Examining the generalizability of inverse surrogate models for different geometries and locations

Liam Jowett-Lockwood, Ralph Evins

2025

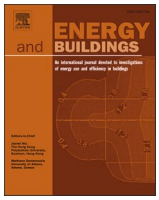
Original citation:
Jowett-Lockwood, L. Evins, R. (2025) Examining the generalizability of inverse surrogate models for different geometries and locations. *Energy and Buildings* (335) https://doi.org/10.1016/j.enbuild.2025.115539

Energy & Buildings 335 (2025) 115539

Contents lists available at ScienceDirect

Energy & Buildings

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



Examining the generalizability of inverse surrogate models for different geometries and locations

Liam Jowett-Lockwood [a,b,c], Ralph Evins [a,b,*]

[a] Energy in Cities Group, Department of Civil Engineering, University of Victoria, V8P 5C2, Canada
[b] Institute for Integrated Energy Systems, University of Victoria, V8P 5C2, Canada
[c] RJC Engineers, V8W 2G4, Canada

ARTICLE INFO

Keywords:
Surrogate modelling
Building retrofit
Inverse modelling

ABSTRACT

While building surrogate modelling has been shown to accurately replicate the outputs of computationally intensive building energy modelling, successfully adopting surrogate modelling in practice still has challenges. As surrogate models are machine learning models, they require an extensive quantity of training data in order to train effectively. The process of acquiring training data often requires numerous simulation runs of a building energy model. To offset this issue, surrogate models that demonstrate a suitable level of generalizability can be applied successfully to multiple projects without the need for the further generation of data.

This study examines the generalizability of multiple inverse surrogate models. Inverse surrogate modelling is a more difficult task than traditional surrogate modelling as it tries to extract building energy model inputs from output data. As the output data required to do this is often comprehensive, deep learning models are preferred. For the inverse surrogate models, a basic deep artificial neural network, convolutional neural network, recurrent neural network and transformer were examined. Output data in this study consisted primarily of temperature and energy time series data with input data being building energy model parameters reflective of thermally important building characteristics.

Generalizability is assessed by first training the inverse surrogate models on data from 3 separate building energy models. Each of the building energy models contain geometry that is randomly scaled. Additionally we examine training the inverse surrogate models on building energy model data produced with multiple locations as well as on data from all building energy models at once. Parameters relating to the building envelope demonstrated the highest prediction performance among the models, whereas the prediction performance for less influential parameters was more varied depending on the inverse surrogate model. Overall, the convolutional neural network typically outperformed the other models with the recurrent neural network and transformer producing slightly worse performance. The artificial neural network was unable to accurately predict parameters outside of a select few that were highly influential to the time-series data. In the cases of training with data from multiple locations or all buildings at once, prediction performance decreased, however several parameters remained predictable.

1. Introduction

1.1. Background

Over the last few decades, Building Energy Modelling (BEM) techniques and methods have been rigorously applied to assist with early-stage building design, optimization, energy savings etc. [1]. With the Canadian Government funding new construction to address the needs of Canada's housing demand [2], the need for effective building performance simulation continues to grow. This is further exacerbated by the impact buildings have on Green House Gas (GHG) emissions. Globally, the continued growth of construction activities have led to increasing not only the GHG emissions from the act of constructing the buildings themselves, but have also contributing to the all-time highs of CO_2 emissions from building operational energy [3]. BEM software has been continuously developed throughout the last several years to further assist practitioners. Common BEM software programs include EnergyPlus [4], IES-VE [5], and DesignBuilder [6]. These computationally

* Corresponding author at: Energy in Cities Group, Department of Civil Engineering, University of Victoria, V8P 5C2, Canada.
E-mail address: revins@uvic.ca (R. Evins).

https://doi.org/10.1016/j.enbuild.2025.115539
Received 5 November 2024; Received in revised form 13 February 2025; Accepted 27 February 2025
Available online 6 March 2025

intensive software tools apply a complex set of inputs to develop a computerized model of the building from which a series of physics equations can be activated to calculate consumptions and emissions overtime.

BEM can be applied to various stages throughout the development of a building, including the preliminary design stage, developed design and post construction [7,8] In the preliminary design stage process, building designers can rely on BEM to provide necessary insight into preliminary energy performance. While a comprehensive building energy simulation model can be considered too uncertain, due to the substantial number of known inputs required, additional BEM software has been introduced recently to aid decision making at this design stage [9]. Later in the design stage when more information of the building's design has been decided, a comprehensive energy model is more appropriate and practitioners can review the potential energy performance and determine if it is within desired limits. Once the post construction phase has been achieved, the use of digital twins becomes more apparent.

In each of these stages, the computationally intensive nature of BEM can be problematic. Running BEM simulations are often met with serious runtimes for many models, thereby hindering tasks which may necessitate multiple simulations. While simpler models alternatively used in the early design stage are more immune to high computational complexity problems, these models are more simplistic and lack some of the technical rigour that could be eventually desired. Furthermore, even at the design stage, some desirable input combinations (e.g., varied geometry combinations and different mechanical systems), may not be possible within the limitations of simpler models, thereby potentially weakening their usability. The issues of computationally intensive BEM can extend to later design stages as well, where finding suitable energy efficiency against increased costs can often require multiple simulation runs with varying inputs.

The performance of buildings after completion often do not align with the original models [8]. The common solution are calibrated building energy models adjusted such that their outputs are similar enough to those observed in reality. Creating a calibrated energy model is an iterative process for which input values are varied either manually or automatically until suitable outputs are acquired [10]. This iterative process involves the repeated use of simulations, for which the process can be beset again by long computational runtimes.

1.2. Surrogate modelling

As a response to the hinderances that BEM currently face, an alternate method has emerged whereby a traditional BEM model is replaced with a building Surrogate Model (SM). A building SM is a Machine Learning (ML) model designed to replicate the performance of a computationally intensive BEM model while addressing the issue of long computational runtimes [11]. Instead of completing complex physics equations reminiscent of BEM, a trained building SM functions by placing accurate predictions on BEM output values almost instantaneously when provided relevant BEM inputs. Building SMs have demonstrated high accuracy in matching outputs from comprehensive BEM models [11]. Common examples include Artificial Neural Networks (ANNs) and Support Vector Machines. The creation of a SM can be organized into the following steps [12]:

1) Acquisition of training data.
2) Processing of training data
3) SM training
4) SM validation

ML models require a high volume of samples each composed of input and output data in order to train efficiently. For a building SM, input data would consist of numerical building properties referred to as parameters which can include the conductivity value of the wall insulation or the flow rate of infiltrating air. Output data would be calculated BEM outputs, such as energy consumption or internal temperatures. Each sample is composed through their own BEM simulation for which the input parameter values are applied to the actual BEM model (that the SM is trying to replicate) to obtain the corresponding output data. Acquiring enough training data in this fashion is a lengthy process, as it may consist of thousands of actual BEM simulation runs. While SM predictions are made near instantly, the process of acquiring the training data could be problematic. However, to counteract this, a SM can be considered generalizable if it can be applied to multiple projects or BEM models, thereby reducing the need to always create and train a new SM.

Once a sufficient quantity of training data has been obtained, it becomes necessary to perform data processing. Initially the integrity of the data should be assessed for missing data points or erroneous values. Fortunately, as the training data is generated via simulation, complete data integrity can usually be easily assured. Preprocessing in the form of data scaling should then be completed to prepare the data for training in addition to defining how to split the data into training, validation and test sets.

After training data has been obtained and appropriately processed, model construction and hyperparameter selection must be conducted prior to training. Model construction relates to the determination of which ML model, its components (e.g., layers) and the structuring of them. Hyperparameters are highly influential variables that relate to either the ML model's components or training ability. The ML model cannot learn its own hyperparameter values and it is ideal, though time consuming, to perform hyperparameter tuning by experimenting with different values and model configurations. It is noteworthy that an appropriate model construction and hyperparameter determination will not only influence prediction accuracy but also memory usage and training time among others.

After training and the generation of predictions, the SM is validated for its accuracy. This is often done through the use of statistical error metrics and common examples include the Coefficient of Determination (R^2), Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE).

1.3. Surrogate modelling with deep learning models

With advancements in the ML domain, the rise of Deep Learning (DL) models and methods have provided recent SMs increased potential for successfully learning difficult tasks. DL includes a subset of ANN models that utilize multiple layers to continuously extract information. This is usually coupled with a more complex set of neurons allowing for advanced operations that provide higher performance within domains that contain large high-dimensional data, such as multiple time series, high resolution images, and complex text [13] Compared to shallow learning methods, DL methods have had less traction for BEM related tasks, however this has improved in recent years partly because of increasingly available powerful computing hardware [11,14]. Common examples include multilayered ANNs, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Sections 1.3.1–1.3.3 provide a brief description of their corresponding DL model as well as related studies. Section 1.3.4 describes the added complexity of the method of inverse surrogate modeling and related studies.

1.3.1. Basic deep artificial neural networks

As suggested in their name, ANNs attempt to replicate the decision process of neurons within the central nervous system of biological animals [15]. Reminiscent of synapses in a brain, connections of artificial neurons within an ANN produce signals or values that are adjusted by a weight that is configured during the learning process. These neurons or nodes are collected in layers. Whereas early versions of ANNs utilized only one layer (shallow models), formulation of the back propagation algorithm enabled effective training of multi-layered ANNs [15].

With the rise in computing power and shear availability of data over the last several decades, DL ANNs have shown substantial success over

other ML methods when tasked with large complex problems [16]. The main component of a basic ANN is the Fully Connected or Dense Layer. As shown in Fig. 1, each node in the Dense layer forms a weighted connected to the each of the nodes in the previous layer, hence being "fully connected".

Olu-Ajayi, et al. examined the ability of a variety of ML models, including both deep and shallow ANNs, to predict annual energy consumption for different types of residential buildings [17]. Their study concluded that the deep learning ANN produced the best results overall, however, some other models were comparatively accurate and the building type was negligible on performance. Suryanarayana et al. compared DL to conventional ML methods for thermal load forecasting of district heating networks [18]. Their DL model consisted of a simple ANN with 2 hidden layers, while other models included a polynomial linear regression model and a ridge regression model. Their study examined two case studies involving different district heating networks in Sweden and in both studies, their DL model outperformed its simpler counterparts with a Mean Absolute Percentage Error (MAPE) of 8.08 % and 4.15 % respectively.

1.3.2. *Recurrent neural networks*

RNNs are a subset of ANNs that are specialized for learning information from sequential data. The significance of the model is that recurrent connections are placed between nodes in the layer such that information from the previous time step is sent forward, which enables the nodes at previous time steps to directly influence the output of future time steps [19]. This allows the model to learn sequential data more strongly compared to an ANN, as there is an immediate connection between each sequential value. This process is illustrated in Fig. 2.

The major hinderance traditional RNNs face is the issue of the vanishing (or sometimes exploding) gradient. In cases of the vanishing gradient, long-term dependences become very difficult or time consuming for the model to learn as exponentially smaller weights are provided to long-term interactions [19]. As a response, different variants of RNNs, namely Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) models have seen prominence over traditional RNNs. LSTM models make use of a memory cell that modifies the RNN architecture by enabling gating signals to help handle challenges that can arise during model training [20].

Fan et al. examined multiple RNNs and input data strategies for short-term building energy predictions [21]. Their models included regular and bidirectional GRU and LSTM models as well as the possible inclusion of a 1D-convolutional layer beforehand. They note that a direct approach, where separate models are created for each time step, performed the best without significant computational burden and both the GRU and LSTM models provided better preservation of long-term temporal dependencies. They also noted that bidirectional operations were beneficial for improving prediction accuracies for cooling loads.

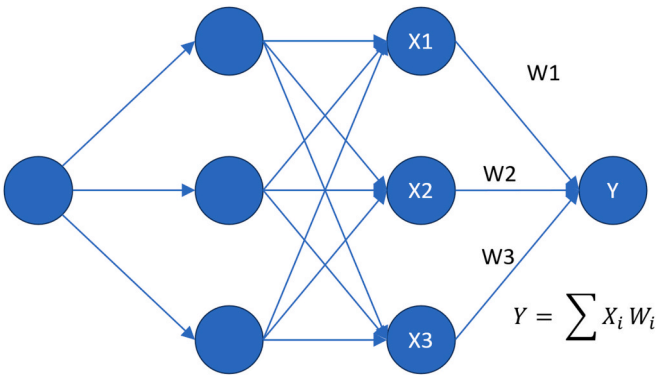


Fig. 1. Simple ANN example. This example features two dense layers in the middle (hidden layers).

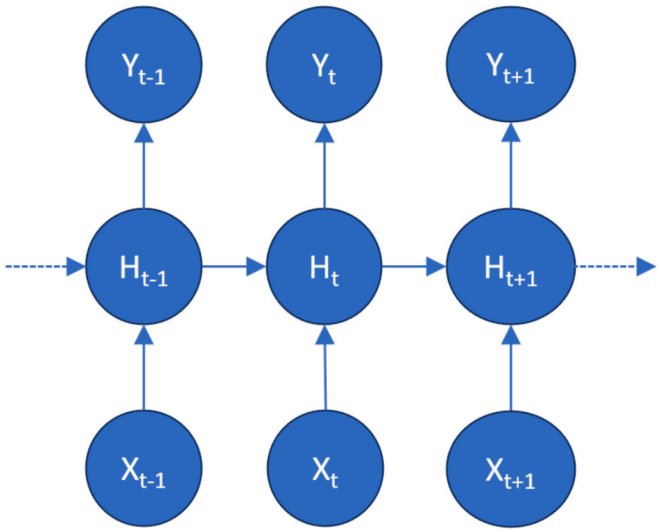


Fig. 2. RNN configuration. Inputs (X) are fed into the hidden nodes (H) which influence future outputs (Y).

Jung et al. developed a multilayer attention-based GRU for the purposes of short-term load forecasting [22]. Making the GRU attention-based was chosen to help enable the model to understand which parts of the input sequence are the most important for load forecasting. When tested for three separate buildings they noted that their model outperformed others such as conventional LSTM and GRU models. Jang et al. employed three different LSTM models for the purposes of predicting heating energy consumption [23]. The first model had input data consisting of building environmental data, the second model also had building environmental data as well as outdoor environmental data and the third model had both sets of environmental data along with operation patten data. They found that the third model outperformed the other two while having the additional strength of better handling situations where the energy consumption experienced a sudden change.

1.3.3. *Convolutional neural networks*

Whereas RNNs are specialized to handle sequential data, CNNs are a particular type of ANN for the purpose of efficiently processing data with a grid-like topology [19]. Data in this form can include images or multiple concurrent time series. CNNs rely on a mathematical process known as convolution, which involves multiplying the input matrix by another, usually smaller, matrix referred to as the kernel matrix. The process is completed by multiplying the kernel matrix along portions of the input matrix (Fig. 3). The unique property of the kernel matrix is that because it moves along the input, its weights are effectively shared. For this reason, CNNs are typically much more memory efficient than similar deep ANNs.

Somu et al. incorporated k-means clustering into a hybrid CNN-LSTM model with the intention of building energy consumption forecasting [24]. The K-means clustering was performed to help organize the energy consumption trends, the CNN to extract features from non-linear interactions impacting energy consumption and the LSTM component to help control the long-term dependencies within their time-series data. When applied to a case study involving collected building energy consumption time series data, their hybrid model outperformed more conventional CNN and LSTM models. Westermann et al. developed a CNN based SM with the intention of predicting heating or cooling energy demand regardless of location [25]. The CNN processes weather-based time-series data before receiving additional inputs in the form of building parameters. When tested on unseen locations, the model experiences a MAPE of less than 3 %.

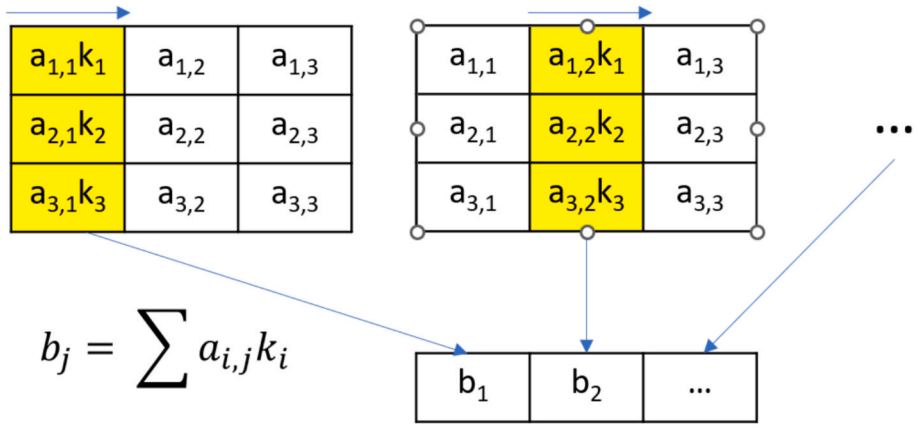


Fig. 3. Dimensional Convolutional Process. The kernel matrix is multiplied by sections of the input matrix to form the output. The kernel moves along during the convolutional process as indicated by the highlighted cells.

1.3.4. Inverse surrogate modelling

As deep learning methods exhibit their largest potential with high-dimensional data, this gives prominence to the potential of successfully applying DL to inverse modelling. For BEM, inverse modelling is the notion of training a SM on BEM outputs to find inputs instead of the usual other way around. As the inverse task is being applied, a SM that trains and predicts this way can be referred to as an Inverse Surrogate Model (ISM). This is often a much more difficult task as it is not entirely possible to map exact input values to relevant physics equations based on only output data as a different combination of values could result in the same output. Fortunately, many BEM outputs can be expressed as a time series, or as outputs of a similar scale, and with a DL model's ability to extrapolate features by inferring the time series, obtaining individual values has a heighted possibility.

While inverse surrogate modelling is less common than the forward approached, examples in literature have examined the potential of ISMs for predicting building parameters. Ferreira et al, examined the applicability of an ANN based ISM used to predict various BEM input parameters related to heating and cooling loads [26]. Their study involved formulating the data with change point models to serve as input data to the ISM and performance was assessed with a testing dataset of 3000 samples. Herbinger et al. developed a SM in the form of an ANN to calibrate a building energy model [27]. Their method is unique in that, once trained, the model uses itself to calibrate building parameter values via gradient descent, resembling an ISM. They compared their SM to a powerful ML optimizer in a controlled case study and found that their SM surpassed the ML in performance. They also compared the performance of the models in a real metered data case study and while it was more comparable, the SM was more consistent.

1.4. Areas of improvement

While DL models provide practitioners increased opportunities for SM usage, they still face the same fault of requiring numerous amounts of training data. With growing popularity, it remains important that SMs express a suitable level of generalizability so that they can be readily applied to multiple projects without requiring the lengthy process of generating new training data each time. Generalizability relates to a ML model's ability to place accurate predictions on unseen data [19]. In the context of this study, we are focused on the generalizability that relates to multiple buildings (i.e. unseen geometry).

The intention of this study is to explore the generalizability of multiple DL models for building SM applications by examining the robustness of each model when trained on variable geometric BEM models. As inverse modelling is typically a more difficult task than forward modelling, this study assesses the performance of the DL models as ISMs to try to understand how DL models can benefit this task. The BEM models used to create training data in this study are partly inspired by commercial office buildings and their basic geometry are randomly scaled to allow for a scenario where a SM needs to be used for another project. Additionally examined is the impact to prediction performance when trained on data from a BEM model simulated in multiple locations or being trained on data from all BEM models at once.

This paper compares an ANN, RNN, CNN and a transformer architecture. To the authors knowledge, this is the first study to examine the potential of multiple deep learning models within the context of assessing generalizability as inverse surrogate models. Both RNNs and CNNs were selected as they are well suited for an ISM trained on time series data, while a basic deep ANN helps provide a performance benchmark. In recent years, transformers have gained increased attention in machine learning communities for their ability to surpass RNNs and other ANN models for sequential text related problems [28]. Unlike RNNs, the transformer is able to model the relationship between all inputs in a sequence regardless of placement by applying a self-attention mechanism. This enables the model to understand the influence of elements more easily in the sequence on other potentially far away elements when compared to a RNN or CNN.

2. Methodology

The methodology of this study is broken into four sections: 3.1 training data organization, 3.2 geometric model design and development, 3.3 ISM model construction and 3.4 error metrics for evaluation. Fig. 4 provides an overview of the methodology used.

Organizing training data relates to deciding on appropriate BEM model parameters and BEM output data to collect along with adjustments and necessary preprocessing. Geometric model construction examines the geometry of the BEM models themselves and other various aspects of them including scalability. ISM model construction relates to the determination of ISM layers and various hyper parameters. Model evaluation describes the various error metric equations used to evaluate the models in the results section of this study.

2.1. Training data organization

When training data is selected, it is imperative that there exists a strong connection between the inputs and the outputs. For the ISMs, this requires that the building parameters selected make a noticeable impact on the BEM outputs otherwise the ISM will struggle to learn the corresponding parameter values. This study prominently uses internal temperature time series data as an ISM input, as they are significantly affected by building envelope properties. Section 3.1.1 describes the parameter selection as they are the most influential on the chosen time series, which is then followed by describing the time series selected as

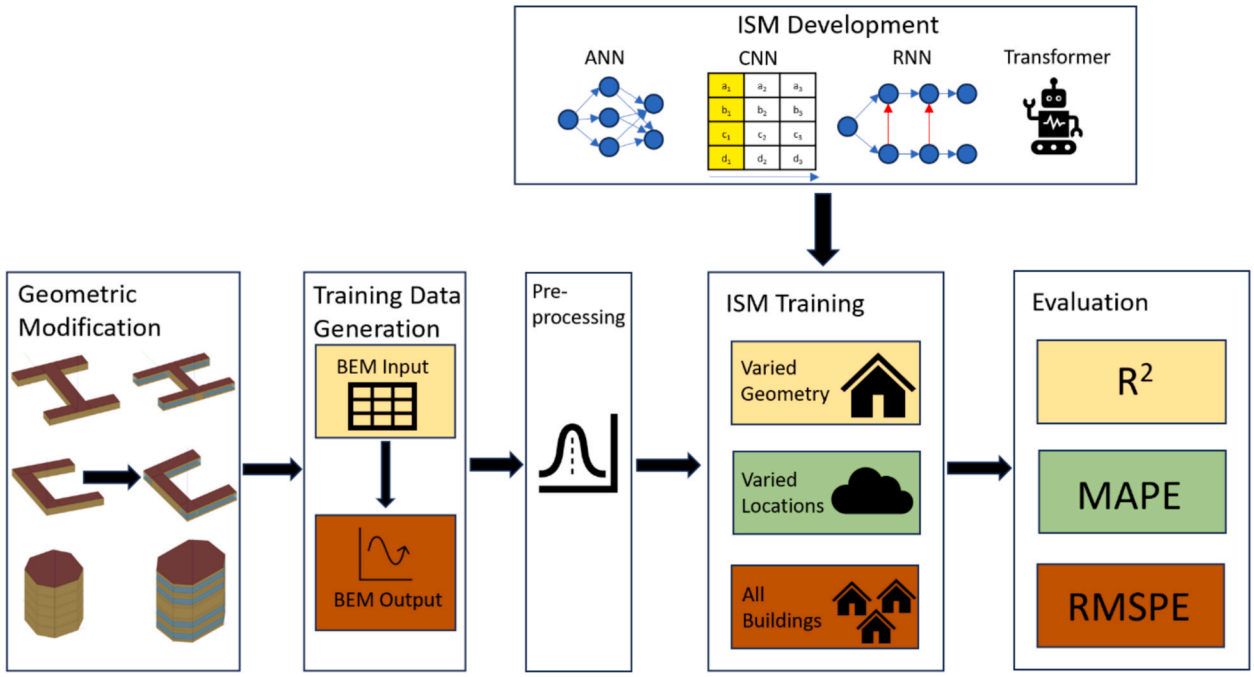


Fig. 4. Study methodology. Three different scalable core office shapes are used to produced training data. Prediction performance is examined with 4 different ISMs: ANN, CNN, RNN and a transformer. Prediction performance with each of the 3 core shapes is assessed as well as training for multiple locations and all shapes at once.

well as their characteristics in Section 3.1.2 and then how preprocessing is performed in Section 3.1.3. Lastly, we discuss the approach of acquiring training data when testing with multiple locations at once (Section 3.1.4).

2.1.1. *BEM parameter selection*

The varied parameters used in this study are provided in Table 1 along with their ranges. The first eight parameters are typical BEM inputs that all have a varying impact on the internal temperatures of the BEM model. Predicting these parameters represents the primary task of the ISMs. The remaining four parameters are implemented with the intention of assisting with the development of suitable training data. Their values are additionally predicted by the ISMs, however, given their significant relevance to internal temperatures, the prediction performance on them is anticipated to be higher than the others.

The ranges in Table 1 are based partly on those in [25] and selected to range from highly to moderately influential for internal temperatures. The conductivity parameters influence the overall effectiveness of the building envelope and their individual impact is additionally contributed by the shape modification of the structure. Increases in height lead to increases in wall and window surface area while the roof surface area remains constant, thereby causing an increase in the contribution of overall building envelope heat flow as a result of wall insulation and window conductivity. It should additionally be noted that the weight category of each sample remains consistent.

Both the Lighting Energy Power Density and the Equipment Energy Power Density parameters influence the internal temperatures via internal gains from lighting and equipment respectively and as a result they are expected to have less impact than the building envelope parameters. Predicting these parameters will examine the ability to accurately discern smaller influences. For all BEM models, the setpoint schedule for the lighting was turned on fully between 5:00 AM to 8:00 PM and completely turned off otherwise. To differentiate it, the equipment energy setpoint schedule was turned on at all times. Differences in prediction performance between the two parameters would suggest that the ISMs would either be impacted or not with a varied setpoint schedule.

While the People Quantity parameter is similar to the lighting and equipment parameters, it is unique in that it directly influences the Ventilation Flow Rate. The Ventilation Flow Rate parameter represents the rate of the intended flow of air into the zone and is increased depending on the number of personnel. The Infiltration Flow Rate parameter is instead the unintended rate of airflow into a zone and is therefore unaffected by the number of personnel and is highly impactful on internal air temperatures.

Instances where the internal temperature changes are those which are the most valuable for the ISMs, which typically occur when the heating setpoint changes. When the temperature decreases rapidly, it can suggest that the combination of building envelope parameter values are poor for an energy efficient structure, while slow temperature changes during the same weather period would indicate a more energy efficient structure. The heating setpoint schedule begins at midnight with 16 °C until 6:00 AM where it is increased to 20 °C and then further increased to 23 °C at 7:00 AM. At 9:00 PM the heating setpoint reverts back to 16 °C.

The Ideal Air Loads system component in EnergyPlus is used to represent the heating supply. This component functions as a highly efficient Heating, Ventilation and Air Condition (HVAC) unit and can be

Table 1
BEM Parameters and Ranges. Geometric scale parameters are set as a percentage to retain consistency.

Parameter	Ranges	Units
Wall Insulation Conductivity	0.01–0.1	W/mK
Roof Insulation Conductivity	0.01–0.1	W/mK
Glass Conductivity	0.005–0.03	W/mK
Lighting Energy Power Density	8–12	W/m^2
Equipment Energy Power Density	8–12	W/m^2
People Quantity	0.025–0.05	$People/m^2$
Ventilation Flow Rate	5–10	L/sPerson
Infiltration Flow Rate	0.1–1	L/sm^2
Maximum Heating Air Flow Rate	25–250	L/s
Geometric Scale X Direction	0.01–1	
Geometric Scale Y Direction	0.01–1	
Geometric Scale Z Direction	0.01–1	

used when developing a full HVAC model is not necessary [29]. The Maximum Heating Air Flow Rate parameter limits the ability of the component, which prevents the BEM model from reacting too rapidly to setpoint changes, therefore increasing ISM trainability.

As the different BEM models were modified by different absolute values, the geometric parameters were organized as a factor to keep consistency among them. Scaling by a value of 1 would apply the maximum increase, whereas 0 would apply the maximum decrease.

2.1.2. Time series composition

As the temperature within a BEM model fluctuates over time as a result of thermal setpoint changes and outdoor temperature, the parameters in Table 1 provide a strong influence on the temperature within the model. In this study, the time series selected were interior temperature for each zone (5 total), outdoor air drybulb temperature and the heating energy provided (Table 2).

Instances of temperature decay when the heating setpoint is lowered are more important to the learning process than instances of temperature rise as during decay, the rate of temperature change is not influenced by heat being provided by the HVAC system. In this study, instances of heat loss are referred to as decay curves while instances of temperature rise (usually between a lower heating setpoint and a higher heating setpoint) are referred to as rising curves. It is important to note that the Maximum Heating Flow Rate parameter remains useful even in situations of decay as heating would be applied once the new setpoint is acquired to prevent further loss of heat. Rising curves remain beneficial, however more unknown influences will occur. For example, while solar radiation may only play a small role during decay curves, as they will often happen at night, their presence will frequently be noticeable during rising curves which will more likely be present during daylight hours.

Training with each BEM model individually, as well as all together, was attempted with only decay curves. Training with a combination of decay and rising curves was done when training the ISMs on a BEM model for multiple locations. The intention was that the additional time series data would assist the ISMs in understanding location and help accurately predict parameter values. When training with a combination of decay and rising curves, the additional time series listed in Table 2 were also used. Some of these time series are influential to the internal temperatures (such as Direct and Diffuse Solar Radiation).

Decay curve instances were computed at 10-minute intervals with 12 at a time (2 h) so that the temperature decay could be adequately captured. A requirement that a difference of 0.05 °C between each temperature value for the first 6 values was implemented so that the curves would not be too flat. This requirement was only examined on one selected zone in the BEM models with the assumption that the other zone temperature decay would be similar. As each sample contained a year of data, starting from the beginning of the year, a total of 84 suitable decay curves (1008 values for each time series) were extracted from each sample to resemble one week of training data.

When dealing with both decay and rising curves, the process was similar. Rising curves were also computed with the same minute long intervals and length and the same requirement of a 0.05 °C buffer between the first 6 samples was applied. A combined total of 84 curves was

Table 2
Time-Series ISM Inputs. Cells highlighted in green represent parameters applied to ISMs in all scenarios. Those highlighted in blue are only applied to specific parts of the study.

Time Series	
Zone 1 Internal Air Temperature	Solar Azimuth Angle
Zone 2 Internal Air Temperature	Solar Altitude Angle
Zone 3 Internal Air Temperature	Zone 1 Air System Sensible Heating Energy
Zone 4 Internal Air Temperature	Zone 2 Air System Sensible Heating Energy
Zone 5 Internal Air Temperature	Zone 3 Air System Sensible Heating Energy
Outdoor Air Drybulb Temperature	Zone 4 Air System Sensible Heating Energy
Heating Energy	Zone 5 Air System Sensible Heating Energy
Outdoor Air Wetbulb Temperature	District Cooling Energy
Diffuse Solar Radiation	Outdoor Air Relative Humidity
Direct Solar Radiation	Precipitation Depth

again used, with rising curves being interwoven with the decay curve data.

2.1.3 Data preprocessing

For each BEM model in this study, a total of 5000 samples were used to train the ISMs. An additional 5000 samples were used to train the ISMs in the case of varied location. With each set of 5000 samples, 1000 samples were reserved for the test set and 800 samples for the validation set, leaving 3200 samples for the remainder of the training set. When training on data from each BEM at once, a total of 6000 samples (2000 from each) was used.

Prior to training any of the models, preprocessing was applied to the data in the form of normalization. Preprocessing data such that it resembles standard normally distributed data is commonly performed on ML models as poor training ability can result otherwise [30]. Given that the data in this study consists of various timeseries, the mean and standard deviation was computed across all samples in the training set for each individual time series. These values were then used to normalize both the training and test set with the test set using the training set mean values and standard deviations. Preprocessing was only conducted for training purposes. When model predictions were acquired, values were then unnormalized prior to error metric calculations.

2.1.4 Varied locations

Regarding locations, the city of Victoria, Canada was the only city used to create training data with each BEM model. When investigating ISM performance when trained on data from a single BEM model in multiple locations, only North American cities were examined, which offered a suitable selection of varying climates. Varying location was simply completed by using a different weather file for each location. While determining location is a categorical problem instead of a regression problem, to keep consistency, the ML models predicted latitude and longitude values. The cities chosen are provided in Table 3.

2.2 Geometric model design and development

Each of the three BEM models were based on commercial office building configurations (Fig. 5). As only the overall dimensions of the BEM models were modified in the model (length, width and height), we separately created different core shapes so that it could be understood whether an ISM is universally strong as a parameter predictor for multiple shapes or only specific ones. Each BEM model was composed of a total of five thermal zones, between which conduction occurs during simulations and temperature time series data is collected. Windows were included with a Window to Wall Ratio (WWR) of 50 %. The WWR was held constant regardless of the change in wall size (i.e., windows would scale along with the wall). Windows were omitted from one zone (Z3) in each model to help the model differentiate the impact of them in regards to other parameter influences.

Table 3
Locations used for varied weather data and their climate zone [31].

Location	ASHRAE Climate Zone
Victoria, BC, CAN	4C
Edmonton, AB, CAN	7
Winnipeg, MB, CAN	7
Toronto, ON, CAN	5A
Anchorage, AK, USA	7
Los Angeles, CA, USA	3C
Denver, CO, USA	5B
Miami, FL, USA	1A
Las Vegas, NV, USA	2B
Austin, TX, USA	2A

The H-Shape model serves as the benchmark for comparisons, as each of its zones are scaled accordingly. The C-Shape model retains large similarity to the H-Shape, however scaling is reduced as Z2 and Z4 zones only shift along to accommodate scaling of the other zones and Z1, Z3 and Z5 only increase or decrease in their longer dimensions. The Octagonal-Shape model, differs from the others as it is the only model with multiple stories and nonrectangular zones. When the height is scaled, the height of each zone is modified, thereby significantly affecting the overall height of the structure. As the models are scaled differently, it may prove insightful in whether the degree of scaling affects predictions.

2.3 ISM model construction

While the ISMs have fundamental differences between each other, a learning rate of 0.001, a batch size of 64 and a choice of 150 epochs was used during training for each model. Furthermore, the Adam optimization and loss function of Mean Square Error (MSE) were implemented for each ISM. Outputs for each ISM and scenario are those listed in Table 1. Each ISM has the same inputs from Table 2, which the exception of examining performance with decay and rising curves as well as the varied locations scenario for which additional inputs are provided.

Each model features some dense layers and dropout layers. Dropout layers are a computationally inexpensive method used to prevent overfitting by temporarily eliminating nodes during training [19]. Overfitting occurs when the ML model picks up noise that only exists in the training set during the training process, which leads to overpromising results. Along with the dropout layers, overfitting was prevented by Early Stopping, that would stop training prematurely if it was observed that the validation loss was no longer decreasing.

Hyperparameter values were chosen based on a similar study [25] as well as past experience. It is possible that different ISMs may benefit from different hypermeter considerations, which would incentivize hyperparameter tuning. Given the number and complexity of the ISMs used, hyperparameter tuning and varying of ISM model constructions would be complicated and is left for future work.

2.3.1 Artificial neural network

As the ANN serves as the benchmark for comparisons between other ISMs, it was decided to keep the ANN simple. The combination of Dense and Dropout layers that forms the model is illustrated in Fig. 6. As the Dense layers are ignorant of the sequential nature of the data, inputs were simply formatted into a single row and fed into the model 1-dimensionally. The number of nodes contained in each hidden layer were twice the length (1008) with the exception of the last Dense layer before the output which was a quarter of this (252).

Aside from the final Dense layer, each hidden Dense layer uses the Rectified Linear Unit (ReLU) activation function. This remains the same for the other ISMs. An activation function is commonly applied to layer outputs to provide non-linearity. The Rectified Linear Unit (ReLU) is one of the most common activation functions for ANNs whereby values less than 0 are instead replaced with 0 and values greater than 0 remain unaffected as shown in Eq. (1) [19].

$$y = MAX(0, X) \tag{1}$$

2.3.2 Recurrent neural network

Compared to the ANN, the RNN in this study employes three LSTM layers early in the model to help with the sequential inputs. As shown in Fig. 7, interwoven between the LSTM layers are dropout layers followed by a series of dense and dropout layers. 2-dimensionality is retained throughout the model until the global average pooling layer near the end of the model converts the data to 1-dimension. After another dense layer, outputs are produced by the model.

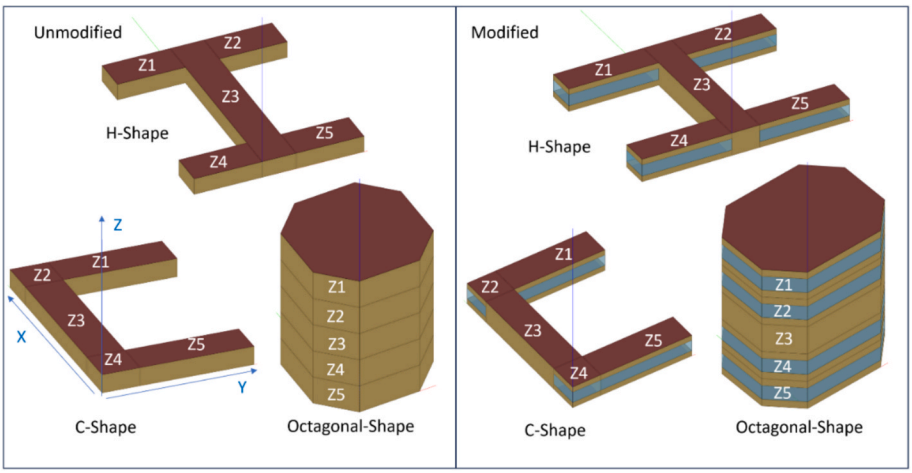


Fig. 5. BEM geometries. Zones are labeled Z1 through Z5. Windows were omitted from Z3 in each model to provide variation between the zones.

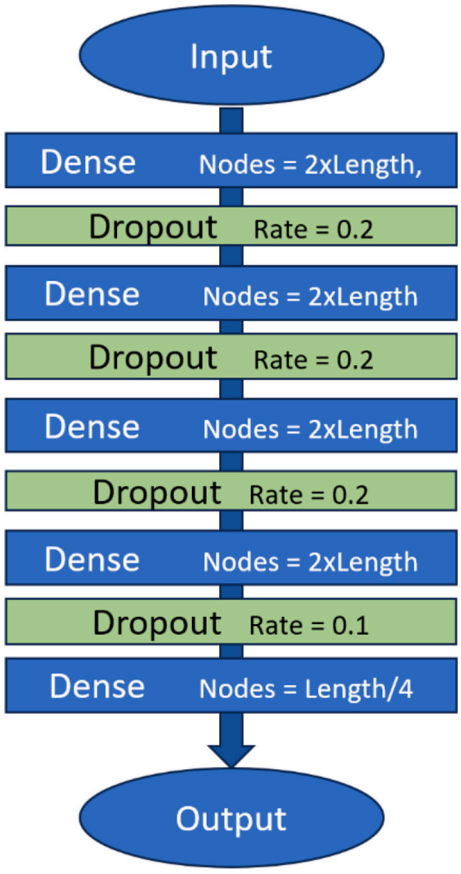


Fig. 6. ANN Construction. Dense layers are alternated with dropout layers.

2.3.3 Convolutional neural network

The CNN used in this study is partly based on the model used in [25]. As shown in Fig. 8, the CNN is formulated into two main blocks of several convolutional and assisting layers. Following each convolution layer is a batch normalization layer that is provided before the ReLU activation. Batch normalization provides reparameterization within a network to help benefit updates as they occur across multiple layers during training [19]. To help prevent the occurrence of the vanishing gradient issue, residual connections were provided at the end of each block creating another connection to the input preceding them. As with

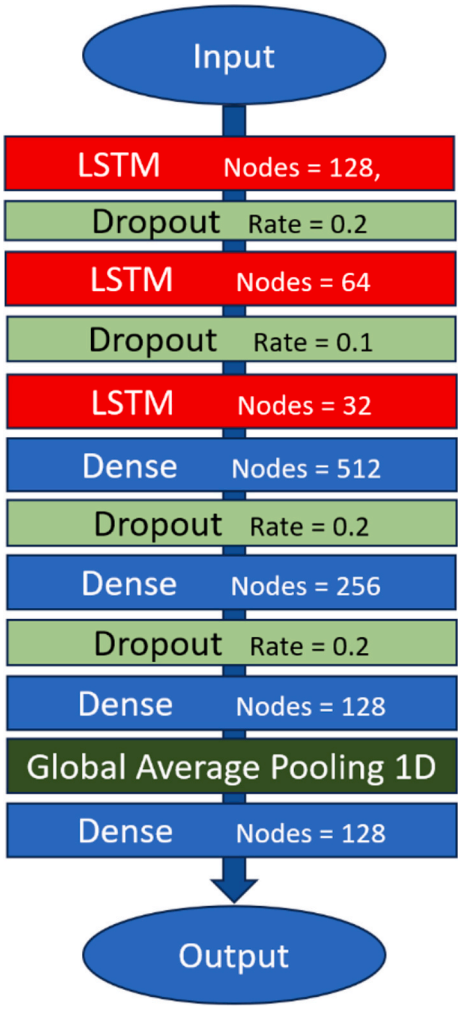


Fig. 7. RNN construction. This model features 3 LSTM layers to provide greater training with the time series sequential data. The pooling layer at the end is implemented to convert the data to 1-dimension.

the RNN, a global average pooling layer is provided near the end of the model to transfer the data from 2-dimensional to 1-dimensional.

1-dimensional convolution was used throughout this model. Compared to its 2-dimensional counterpart, 1-dimensional CNNs have

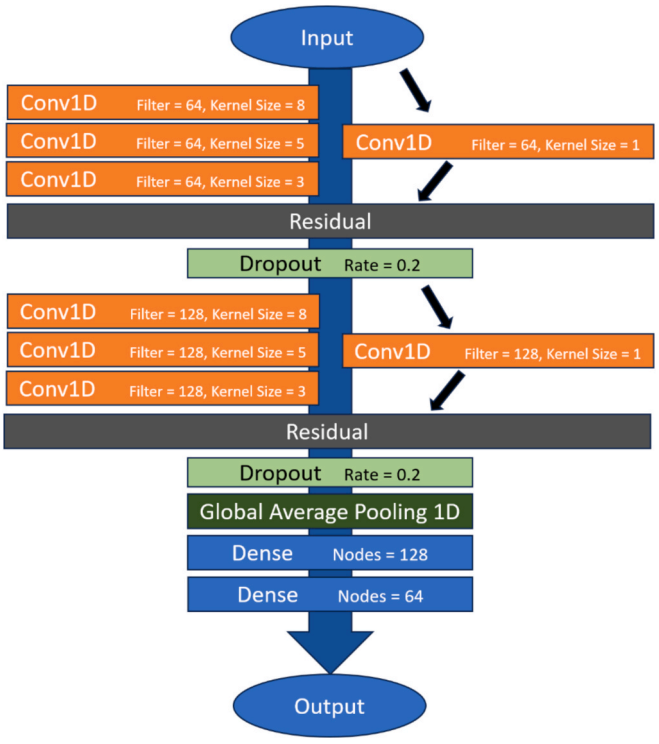


Fig. 8. CNN construction. In two occasions, three convolutional layers with decreasing kernel sizes are computed in parallel with a single convolutional layer with a kernel size of 1 which are then connected with a residual layer.

substantially less computational complexity, which makes them well-suited for lower cost tasks involving 1-dimensional signals [32].

2.3.4 Transformer

The attention mechanism employed by transformer models can be viewed as a mapping of an input query to a dictionary of key-value pair to produce an output which is formed as a weighted summation of the values [33]. In this study, we utilize global self-attention for our time series inputs. Global self-attention involves feeding the model inputs as both the query and key-value pairs (Fig. 9), thus enabling values throughout the time series to be influential [28]. This effectively allows each element in the input to interact with each other element directly with all the outputs being computed in parallel.

The transformer architecture used in this study is provided in Fig. 10. The model includes several multi-head attention layers followed by a residual connection to its input and then a linear normalization layer. Its structure is partly based on [33]. Multi-head attention allows for multiple representations of the inputs into the attention mechanism, which enables it to learn different features from each input [28]. In this study, to keep computational cost low during training, we used 2 heads with a size of 128. Similar to batch normalization in the CNN, the linear normalization layer helps maintain trainability throughout the learning process by helping maintain normalized values throughout the model. A small one-dimensional CNN block is applied to the end this model to help manage the sizeable data still present after the final attention block. As with the CNN and RNN before, Global Average Pooling was then applied.

2.4 Error metrics for evaluation

Error metrics provide a statistical means of inferring ML model performance. These equations receive pairings of ML predictions and their corresponding correct value as input. By including multiple error metrics, a deeper understanding of model performance can be obtained, as some models may perform better or worse across different metrics. This study examines performance with three different error metrics: the Coefficient of Determination (R^2), Mean Absolute Percentage Error (MAPE) and the Root Mean Square Percentage Error (RMSPE):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{3.2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{3.3}$$

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \tag{3.4}$$

The three variables in the above equations, y_i, $\hat{y}_i$, $\overline{y}_i$ represent the actual values, the predicted values and the mean of the actual values respectively.

The Coefficient of Determination provides an assessment of how well the predicted values fit the actual values. The R^2 value typically ranges from 0 to 1, where a value of 1 indicates a perfect fit. As the error metric represents the degree of fit of the model, one of its benefits is the ease of comparing scores across multiple studies [34].

MAE and RMSE are statistical error metrics extensively used for evaluating model performance. There are various arguments and discussions on the applicability of the RMSE and MAE in comparison to each other and which to use [35–37] and therefore both are considered in this study. Referred to as relative or percentage errors, MAPE and RMSPE are preferred in this study over their absolute error equivalents

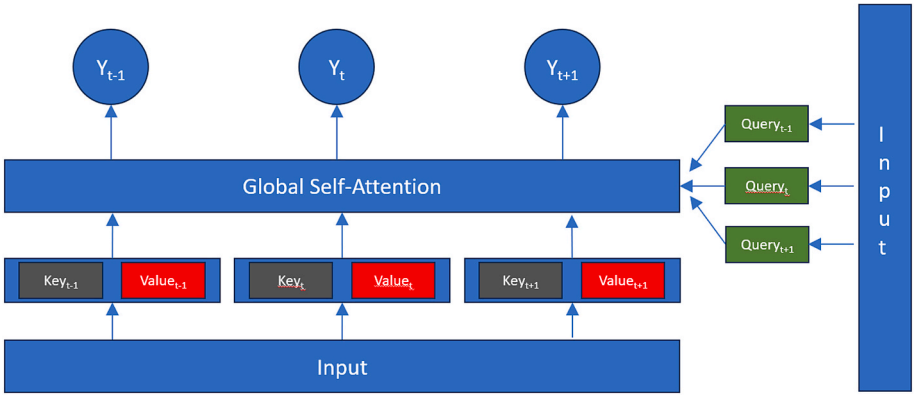


Fig. 9. Global Self-Attention mechanism. The query and key-value pairs are provided to the global self-attention layer. In our case this is a Multi-head Attention layer.

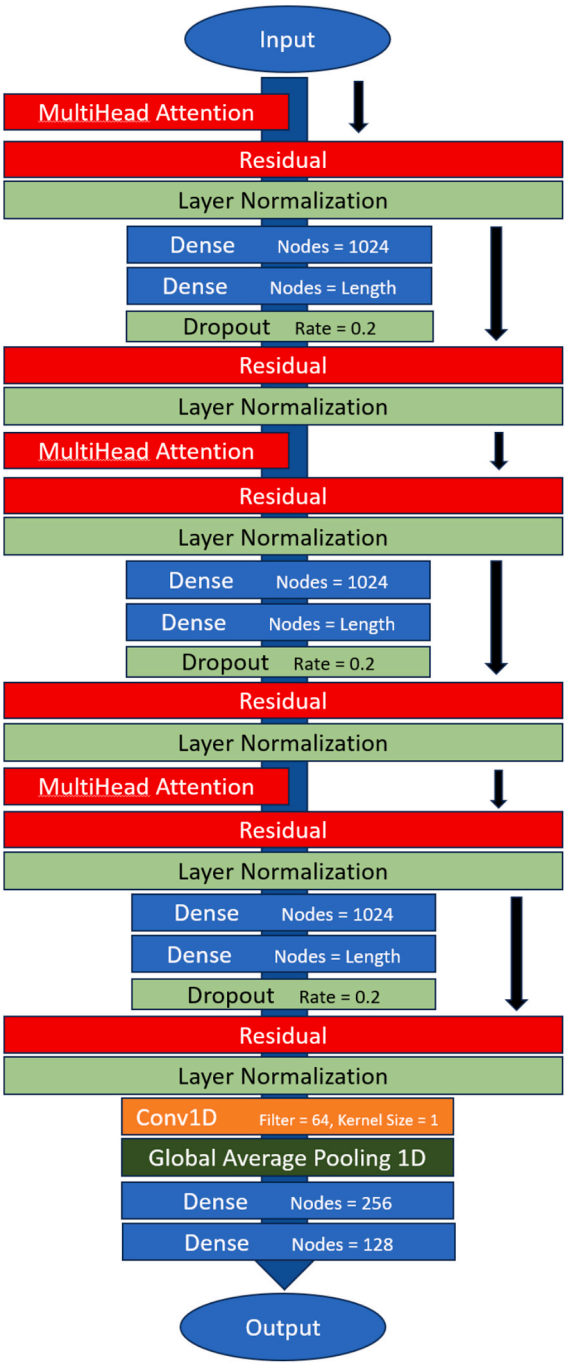


Fig. 10. Transformer construction. Arrows clarify a connection between layers. A single convolutional layer is provided near the end of the model to assist with the lengthy data still present.

(MAE and RMSE) as the parameter values being predicted have different units and scales. Expressing errors as a percentage helps ensure easy comparison between them.

MAPE is commonly used when relative errors are needed; reasons include its intuitive interpretation and adaptability for forecasting applications [38]. The quadratic term in the RMSPE equation helps emphasize instances where significant differences between values exist. In comparisons between models, situations where MAPE scores are similar, but RMSPE scores are farther apart, may signify that a higher percentage of outliers are present in the predictions of one model compared to the other.

3 Results

The results of this study are presented in three sections. Firstly, the performance of the ISMs are examined on each of the three geometric models individually. By doing so, we can assess which ISMs are the best performing and if performance greatly differs between geometric models. The second section of the results examines the prediction ability of the ISMs when trained on BEM data from multiple locations. Also included is a comparison between training with only decay curves from the first section and training with decay and rising curves. The last section of the results examines the performance of the ISMs when trained on data from all the BEM models at once. Only one location is used as it was expected that the decrease in ISM performance would already be significant.

3.1 H-Shape model results

Table 4 gives the error metric results for the ISM models when trained on data from the H-Shape BEM model. Overall, the CNN outperforms the other models with the RNN and transformer exhibiting similar performance to each other. Unsurprisingly, while the ANN is able to have decent prediction performance with a few parameters, it is severely lacking in performance compared to the other ISMs. This shows that the ANN is often unable to discern the influence of less impactful parameters.

Based on the R^2 error metric, predictions for the Glass Conductivity, Infiltration Flow Rate and Maximum Heating Flow Rate parameters preform the most consistently strong. Aside from the Maximum Heating Flow Rate, this can be attributed to their involvement as a component of the building envelope and thereby having a significant impact on the internal temperature of the structure. The Wall Insulation Conductivity and Roof Insulation Conductivity remains highly predictable for both the CNN and RNN, though the performance with the transformer is significantly worse. It is unclear why this occurs, however, as the parameters influence surfaces of every zone, the transformer may struggle to separate the influence of it.

While the R^2 values for the Lighting Energy Power Density and Equipment Energy Power Density are lower than some of the more predictable parameters for the RNN and transformer, the error for the other metrics is among the lowest. This is the same for Equipment Energy Power Density for the CNN. This would suggest that while a goodness of fit is not as achievable as some of the other metrics, their predictions remain relatively consistent (i.e., lack of outliers). This also remains especially true for the ANN, as the R^2 scores suggest practically no fit, however both the MAPE and RMSPE remain low.

Additionally, poor prediction performance on the Ventilation Flow Rate parameter is partly a result of the low performance for the People Density parameter. As the Ventilation Flow Rate is tied to the People Density, making accurate predictions remains difficult as the ISMs struggle to learn the influence of the quantity of people.

Another notable aspect of the results provided in Table 4 is how significantly large the RMSPE are for each of the predictions on the geometric parameters, even though the R^2 remains relatively solid throughout (ignoring the ANN). Given that there is also a large discrepancy between the RMSPE and MAPE scores, this implies that significant outliers have occurred.

Lastly, Fig. 11 demonstrates the decrease in loss over the number of epochs. As the model is trained to predict all parameters, the loss value is influenced by the more difficult parameter values to predict. It can be observed that while the ANN experiences an immediate sudden drop in loss, it plateaus higher and more rapidly than the other models. No more values are reported when the Early Stopping is reached. Due to their similar structures, both the ANN and RNN exhibit less fluctuations than the CNN and transformer. Fluctuations with the transformer possibly help explain the poor prediction performance with some parameters (e.g. Roof Insulation Conductivity) as the consistency in making accurate

Table 4
H-Shape results table.

H-Shape	R^2				MAPE				RMSPE			
Scenario	ANN	CNN	RNN	TRA	ANN	CNN	RNN	TRA	ANN	CNN	RNN	TRA
Wall Insulation Conductivity	0.30	0.84	0.85	0.55	49.2	18.8	18.4	40.8	81.3	28.0	34.8	58.2
Roof Insulation Conductivity	0.35	0.91	0.83	0.21	49.6	15.2	19.8	47.9	84.9	25.0	33.3	61.9
Glass Conductivity	0.58	0.96	0.93	0.93	27.0	7.2	9.8	9.1	39.2	10.2	14.9	14.3
Lighting Energy Power Density	0.11	0.93	0.69	0.63	9.6	2.4	5.0	5.6	11.6	3.2	6.4	7.2
Equipment Energy Power Density	-0.02	0.55	0.62	0.55	10.1	6.8	5.7	6.2	12.1	7.5	7.4	7.5
People Quantity	0.02	0.32	-0.03	0.24	28.8	24.3	30.6	23.7	36.9	33.9	41.0	30.7
Ventilation Flow Rate	0.25	0.72	0.49	0.29	14.9	8.5	11.2	14.6	18.9	11.0	14.7	18.8
Infiltration Flow Rate	0.89	0.98	0.98	0.98	18.1	7.9	6.5	7.8	29.2	13.6	10.3	11.9
Maximum Heating Flow Rate	0.91	0.99	0.98	0.93	16.0	4.3	6.8	13.2	25.6	6.9	10.4	17.5
Scale - X	0.34	0.95	0.94	0.91	>100	28.6	28.5	28.2	>100	83.6	86.6	59.1
Scale - Y	0.06	0.83	0.72	0.81	>100	43.0	54.6	43.9	>100	>100	>100	>100
Scale - Z	0.69	0.97	0.96	0.94	75.6	17.3	19.5	20.6	>100	50.0	57.2	45.9
Mean	0.37	0.83	0.75	0.66	45.97	15.37	18.04	21.80	>100	32.69	37.42	39.00
Range	0.00		1.00		0.0		70.0		0.0		50.0	

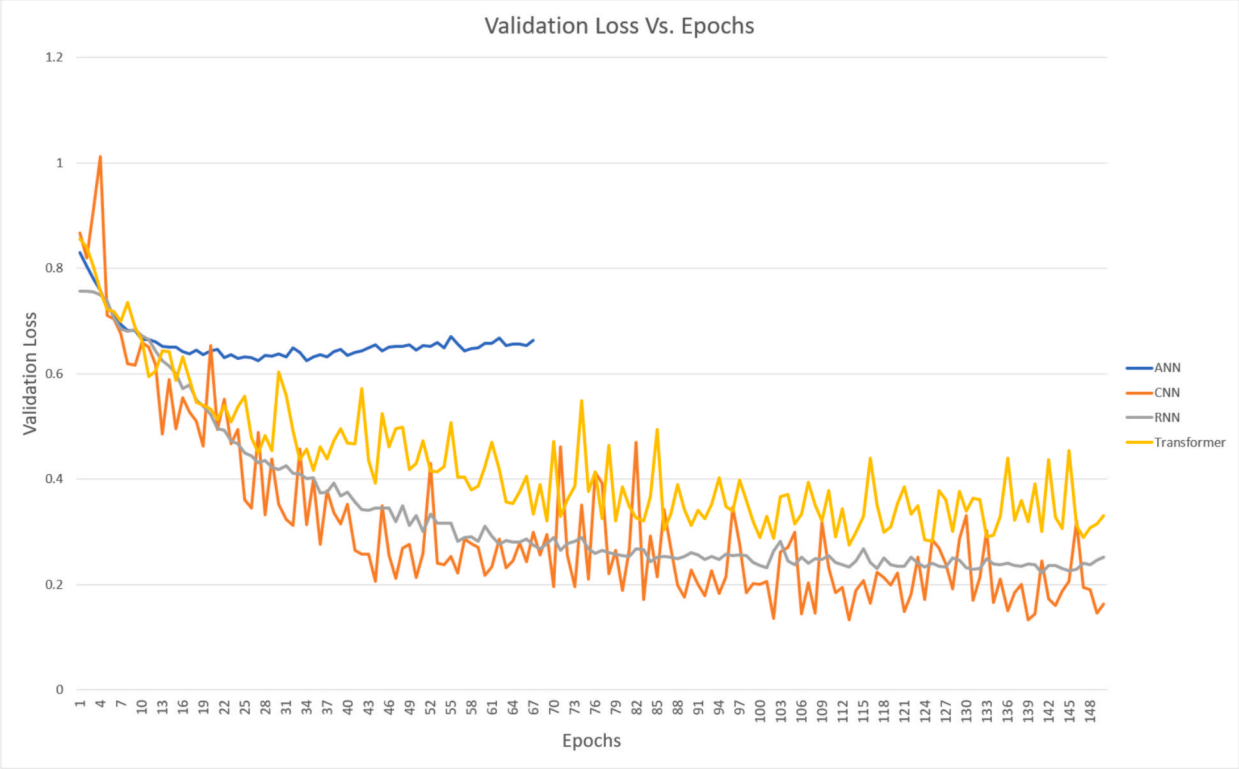


Fig. 11. Validation Loss vs. Epochs.

predictions varies. As neither the CNN, RNN or transformer finish before 150 epochs, the number of training epochs could be increased at the consequence of longer training time.

3.2 H-Shape, C-Shape and Octagonal-Shape result comparisons

R^2 error metric comparison results when the ISMs are trained on data from each of the BEM models are shown in Fig. 12. Findings with MAPE and RMSPE are similar to those in Table 4. Metric values for the geometric parameters have also been omitted for clarity; overall findings are again similar (i.e. high R^2, MAPE and RMSPE scores).

It is observed that switching to a different BEM model for training data does not significantly impact overall performance for the ISMs. An exception to this is the transformer has much stronger predictions for the Wall Insulation Conductivity and Roof Insulation Conductivity with the data from the other BEM models. Overall performance among the ISMs remains similar with the CNN outperforming both the RNN and transformer models, while the ANN poorly performs on all parameters aside from the Maximum Heating Flow Rate and Infiltration Flow Rate.

One of the more notable differences between the BEM models is the higher prediction performance the ISMs have with the H-Shape model data for the Equipment Energy Power Density. Unlike similar parameters (Lighting Energy Power Density and People Quantity), the equipment parameter has a fixed schedule as it remains active overnight. The higher R^2 score indicates that there exists more instances in the H-Shape model data where its influence is noticeable. Interestingly, the Lighting Energy Power Density parameter has noticeably higher predictability for the C-Shape (aside from the CNN), which would suggest that the lighting being active is more frequent. As the equipment parameter would be less isolated in this dataset, it would be expected to be harder to predict.

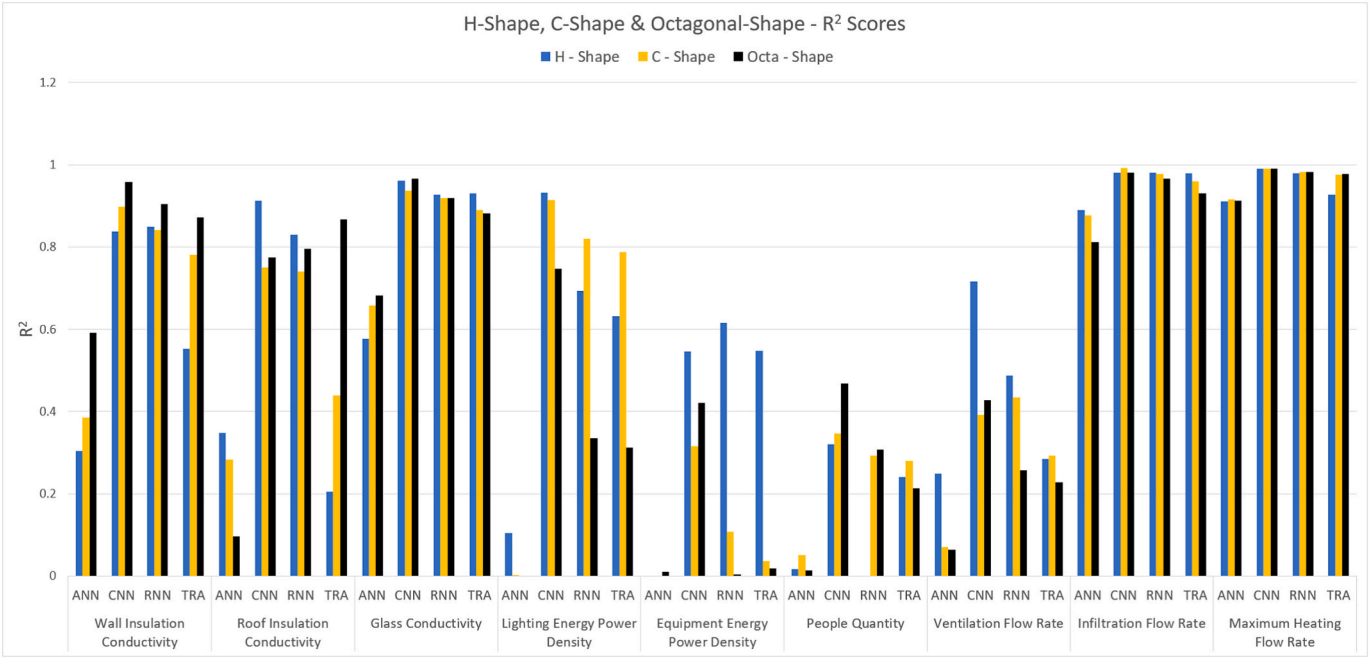


Fig. 12. H-Shape, C-Shape and Octagonal-Shape – R^2 Scores.

Both these parameters are poorly predicted with data from the Octagonal-Shape model. As the floor area remains the same for each zone when scaled, these parameters have no difference in heat provided between zones, making it so that the ISMs have little ability to understand the impact of geometry scaling on them.

3.3 Decay and rising curves and variable locations

Unlike the results provided in Fig. 12, when the ISMs are provided training data with decay and rising curves or from BEM simulations with variable locations, performance differs substantially (Fig. 13). Overall, the ISMs trained on a combination of decay and rising curves exhibit superior performance compared to only being trained on decay curves. The performance improvement can be attributed both to providing the rising curves as well as the overall increase in time series for each sample. Unfortunately, there is a decrease in prediction performance when the ISMs are provided BEM data produced with multiple locations, as is to be expected.

Both the ANN and transformer ISMs experience the largest increase in prediction performance when switched to a combination of decay and rising curves. This shows that the models can take significant advantage of the increase in time series provided, which is especially true in the case of the transformer as it outperforms the CNN in some instances. The ANN still performs poorer than all other ISMs. The CNN performance decreases slightly for several parameters suggesting that increasing the amount of data may be prohibitive for it to learn from, or that it benefits

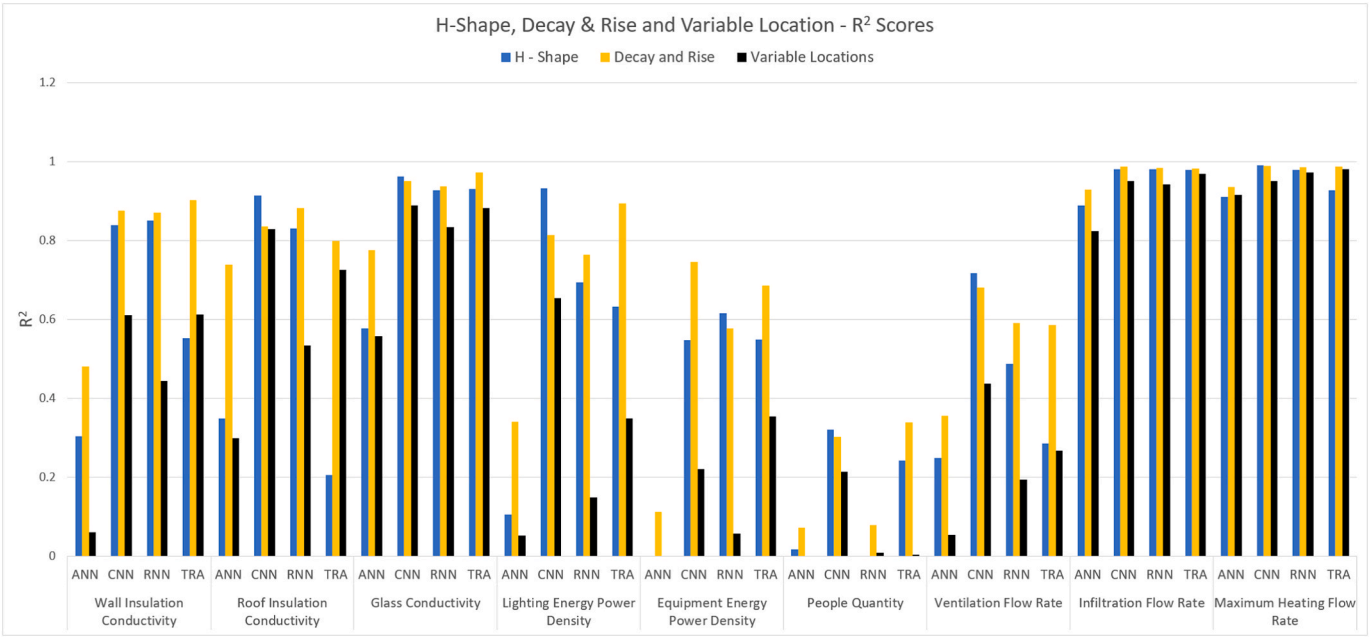


Fig. 13. H-Shape, Decay & Rise, Variable Location – R^2 scores.

more greatly from only using decay curves.

Whereas in Fig. 12 where both the People Density and the Ventilation Flow Rate parameters were similar in terms of performance, this is less apparent when the ISMs are trained on a combination of decay and rising curves. While the performance for the People Density parameter decreased, it increased for the Ventilation Flow Rate. With rising curves, ventilation would become notable as a countermeasure to increases in heat, when scheduled occupancy occurs. With rising curves, the small portion of heat provided by the quantity of people could get lost with the rest of the heat sources.

3.4 All buildings

The results in Fig. 14 show the R^2 score comparison between the ISMs being only trained on the H-Shape model compared to being trained on data from each BEM model at once. Expectedly, the overall prediction performance decreases for practically all parameters. Performance decreases are more significant on parameters that previously observed more varied prediction performance in Figs. 12 and 13 (i.e. Roof Insulation Conductivity, Lighting Energy Power Density and Equipment Energy Power Density).

While weaker, the ISMs still retain the ability to make adequate and sometimes strong predictions for several parameters. Like the results provided in Figs. 12 and 13, the Infiltration Flow Rate, Maximum Heating Flow Rate and Glass Conductivity remain suitable for the ISMs to predict. As only 2000 samples of each BEM model were used for training with all BEM models at once (compared to a full 5000 when trained individually), expanding the training dataset to include more samples of each model may demonstrate a performance benefit.

4 Conclusions

This study examined the generalizability of ISMs in their ability to predict BEM parameters from typical BEM output data. Each of the three BEM models from which training data was generated had scalable geometry requiring the ISMs to be able to decipher parameter values regardless of the exact shape of the building. BEM output data to be fed into the ISM was structured as a collection of time series data in the form of decay curves. Overall, the CNN preformed the strongest with the RNN and Transformer typically performing weaker. The ANN, while a suitable benchmark, struggled to make out the influence of different parameters and therefore its performance was more limited to only being able to predict the values accurately for a small number of parameters. This was easily observed with CNN, RNN, Transformer and ANN having average parameter prediction R2 values of 83 %, 72 %, 81 % and 37 % respectively when trained on data from the H-Shape model. Additionally we examined the prediction performance of the ISMs when trained on a mixture of decay and rising curves, as well as various locations. It was found that the ISMs typically performed more strongly with a combination of decay and rising curves, however, this may be in large part due to the increase in the number of time series provided. Testing with variable locations, as well as training the ISMs on all BEM models at once, demonstrated a reduction in performance, but predictions with several parameters were still satisfactory for the CNN, RNN and Transformer.

In terms of future work, exploring modifications to hyperparameters and the ML model construction for each of the ISMs would potentially prove insightful. While an ideal hyperparameter and model construction combination for one BEM model may not be ideal for another, an examination of this in some form would be beneficial for comparisons between ISMs as well as demonstrate their resilience to further changes. Additional further examination into the predictability of lesser predictable parameters would provide insight into what caused the prediction performance on them to notably fluctuate between BEM models in this study. Lastly, surrogate models have shown extensive use in the domain of building performance optimization [11]. It may additionally prove meaningful to examine the performance of the DL ISMs in comparison to applying an optimization strategy that incorporates SM whereby it tries to match BEM output data with an optimized set of inputs. Doing so, would demonstrate the potential benefits and constraints of either approach.

Overall, ISMs have been shown to provide a promising means of placing accurate predictions on BEM parameters even when the building geometry is varied. The prediction accuracy can vary considerably depending on the ISM model used in addition to the parameter being predicted. Prediction accuracy decreased when the ISMs were trained on data from multiple locations or on each core building geometry at once, however it still remained sufficient for several parameters.

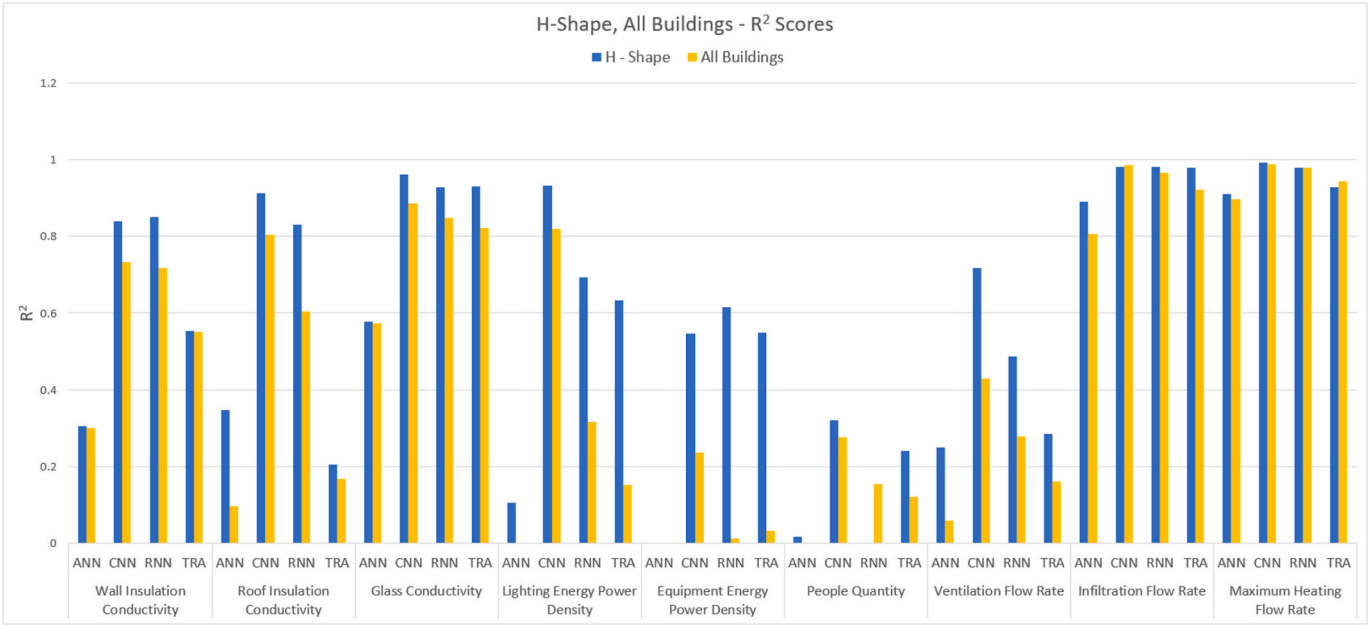


Fig. 14. H-Shape, All Buildings – R^2 scores.

CRediT authorship contribution statement

Liam Jowett-Lockwood: Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Ralph Evins:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ralph Evins reports financial support was provided by Natural Sciences and Engineering Research Council of Canada. Ralph Evins reports a relationship with Building Atlas Ltd that includes: consulting or advisory and equity or stocks. Liam Jowett-Lockwood reports a relationship with RJC Engineers that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] H. Wang, Z. (John) Zhai, Advances in building simulation and computational techniques: a review between 1987 and 2014, Energ. Build. 128 (2016) 319–335, https://doi.org/10.1016/j.enbuild.2016.06.080.

[2] "National Housing Strategy." Cmhc-schl.gc.ca. https://www.cmhc-schl.gc.ca/nhs/guidepage-strategy (accessed Feb. 12, 2024).

[3] 2022 Global Status Report for Buildings and Construction., United Nations Environment Programme. Geneva, Switzerland, 2022.

[4] "EnergyPlus" energyplus.net. https://energyplus.net/ (accessed October 7, 2023).

[5] "VE Virtual Environment." IESVE.com. https://www.iesve.com/software/virtual-environment (accessed Feb. 12, 2024).

[6] "Design Builder." designbuilder.co.uk. https://designbuilder.co.uk/ (accessed Feb 12. 2024).

[7] H. Gao, C. Koch, Y. Wu, Building information modelling based building energy modelling: a review, Appl. Energy 238 (2019) 320–343, https://doi.org/10.1016/j.apenergy.2019.01.032.

[8] E. Fabrizio, V. Monetti, Methodologies and advancements in the calibration of building energy models, Energies 8 (4) (2015) 2548–2574, https://doi.org/10.3390/en8042548.

[9] "PathFinder." Buildingpathfinder.com. https://www.buildingpathfinder.com/ (accessed Feb 12.2024).

[10] A. Chong, Y. Gu, H. Jia, Calibrating building energy simulation models: a review of the basics to guide future work, Energ. Build. 253 (2021) 111533, https://doi.org/10.1016/j.enbuild.2021.111533.

[11] P. Westermann, R. Evins, Surrogate modelling for sustainable building design – a review, Energ. Build. 198 (2019) 170–186, https://doi.org/10.1016/j.enbuild.2019.05.057.

[12] D. Hou, R. Evins, A protocol for developing and evaluating neural network-based surrogate models and its application to building energy prediction, Renew. Sustain. Energy Rev. 193 (2024) 114283, https://doi.org/10.1016/j.rser.2024.114283.

[13] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron. Mark. 31 (3) (2021) 685–695, https://doi.org/10.1007/s12525-021-00475-2.

[14] P.W. Tien, S. Wei, J. Darkwa, C. Wood, J.K. Calautit, Machine learning and deep learning methods for enhancing building energy efficiency and indoor environmental quality – a review, Energy AI 10 (2022) 100198, https://doi.org/10.1016/j.egyai.2022.100198.

[15] Daniel, Graupe, Principles of Artificial Neural Networks, third ed., World Scientific, Singapore, 2013.

[16] A. Géron, Hands-on Machine Learning with Scikit-learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, second ed., O'Reilly, Sebastopol, 2019.

[17] R. Olu-Ajayi, H. Alaka, I. Sulaimon, F.T. Sunmola, S.O. Ajayi, Building energy consumption prediction for residential buildings using deep learning and other machine learning techniques, J. Build. Eng. 45 (2022) 103406, https://doi.org/10.1016/j.jobe.2021.103406.

[18] G. Suryanarayana, J. Lago, D. Geysen, P. Aleksiejuk, C. Johansson, Thermal load forecasting in district heating networks using deep learning and advanced feature selection methods, Energy (oxford) 157 (2018) 141–149, https://doi.org/10.1016/j.energy.2018.05.111.

[19] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, The MIT Press, Cambridge, Massachusetts, 2016.

[20] F.M. Salem, in: Recurrent Neural Networks: From Simple to Gated Architectures, first ed., Springer International Publishing AG, Cham, 2022 https://doi.org/10.1007/978-3-030-89929-5.

[21] 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, https://doi.org/10.1016/j.apenergy.2018.12.004.

[22] S. Jung, J. Moon, S. Park, E. Hwang, An attention-based multilayer GRU model for multistep-ahead short-term load forecasting, Sensors (Basel, Switzerland) 21 (5) (2021) 1639, https://doi.org/10.3390/s21051639.

[23] J. Jang, J. Han, S.-B. Leigh, Prediction of heating energy consumption with operation pattern variables for non-residential buildings using LSTM networks, Energ. Build. 255 (2022) 111647, https://doi.org/10.1016/j.enbuild.2021.111647.

[24] N. Somu, G. Raman, K. Ramamritham, A deep learning framework for building energy consumption forecast, Renew. Sustain. Energy Rev. 137 (2021) 110591, https://doi.org/10.1016/j.rser.2020.110591.

[25] P. Westermann, M. Welzel, R. Evins, Using a deep temporal convolutional network as a building energy surrogate model that spans multiple climate zones, Appl. Energy 278 (2020) 115563, https://doi.org/10.1016/j.apenergy.2020.115563.

[26] S. Ferreira, B. Gunay, A. Ashouri, S. Shillinglaw, Unsupervised learning of load signatures to estimate energy-related building features using surrogate modelling techniques, Build. Simul. 16 (7) (2023) 1273–1286, https://doi.org/10.1007/s12273-023-1005-5.

[27] F. Herbinger, C. Vandenhof, M. Kummert, Building energy model calibration using a surrogate neural network, Energy Build. 289 (2023) 113057, https://doi.org/10.1016/j.enbuild.2023.113057.

[28] U. Kamath, in: Transformers for Machine Learning: A Deep Dive, first ed., CRC Press, Boca Raton, Florida, 2022 https://doi.org/10.1201/9781003170082.

[29] "Group – Zone Forced Air Units" bigladdersoftware.com https://bigladdersoftware.com/epx/docs/8-0/input-output-reference/page-032.html#zonehvacidealloadsairsystem (accessed October 7, 2023).

[30] "6.3. Preprocessing data." scikit-learn.org https://scikit-learn.org/stable/modules/preprocessing.html.

[31] ANSI/ASHRAE Addendum a to ANSI/ASHRAE Standard 169-2020 – Climatic Data for Building Design Standards, 169-2020, ASHRAE, Atlanta, GA, USA, Oct. 2021. [Online]. Available: https://www.ashrae.org/file%20library/technical%20resources/standards%20and%20guidelines/standards%20addenda/169_2020_a_20211029.pdf.

[32] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D.J. Inman, 1D convolutional neural networks and applications: a survey, Mech. Syst. Signal Process. 151 (2021) 107398, https://doi.org/10.1016/j.ymssp.2020.107398.

[33] Vaswani, Ashish, et al., Attention is all you need. Advances in neural information processing systems, 2017, doi: 10.48550/arXiv.1706.03762.

[34] T. Østergård, R.L. Jensen, S.E. Maagaard, A comparison of six metamodeling techniques applied to building performance simulations, Appl. Energy 211 (2018) 89–103, https://doi.org/10.1016/j.apenergy.2017.10.102.

[35] D.S.K. Karunasingha, Root mean square error or mean absolute error? Use their ratio as well, Inf. Sci. 585 (2022) 609–629, https://doi.org/10.1016/j.ins.2021.11.036.

[36] T.O. Hodson, Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, Geosci. Model Dev. 15 (14) (2022) 5481–5487, https://doi.org/10.5194/gmd-15-5481-2022.

[37] T. Chai, R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature, Geosci. Model Dev. 7 (3) (2014) 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014.

[38] A. de Myttenaere, B. Golden, B. Le Grand, F. Rossi, Mean absolute percentage error for regression models, Neurocomputing (Amsterdam) 192 (2016) 38–48, https://doi.org/10.1016/j.neucom.2015.12.114.