly illustrates how cooling and heating loads respond differently to identical building features, with some features showing strong model-specific importance while others maintain consistent relevance across all prediction targets.

Surface-related features exhibit the most pronounced model-specific variations, with cooling loads showing maximum sensitivity to roof thermal performance while heating loads demonstrate primary sensitivity to floor thermal coupling. This divergence reflects the directional nature of thermal transfer mechanisms, where upward heat flow through roof assemblies dominates cooling load generation while downward heat flow through floor systems drives heating load requirements. The facade surfaces show intermediate importance levels across all models, indicating their consistent but secondary role in determining energy consumption compared to horizontal surfaces.

Moreover, geometric parameters, represented by the Aspect Ratio feature, demonstrate moderate but consistent importance across all three models. This pattern indicates that building shape characteristics influence energy performance through multiple mechanisms including surface area to volume ratios, thermal bridge configurations, and natural ventilation potential. The consistent geometric importance across different load types suggests that modular building proportions affect both heating and cooling performance through fundamental building physics relationships.

5.4.3. Solar irradiance feature dependencies

Our analysis reveals a notable absence of solar irradiance features among the top-ranking importance factors in the total load model, which raises important questions about the model's representation of solar thermal effects. While the individual cooling and heating load models may implicitly capture solar effects through surface property encodings, the total load model appears to rely primarily on envelope thermal performance characteristics rather than direct solar radiation inputs. This pattern suggests that the total load model may be capturing solar effects through indirect mechanisms embedded within the surface property encodings rather than treating solar irradiance as independent variables. The six-surface property encoding system described in the methodology incorporates construction settings that include BIPV integration, which may inherently account for solar thermal effects without requiring separate irradiance variables. This could potentially create feature collinearity between the categorical construction settings (C0-C3) and continuous irradiance values (D0-D9), where both features convey overlapping information about solar exposure. XGBoost's tree-based algorithm tends to favor the categorical construction settings when both provide similar predictive power, resulting in lower apparent importance for direct irradiance measurements despite their physical significance. Additionally, the spatial shading equivalent radiation methodology employed in the feature engineering process may have preprocessed solar effects into the surface property classifications, reducing the apparent importance of direct solar measurements.

The reduced prominence of solar irradiance features in the total load model may also reflect the balancing effects of heating and cooling loads throughout annual operation cycles. During cooling periods, increased solar irradiance directly increases cooling loads through enhanced heat gain, while during heating periods, increased solar irradiance can reduce heating loads through beneficial solar gains. These opposing effects may result in solar irradiance features showing lower apparent importance in total load prediction compared to envelope thermal properties that consistently influence energy consumption in the same direction regardless of season. This observation highlights the complexity of feature importance interpretation in multi-seasonal energy prediction models and suggests that future model development should investigate the explicit inclusion of temporal solar irradiance patterns to enhance model interpretability and ensure proper representation of BIPV system interactions with building thermal performance.

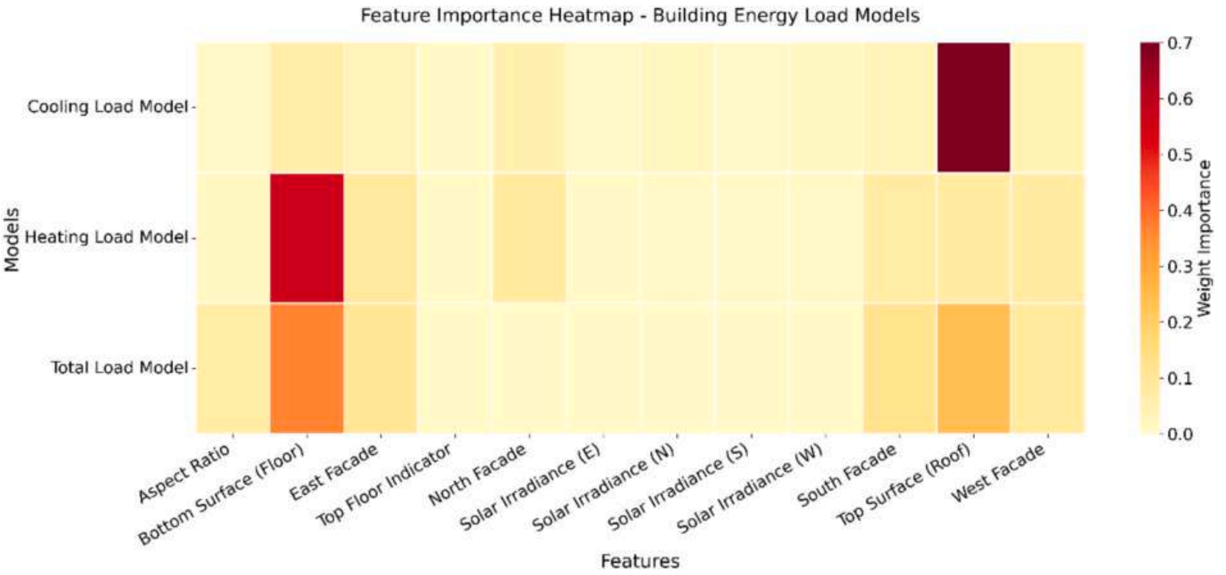


Fig. 15. Summarized heatmap presentation for all features in different models.

5.5. *Integration with parametric design platform*

The integration of machine learning energy prediction capabilities with parametric design platforms represents a significant advancement in building design workflow efficiency and capability. Fig. 16 illustrates the implementation of the machine learning model within the Grasshopper environment, demonstrating seamless integration that enables real-time energy feedback during design development processes.

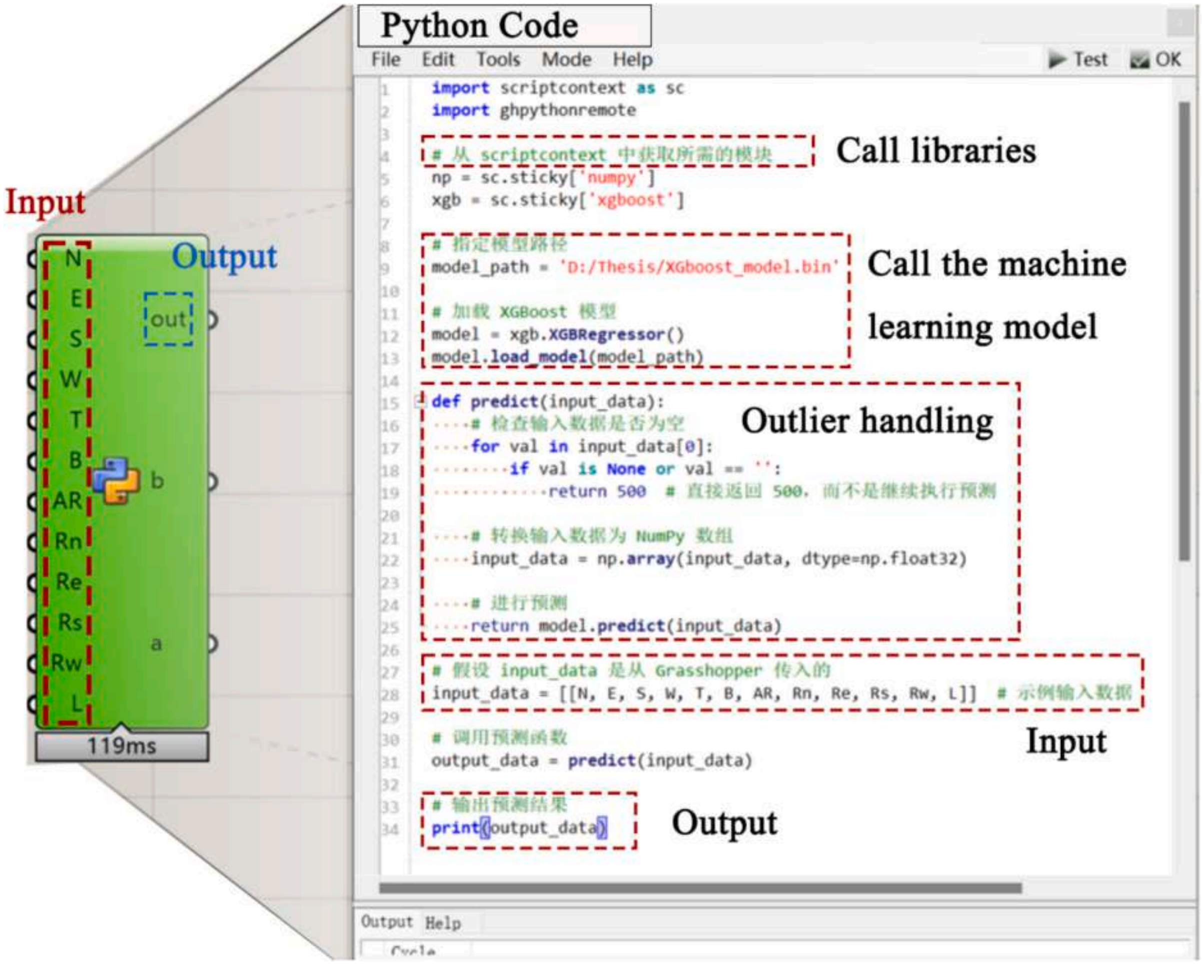
The Python scripting component visible in Fig. 16 requires three primary input categories from the parametric model: geometric parameters including module dimensions, aspect ratios, and top floor indicators extracted from the Rhino 3D geometry, surface properties including six-surface encodings including surface types T0-T3, boundary conditions B0-B2, construction settings C0-C3, and WWR specified through user interface sliders, and environmental parameters including solar irradiance values D0-D9 calculated from site location and building orientation. The script packages these inputs into the feature vector format expected by the trained XGBoost model, executes the prediction using the stored model file, and returns three output values: predicted heating load, cooling load, and total energy consumption in kWh/m^2. These predictions are immediately visualized in the Grasshopper canvas through numerical display panels and can be connected to optimization components for automated design space exploration. Designers use this real-time feedback to iteratively adjust module configurations, window-wall ratios, and BIPV placement while observing immediate energy performance implications, enabling energy-informed decisions during early design stages when modifications are most cost-effective.

The integration methodology employs Python scripting within Grasshopper to access the trained XGBoost models through remote component interfaces. This approach enables direct model execution within the parametric design environment without requiring external software dependencies or complex data transfer protocols. The integration maintains full compatibility with existing Grasshopper workflows while adding comprehensive energy prediction capabilities.

Real-time performance assessment demonstrates responsiveness with energy predictions updating immediately as geometric parameters are modified within the Grasshopper interface. This immediate feedback capability enables designers to observe energy performance implications of design modifications in real-time, facilitating energy-informed design decisions throughout the conceptual design phase. The rapid response enables iterative design exploration that was previously impractical due to computational limitations of physics-based simulation. The integration supports both single module analysis and complete building assessment through automatic aggregation of individual module predictions. This scalability enables application across diverse project scales from small residential modules to large-scale modular construction projects without computational performance degradation. Parametric optimization integration enables automated design space exploration through integration with genetic algorithm and multi-objective optimization tools available within the Grasshopper ecosystem.

5.6. *Limitations and Future Directions*

While this research demonstrates effective machine learning-based energy prediction for BIPV-integrated modular buildings, several



```python
import scriptcontext as sc
import ghpythonremote

# 从 scriptcontext 中获取所需的模块
np = sc.sticky['numpy']
xgb = sc.sticky['xgboost']

# 指定模型路径
model_path = 'D:/Thesis/XGboost_model.bin'

# 加载 XGBoost 模型
model = xgb.XGBRegressor()
model.load_model(model_path)

def predict(input_data):
    # 检查输入数据是否为空
    for val in input_data[0]:
        if val is None or val == '':
            return 500  # 直接返回 500，而不是继续执行预测

    # 转换输入数据为 NumPy 数组
    input_data = np.array(input_data, dtype=np.float32)

    # 进行预测
    return model.predict(input_data)

# 假设 input_data 是从 Grasshopper 传入的
input_data = [[N, E, S, W, T, B, AR, Rn, Re, Rs, Rw, L]]  # 示例输入数据

# 调用预测函数
output_data = predict(input_data)

# 输出预测结果
print(output_data)
```

Fig. 16. Machine learning model integration in Grasshopper platform of Rhino 7.

limitations warrant discussion. The model's scalability to very large buildings with hundreds of modules has not been extensively validated. Although the decomposition strategy theoretically supports aggregation of numerous individual modules, the cumulative prediction errors and computational memory requirements for extremely large assemblies (exceeding 50–100 modules) remain unexplored and may require hierarchical prediction strategies or model ensemble approaches to maintain accuracy. The current framework is optimized for standardized rectangular modules with consistent dimensions and construction specifications. Non-standard modules with irregular geometries, curved surfaces, or highly customized construction assemblies may not be adequately represented by the existing six-surface encoding system, requiring extension of the feature engineering framework with additional geometric descriptors or surface subdivision strategies. Modules with significantly different thermal properties or unconventional BIPV integration configurations may fall outside the model's training distribution, necessitating either model retraining with expanded datasets or development of transfer learning approaches.

Moreover, while the model demonstrates strong cross-climate performance in simulation-based validation across four climate zones, additional experimental validation across diverse geographic locations and building typologies would strengthen confidence in real-world applicability. The current approach focuses on steady-state annual energy predictions and does not explicitly model transient thermal behavior or short-term load forecasting, limiting its applicability for real-time building control that requires hourly or sub-hourly predictions. The model also assumes typical occupancy patterns and does not account for atypical user behaviors or operational scenarios that may significantly deviate from standard conditions. Future research should address these limitations by validating model performance on larger building assemblies, extending the feature engineering framework for non-standard geometries, conducting multi-site experimental validations, and developing temporal prediction capabilities for building control applications.

6. Conclusions

This research presents a comprehensive machine learning-based rapid energy prediction method specifically designed for BIPV-integrated modular buildings, addressing critical computational limitations of traditional physics-based simulation approaches while leveraging the unique structural characteristics of modular construction systems. Some major findings in this research are as follows:

- The proposed modular building decomposition approach achieves prediction accuracy within ± 10% compared to integrated modeling across various climate zones, with R^2 values exceeding 0.90, demonstrating the feasibility of individual module analysis for system-level energy prediction.
- The XGBoost-based model significantly outperforms alternative approaches, achieving R^2 values of 0.9763, 0.9466, and 0.9374 for heating loads, cooling loads, and total energy consumption respectively, compared to maximum R^2 values of 0.93, 0.86, and 0.87 from competing algorithms.
- Prediction speeds exceed 2,000 × faster than traditional physics-based simulation, with individual predictions completing within milliseconds, enabling real-time design iteration and comprehensive optimization during conceptual design phases.
- Case study validation using the C-Smart building demonstrates mean absolute errors below 1.5°C and RMSE values of 0.80–1.70°C across all measurement locations, confirming model accuracy and reliability.
- Feature importance analysis reveals cooling loads dominated by roof surfaces, heating loads by floor surfaces, and total loads showing balanced roof-floor importance, with solar irradiance features having reduced prominence due to opposing seasonal effects on heating versus cooling demands.

The implementation of our proposed method with Grasshopper parametric design platforms provides immediate energy feedback, facilitating energy-informed design decisions and supporting widespread adoption of sustainable modular construction practices. Future research should explore applications to other building typologies and integration with advanced building control systems to further enhance practical value for sustainable building design. The advancement presented in this work overcomes processing limitations to energy performance optimization for modular buildings and supports the broader adoption of sustainable modular construction practices by providing practical tools for energy-informed design decision-making.

CRediT authorship contribution statement

Yiqian Zheng: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Biao Yang:** Validation, Project administration, Formal analysis, Data curation. **Miaomiao Hou:** Supervision, Resources, Methodology, Investigation, Funding acquisition. **Yi Zhang:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation. **Yuekuan Zhou:** Writing – review & editing, Visualization, Resources, Project administration. **Xing Zheng:** Project administration, Methodology, Data curation. **Pengyuan Shen:** Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

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

Acknowledgement

This research is partly supported by the Shenzhen Fundamental Research Program JCYJ20250604180231041.

Appendix

APPENDIX A: Data generation and constraints handling.

A.1 Latin Hypercube sampling methodology

The generation of comprehensive training datasets for modular building energy prediction requires systematic sampling across the multi-dimensional parameter space defined by the feature engineering framework. Latin Hypercube Sampling (LHS) is employed to ensure efficient and representative coverage of the parameter space while minimizing the total number of required simulations. LHS provides superior space-filling properties compared to random sampling, ensuring that each input variable is sampled uniformly across its entire range.

For modular building applications, the LHS implementation addresses 12 primary input parameters including six-surface property encodings, geometric characteristics, and solar irradiance values as systematically detailed in the previous Table 3. The selection of samples balances three competing requirements: adequate coverage of the multi-dimensional parameter space, computational cost of physics-based simulations, and machine learning training data requirements (XGBoost typically requires thousands of samples for robust generalization across diverse conditions). Our preliminary testing with smaller sample sizes (1,000–2,000) resulted in inadequate representation of parameter space corners and reduced prediction accuracy, while larger sample sizes (10,000 +) provided diminishing returns in model performance relative to the doubled computational time. The 5,000-sample size also aligns with machine learning best practices for tabular data, where the sample-to-feature ratio of approximately 400:1 (5,000 samples for 12 primary features) ensures sufficient data density to prevent overfitting while capturing non-linear relationships. The sampling procedure generates 5,000 unique parameter combinations, providing sufficient diversity for robust model training while maintaining computational feasibility for energy simulation execution using LHS. The mathematical foundation of LHS ensures uniform distribution across the parameter space:

$$x_{ij} = \frac{\pi_j(i) - u_{ij}}{n}$$

where x_{ij} represents the normalized parameter value for sample i and parameter j, $\pi_j(i)$ denotes a random permutation of integers from 1 to n for parameter j, and u_{ij} represents a random number between 0 and 1.

A.2 Variable combination and constraint handling

The variable combination procedure addresses the complex interdependencies between different building parameters while ensuring physical realism and constructability constraints. According to Table 3, geometric constraints ensure that window-wall ratios remain within feasible ranges for different surface types and orientations. The detailed encoding matrix shows that air boundary surfaces, representing open connections between modules, are constrained to zero window-wall ratios, as these surfaces cannot accommodate conventional fenestration systems. Similarly, ground contact surfaces are restricted to zero window-wall ratios due to their below-grade positioning.

Construction assembly constraints link surface types with appropriate construction specifications as detailed in Table 3. Wall surfaces can accommodate either standard wall assemblies (construction setting 0) or photovoltaic-integrated assemblies (construction setting 1), while roof surfaces are limited to roof-specific construction types (construction setting 2). Floor surfaces employ specialized assemblies designed for ground contact conditions (construction setting 3).

The constraint validation process systematically evaluates each generated parameter combination against predefined feasibility criteria, eliminating invalid combinations while preserving the uniform distribution characteristics of the LHS procedure. The research indicates that approximately 15% of initially generated combinations are rejected due to constraint violations, requiring iterative generation to achieve the target sample size of 5,000 valid configurations.

The validated parameter combinations are subsequently processed through physics-based energy simulation using Honeybee and Grasshopper platforms, to generate corresponding energy consumption values. This simulation process requires approximately 24 h for comprehensive dataset generation, with each individual simulation completing within 20–30 s. The resulting dataset provides the foundation for machine learning model training and validation across diverse modular building configurations and climate conditions.

APPENDIX B: Hyperparameter optimization procedures

Hyperparameter optimization employs a systematic grid search approach combined with Bayesian optimization techniques to identify optimal parameter configurations for different climate zones and building types. The optimization process considers multiple hyperparameters that significantly influence model performance, including learning rate, maximum tree depth, minimum child weight, subsample ratio, and regularization parameters.

The learning rate parameter controls the contribution of each tree to the final prediction, with lower values requiring more trees but potentially achieving better generalization:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta \cdot f_t(x_i)$$

where η represents the learning rate, $f_t(x_i)$ denotes the prediction from the t-th tree, and $\widehat{y_i^{(t)}}$ represents the prediction after t iterations. The hyperparameter optimization process follows a structured two-stage approach that balances computational efficiency with comprehensive parameter space exploration. In the first stage, grid search establishes baseline parameter ranges by evaluating discrete combinations of key hyperparameters: learning rate values of 0.01, 0.05, 0.1, and 0.2; maximum tree depths from 3 to 10; minimum child weights of 1, 3, 5, and 7; and subsample ratios of 0.6, 0.8, and 1.0. This initial grid search evaluates 768 parameter combinations using 5-fold cross-validation, identifying promising parameter regions based on validation set R^2 values. The second stage applies Bayesian optimization within the promising regions identified in stage one, using Gaussian process priors to model the relationship between hyperparameters and model performance. This Bayesian optimization adaptively samples parameter combinations that maximize expected improvement in validation accuracy. The final hyperparameter configuration is selected based on a weighted criterion that considers validation accuracy, training time, and model complexity to ensure practical applicability in real-time design workflows. For the climate zones studied, optimal configurations converge to learning rates of 0.05–0.08, maximum depths of 6–7, minimum child weights of 3–5, and subsample ratios of 0.8–0.9, with specific values varying by approximately 10–15% across different climate zones to account for regional variations in energy consumption patterns. This systematic optimization approach ensures that model performance is not limited by suboptimal hyperparameter selection while maintaining computational tractability for practical implementation.

APPENDIX C: Statistical significance testing procedures and results

F.1 Bootstrap resampling methodology

To establish statistical rigor in performance comparisons, we conduct comprehensive significance testing using bootstrap resampling procedures. The bootstrap methodology provides robust estimates of performance metric variability and enables construction of confidence intervals without assuming specific probability distributions.

The bootstrap procedure operates through five sequential steps. First, from the validation dataset containing n samples, we randomly select n samples with replacement to create a bootstrap sample, where some original samples may appear multiple times while others may not appear at all. Second, we train each comparison model (Random Forest, SVM, ANN) on the bootstrap sample using their respective optimized hyperparameters. Third, we calculate performance metrics (R^2, MAE, RMSE) on the out-of-bootstrap samples, which are samples not selected in the bootstrap sample and constitute approximately 36.8% of the original data. Fourth, we repeat the first three steps for 1,000 iterations to generate sampling distributions of performance metrics. Fifth, we calculate 95% confidence intervals using the percentile method, taking the 2.5th and 97.5th percentiles of the bootstrap distribution.

F.2 Paired t-Test procedures

Paired t-tests assess whether the performance difference between XGBoost and each baseline model is statistically significant. The test operates on prediction errors rather than performance metrics directly. The null hypothesis (H_0) states that the mean prediction error of XGBoost equals the mean prediction error of the baseline model, while the alternative hypothesis (H_1) states that the mean prediction error of XGBoost differs from the mean prediction error of the baseline model.

For each validation sample i, we calculate the XGBoost error as:

$$e_{XGB,i} = |y_{actual,i} - y_{XGB,i}|$$

the baseline model error as:

$$e_{baseline,i} = |y_{actual,i} - y_{baseline,i}|$$

and the difference as:

$$d_i = e_{XGB,i} - e_{baseline,i}$$

The t-statistic is then calculated as:

$$t = \frac{\overline{d}}{s_d/\sqrt{n}}$$

where $\overline{d}$ represents the mean difference across all validation samples, s_d denotes the standard deviation of differences, and n indicates the number of validation samples. The degrees of freedom for the *t*-test equal n-1, and the p-value is calculated using the two-tailed t-distribution. Statistical significance is assessed at the $\alpha = 0.001$ level, corresponding to 99.9% confidence.

F.3 Statistical test results

Table F.1 Statistical significance analysis of model performance.

Model	Heating Load R^2	95% CI	Cooling Load R^2	95% CI	Total Energy R^2	95% CI	p-value vs. XGBoost
XGBoost	0.9763	[0.9721, 0.9798]	0.9466	[0.9401, 0.9524]	0.9374	[0.9305, 0.9437]	–
ANN	0.9519	[0.9463, 0.9569]	0.8592	[0.8486, 0.8692]	0.8796	[0.8697, 0.8889]	$p < 0.001$
Random Forest	0.8588	[0.8497, 0.8674]	0.8616	[0.8511, 0.8715]	0.8553	[0.8449, 0.8651]	$p < 0.001$
Decision Tree	0.9269	[0.9198, 0.9334]	0.7583	[0.7444, 0.7716]	0.7645	[0.7509, 0.7775]	$p < 0.001$

The 95% confidence intervals are computed using bootstrap resampling with 1,000 iterations, providing robust estimates of metric variability. The paired t-tests compare prediction errors (|y_actual − y_predicted|) between XGBoost and each baseline model across all validation samples. All comparisons yield p-values below 0.001, indicating that XGBoost's superior performance is statistically significant at the 99.9% confidence level. The confidence intervals for XGBoost show relatively narrow ranges (e.g., [0.9721, 0.9798] for heating load R^2), demonstrating consistent high performance across different subsets of the validation data.

Data availability

Data will be made available on request.

References

[1] P. Shen, Y. Li, X. Gao, S. Chen, X. Cui, Y. Zhang, X. Zheng, H. Tang, M. Wang, Climate adaptability of building passive strategies to changing future urban climate: a review, Nexus 2 (2) (2025) 1–13.

[2] R. Yang, S.T. Imalka, W.M.P. Wijeratne, G. Amarasinghe, N. Weerasinghe, S.D. S. Jayakumari, H. Zhao, Z. Wang, C. Gunarathna, J. Perrie, C. Liu, R. Wakefield, Digitalizing building integrated photovoltaic (BIPV) conceptual design: a framework and an example platform, Build. Environ. 110675 (2023).

[3] W. Chen, S. Yang, X. Zhang, N.D. Jordan, J. Huang, Embodied energy and carbon emissions of building materials in China, Build. Environ. 207 (2022) 108434.

[4] S. Hu, Y. Zhang, Z. Yang, D. Yan, Y. Jiang, Challenges and opportunities for carbon neutrality in China's building sector—Modelling and data, Build. Simul. 15 (11) (2022) 1899–1921.

[5] H.-T. Thai, T. Ngo, B. Uy, A review on modular construction for high-rise buildings, Structures 28 (2020) 1265–1290.

[6] F. Greer, A. Horvath, Modular construction's capacity to reduce embodied carbon emissions in California's housing sector, Build. Environ. 240 (2023) 110432.

[7] M.R. Abdul Kadir, W.P. Lee, M.S. Jaafar, S.M. Sapuan, A.A.A. Ali, Construction performance comparison between conventional and industrialised building systems in Malaysia, Struct. Surv. 24 (5) (2006) 412–424.

[8] J. Meiling, F. Backlund, H. Johnsson, Managing for continuous improvement in off-site construction, Eng. Constr. Archit. Manag. 19 (2) (2012) 141–158.

[9] L. Ding, Y. Zhu, L. Zheng, Q. Dai, Z. Zhang, What is the path of photovoltaic building (BIPV or BAPV) promotion? –the perspective of evolutionary games, Appl. Energy 340 (2023) 121033.

[10] T. Chen, K.F. Tai, G.P. Raharjo, C.K. Heng, S.W. Leow, A novel design approach to prefabricated BIPV walls for multi-storey buildings, Journal of Building Engineering 63 (2023) 105469.

[11] R. Yang, T. Samarasinghalage, Y. Zhao, Data-driven Building Integrated Photovoltaics (BIPV) envelope design optimization, in: In: 2025 IEEE 53rd Photovoltaic Specialists Conference (PVSC), 2025, pp. 0504–0507.

[12] EnergyPlus, EnergyPlus Engineering Reference – the Reference to EnergyPlus Calculations, in, DOE, 2007.

[13] P. Shen, Z. Wang, Y. Ji, Exploring potential for residential energy saving in New York using developed lightweight prototypical building models based on survey data in the past decades, Sustain. Cities Soc. 66 (2021) 102659.

[14] C. Dara, C. Hachem-Vermette, Evaluation of low-impact modular housing using energy optimization and life cycle analysis, Energy Ecol. Environ. 4 (6) (2019) 286–299.

[15] B.J.O. Pasello, R.M.S.F. Almeida, J.D.M. Moura, What does Modular mean? a Systematic Review on Definitions, Ambiguities, and Terminological Gaps in Construction, Buildings 15 (17) (2025) 3017.

[16] L. Lei, S. Shao, L. Liang, An evolutionary deep learning model based on EWKM, random forest algorithm, SSA and BiLSTM for Building Energy Consumption Prediction, Energy 288 (2024) 129795.

[17] M. Kamali, K. Hewage, Life cycle performance of modular buildings: a critical review, Renew. Sustain. Energy Rev. 62 (2016) 1171–1183.

[18] S. Lau, T. Chen, J. Zhang, X. Xue, S. Lau, Y. Khoo, A new approach for the project process: prefabricated building technology integrated with photovoltaics based on the BIM system, in: IOP conference series: earth and environmental science, IOP Publishing, 2019, pp. 012050.

[19] C. Vassiliades, G. Barone, A. Buonomano, C. Forzano, G.F. Giuzio, A. Palombo, Assessment of an innovative plug and play PV/T system integrated in a prefabricated house unit: active and passive behaviour and life cycle cost analysis, Renew. Energy 186 (2022) 845–863.

[20] Y. Chen, A.K. Athienitis, K. Galal, Modeling, design and thermal performance of a BIPV/T system thermally coupled with a ventilated concrete slab in a low energy solar house: Part 1, BIPV/T System and House Energy Concept, Solar Energy 84 (11) (2010) 1892–1907.

[21] Y. Wang, W. Tian, J. Ren, L. Zhu, Q. Wang, Influence of a building's integrated-photovoltaics on heating and cooling loads, Appl. Energy 83 (9) (2006) 989–1003.

[22] M. Li, T. Ma, J. Liu, H. Li, Y. Xu, W. Gu, L. Shen, Numerical and experimental investigation of precast concrete facade integrated with solar photovoltaic panels, Appl. Energy 253 (2019) 113509.

[23] M.A. Mohammed, I.M. Budaiwi, M.A. Al-Osta, A.A. Abdou, Thermo-Environmental Performance of Modular Building Envelope Panel Technologies: a Focused Review, Buildings 14 (4) (2024) 917.

[24] C. Wang, J. Ji, Comprehensive performance analysis of a rural building integrated PV/T wall in hot summer and cold winter region, Energy 282 (2023) 128302.

[25] T. Hwang, J.T. Kim, Y. Chung, Power performance of photovoltaic-integrated lightshelf systems, Indoor Built Environ. 23 (1) (2014) 180–188.

[26] H. Lee, Performance evaluation of a light shelf with a solar module based on the solar module attachment area, Build. Environ. 159 (2019) 106161.

[27] Y. Sun, D. Liu, J.-F. Flor, K. Shank, H. Baig, R. Wilson, H. Liu, S. Sundaram, T. K. Mallick, Y. Wu, Analysis of the daylight performance of window integrated photovoltaics systems, Renew. Energy 145 (2020) 153–163.

[28] P. Redweik, C. Catita, M. Brito, Solar energy potential on roofs and facades in an urban landscape, Sol. Energy 97 (2013) 332–341.

[29] K. Fath, J. Stengel, W. Sprenger, H.R. Wilson, F. Schultmann, T.E. Kuhn, A method for predicting the economic potential of (building-integrated) photovoltaics in urban areas based on hourly Radiance simulations, Sol. Energy 116 (2015) 357–370.

[30] C. Feng, C. Zhang, J. Lu, Y. Zhao, Hybrid data-driven and physics-based fast building cooling demand modeling method for large-scale building demand response control, Journal of Building Engineering 100 (2025) 111808.

[31] K. Amasyali, N.M. El-Gohary, A review of data-driven building energy consumption prediction studies, Renew. Sustain. Energy Rev. 81 (2018) 1192–1205.

[32] Y. Chen, M. Guo, Z. Chen, Z. Chen, Y. Ji, Physical energy and data-driven models in building energy prediction: a review, Energy Rep. 8 (2022) 2656–2671.

[33] Q. Qiao, A. Yunusa-Kaltungo, R.E. Edwards, Towards developing a systematic knowledge trend for building energy consumption prediction, Journal of Building Engineering 35 (2021) 101967.

[34] X. Li, S. Chen, H. Li, Y. Lou, J. Li, A behavior-orientated prediction method for short-term energy consumption of air-conditioning systems in buildings blocks, Energy 263 (2023) 125940.

[35] M. El-Maraghy, M. Metawie, M. Safaan, A. Saad Eldin, A. Hamdy, M. El Sharkawy, A. Abdelaty, S. Azab, M. Marzouk, Predicting energy consumption of mosque buildings during the operation stage using deep learning approach, Energ. Buildings 303 (2024) 113829.

[36] P. Shen, Building retrofit optimization considering future climate and decision-making under various mindsets, Journal of Building Engineering 96 (2024) 110422.

[37] S. Zhan, G. Wichern, C. Laughman, A. Chong, A. Chakrabarty, Calibrating building simulation models using multi-source datasets and meta-learned Bayesian optimization, Energ. Buildings 270 (2022) 112278.

[38] N. Bucarelli, N. El-Gohary, Sensor deployment configurations for building energy consumption prediction, Energ. Buildings 308 (2024) 113888.

[39] C. Song, H. Yang, X.-B. Meng, P. Yang, J. Cai, H. Bao, K. Xu, A novel deep-learning framework for short-term prediction of cooling load in public buildings, J. Clean. Prod. 434 (2024) 139796.

[40] L. Barreñada, P. Dhiman, D. Timmerman, A.-L. Boulesteix, B. Van Calster, Understanding overfitting in random forest for probability estimation: a visualization and simulation study, Diagn. Progn. Res. 8 (1) (2024) 14.

[41] Y.-C. Wu, J.-W. Feng, Development and Application of Artificial Neural Network, Wirel. Pers. Commun. 102 (2) (2018) 1645–1656.

[42] C. Fan, J. Wang, W. Gang, S. Li, Assessment of deep recurrent neural network-based strategies for short-term building energy predictions, Appl. Energy 236 (2019) 700–710.

[43] T. Chen, C. Guestrin, Xgboost: a scalable tree boosting system, in, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[44] S. Li, M. Wang, P. Shen, X. Cui, L. Bu, R. Wei, L. Zhang, C. Wu, Energy Saving and thermal Comfort Performance of Passive Retrofitting measures for Traditional Rammed Earth House in Lingnan, China, Buildings 12 (10) (2022) 1716.

[45] C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Climate Res. 30 (1) (2005) 79–82.