

# Forecasting Building Thermal Demand Using Machine Learning<sup>#</sup>

Corcoran L<sup>1\*</sup>, Carr D<sup>1</sup>, Ugalde-Loo CE<sup>1</sup>

<sup>1</sup> School of Engineering, Cardiff University, Wales, UK

(\*Corresponding Author: CorcoranL@cardiff.ac.uk)

## ABSTRACT

Rising summer temperatures in the UK, alongside stricter building regulations on thermal efficiency, are increasing the importance of residential cooling demand. While very few households currently use active cooling, adoption is expected to grow, placing additional pressure on electricity networks. This paper presents a study combining physics-based modelling with machine learning to forecast building thermal demand using basic dwelling characteristics and weather data. Results show that the method achieves high accuracy in predicting the thermal demand of previously unseen dwellings—showcasing the potential extreme gradient boosting may have in forecasting cooling demand in a warming world.

**Keywords:** building thermal demand, machine learning, extreme gradient boosting, IES VE

## NOMENCLATURE

### Abbreviations

3D	3-dimensional
CSV	Comma separated value
EPW	EnergyPlus Weather
XGBoost	Extreme gradient boosting

## 1. INTRODUCTION

The UK Meteorological Office predicts that the country will face an increase in both the frequency and intensity of summer heatwaves alongside a warmer climate [1]. As a result, cooling demand in dwellings is becoming increasingly important. At the same time, UK Building Regulations continue to mandate higher levels of thermal efficiency [2], which can exacerbate overheating risks. Many residents will turn to installing active cooling measures such as air conditioning units to combat extreme heat in dwellings.

According to the Energy Follow-Up Survey (2021) [3], only 2% of UK homes currently have active cooling devices installed, but this proportion is expected to rise sharply in the coming decades. National Grid, the largest

electrical transmission and distribution operator in the UK, estimates that additional cooling could add up to 39 GW to the system by 2050 on typical weekend days [4].

Although passive cooling strategies such as night purging and the use of shading devices can help reduce cooling loads, they are not always sufficient to maintain indoor temperature conditions at safe levels [5]. This makes it essential to quantify future thermal demand in dwellings and to enable active adjustments to grid operation and planning.

While physics-based simulation models can provide accurate representations of dwelling thermal demand, they are computationally intensive and unsuitable for large-scale or real-time applications. This creates a critical gap: there is currently no scalable method to forecast thermal demand (both cooling and heating) in UK dwellings using only basic input parameters.

Machine learning offers a promising solution to the aforementioned problem, with proven effectiveness in handling complex, tabular data [6]—yet its application to UK residential cooling demand remains limited. This paper addresses this gap by combining a physics-based approach with machine learning to develop a predictive model capable of rapidly forecasting hourly thermal demand across a range of UK dwellings and climates.

The presented work builds on the methodology introduced in [2], which enabled generating thermal demand profiles for representative UK dwellings over a typical year across multiple locations. These thermal demand profiles are here used to train a predictive model using machine learning. The model employs weather forecast data and basic dwelling properties to predict thermal demand in any specified UK dwelling.

The predictive model is based on extreme gradient boosting (XGBoost). This is a machine learning algorithm particularly effective for large, structured datasets [6]. Unlike deep learning methods which use multi-layered neural networks, XGBoost is an ensemble-based approach. The XGBoost model was trained to predict thermal demand from key input parameters: outdoor

<sup>#</sup> This is a paper for the 17th International Conference on Applied Energy (ICAE2025), December 8-12, 2025, Bangkok, Thailand.

temperature, wind speed, solar radiation and dwelling characteristics (orientation, heat transfer coefficients).

Once the model has been trained and provided weather data files are available, the presented approach offers a fast and scalable method to quantify thermal demand for most UK dwellings at any time horizon.

## 2. MATERIALS AND METHODS

### 2.1 Pool of thermal demand profiles

Following the methodology presented in [2], the yearly thermal demand for the most common UK dwelling types (i.e. bungalows, detached houses, flats, semi-detached houses and terraced houses) across five different locations was calculated. Each dwelling was simulated across the full set of dwelling codes adopted in [2] to represent variations in thermal efficiency levels. Four dwelling orientations were considered: north, east, south, west. The thermal demand profiles obtained from the simulations were used to build a comprehensive pool of data. Internal gains were not considered in the simulations and an internal set point temperature of 21°C was used to quantify cooling demand.

The five selected locations (Cardiff, Glasgow, London, Manchester and Plymouth) were chosen to capture the diversity of UK weather zones, based primarily on differences in solar radiation intensity, as illustrated in Fig. 1. Weather data for each location were stored in EnergyPlus Weather (EPW) files, which contain the necessary inputs for modelling, including outdoor dry bulb temperature, humidity, radiation, cloud cover and wind speed. These files were obtained directly from an open-access data repository [7].

Building simulations were conducted using a physics-based modelling approach in IES VE—a commercially licensed building modelling software. The key inputs included: (i) 3-dimensional (3D) dwelling models constructed using images and floor plans obtained from [9]; (ii) physical properties of building elements such as floors, walls, roofs and windows/doors; and (iii) location-specific yearly weather data. With these inputs, IES VE calculated the hourly thermal demand using the heat balance equation and related governing equations.

At the time of writing this paper, the pool of data considered ~185,000 yearly thermal demand datasets, each containing 8,760 hourly values (i.e. one full year of data). Creating this resource required substantial computational capabilities due to the size and resolution of the simulations. This data is specific to UK dwellings as it incorporates construction typologies and thermal performance levels defined by UK Building Regulations.

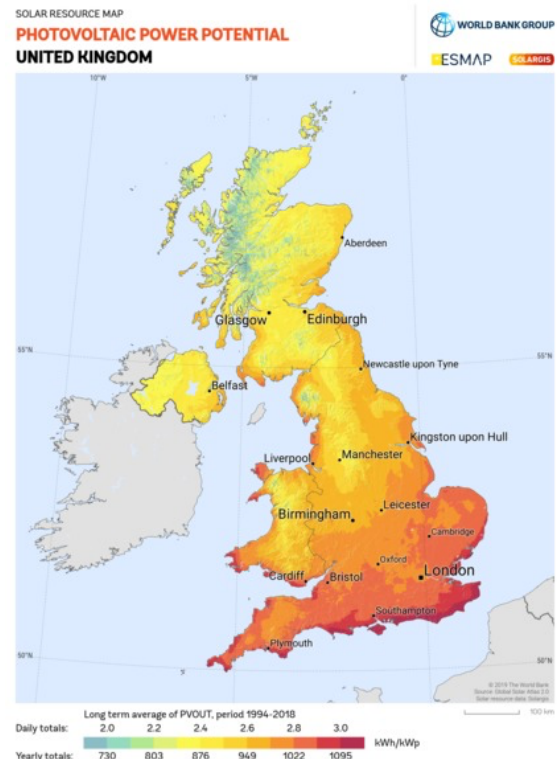


Fig. 1. Photovoltaic power potential in the UK [8].

### 2.2 XGBoost

The predictive model was developed using the XGBoost algorithm. This is a scalable and efficient implementation of gradient boosting that builds an ensemble of decision trees sequentially. Each tree is trained to correct the errors of its predecessors by minimising a loss function using gradient descent. XGBoost incorporates regularisation to reduce overfitting and is particularly well-suited for structured tabular data, making it a strong choice for capturing complex, non-linear relationships between input variables [10].

The training data consisted of the thermal demand profiles described in Section 2.1, with the following variables provided on an hourly basis:

- Outdoor dry bulb temperature
- Outdoor dew point temperature
- Global radiation
- Diffuse radiation
- Wind speed and direction (vector format)
- Total sky cover
- Total heat transfer coefficient
- Building orientation
- Building volume
- Dwelling type
- Building thermal demand (target variable)

As XGBoost is a gradient boosting method, feature normalisation was not required. However, cyclical time variables (e.g. hour of day, day of year) were normalised to better represent their periodic nature.

To evaluate model performance, the data pool was partitioned into training and validation subsets using an 80/20 split. Rather than a random division, one location from the dataset was designated as the validation set to ensure independence. To select the most appropriate UK location for the model validation, correlations of global radiation profiles were compared across sites. Global radiation was chosen for this comparison as it provides the clearest indication of sunrise and sunset times, with differences between locations primarily reflected in radiation intensity. Cardiff and Plymouth exhibited the highest similarity, with a correlation coefficient of 0.85; therefore, thermal demand data for Plymouth was selected as the validation dataset.

Model optimisation was carried out iteratively, involving heuristic adjustments to both data preparation and model hyperparameters. Parameters tuned included learning rate, maximum tree depth and the number of trees (i.e. estimators) in the ensemble model.

For the purposes of this paper, the focus was placed on a detached house typology. This allowed for a more detailed tuning and validation of the model, while also avoiding the inclusion of a large number of similar results, thereby improving clarity and interpretability.

A test case was conducted using an unseen detached house located in a different city, Leeds, with a north-east orientation not present in the training data. The only available dwelling information other than the EPW file was the following:

- Dwelling type
- Dwelling volume
- Ground, wall, window and roof U-values and surface areas
- Orientation

The dwelling in Leeds was also modelled in IES VE following the methodology described in [2], in which floor plans were used to construct a 3D model assigned with unique U-values. Simulated thermal demand profiles were then compared with the XGBoost predictions to assess the model's ability to generalise beyond the training database.

The following section outlines the overall modelling approach and data handling workflow, including the use of Python and Parquet files, which were used to store and organise the simulation data efficiently.

## 2.3 Modelling Approach

All the modelling work carried out for this paper was done using Python. It was selected due to the widely available libraries for data handling and machine learning. The thermal demand data generated were stored locally as comma separated value (CSV) files.

The workflow consisted of three stages:

1. *Data preparation*: this stage involved extracting the relevant weather data from the EPW files, the thermal demand from each of the CSV files and assigning the correct dwelling parameters to each row within the input dataframe. The prepared data are then stored in a dataframe and saved as a Parquet file ready for the model.
2. *Model training and optimisation*: XGBoost would read and conduct training on each individual Parquet file to minimise memory usage. These models were run multiple times adjusting the input data and the model hyperparameters.
3. *Validation and testing*: Model predictions for one location (Plymouth) were compared against the simulated thermal demand to evaluate performance. These results were then compared to the simulated results. A test case was also carried out using a detached house in Leeds.

This approach allowed for rapid experimentation with multiple model configurations while maintaining reproducibility and scalability for larger datasets.

## 3. RESULTS AND DISCUSSION

### 3.1 Validation results

The model was trained using the methodology described in Section 2.3. A scatter plot of the predicted thermal demand versus the simulated thermal demand for a detached house in Plymouth is shown Fig. 2.

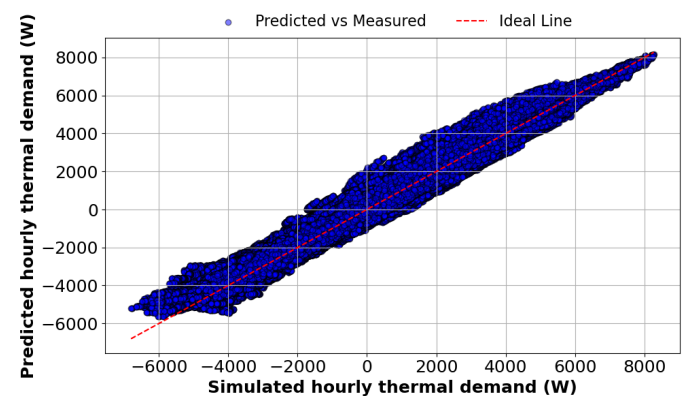


Fig. 2. Scatter plot of XGBoost-predicted versus simulated thermal demand for a detached house in Plymouth (validation dataset).

As shown in Fig. 2, results from the predictive model show an extremely close agreement with the simulated data. This is further corroborated with a coefficient of determination score of  $R^2 = 0.987$ . At the extremes, representing very high heating or cooling demand, the model tends to underpredict the loads. In the case of cooling, since demand is represented as a negative value, such underprediction appears as predicted values having a larger negative amplitude than the ones obtained from simulations. This can be appreciated from the error distribution shown in Fig. 3.

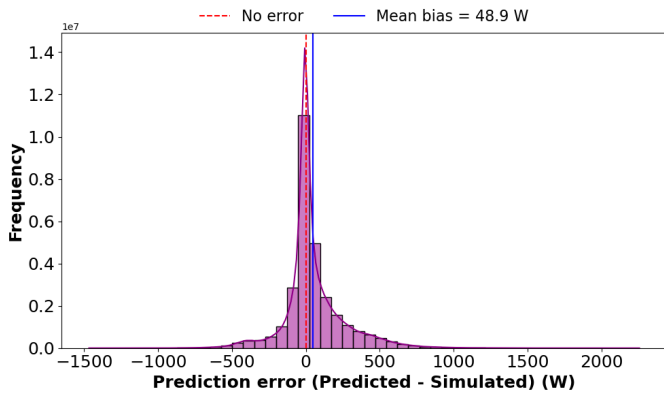


Fig. 3. Error distribution of XGBoost-predicted versus simulated hourly thermal demand for a detached house in Plymouth.

As shown in Fig. 3, the mean bias indicates an average overprediction of 48.9 W, which arises because the model underpredicts the magnitude of cooling demand. This is also visible in Fig. 2, where a greater density of deviations from the ideal line occurs for cooling cases, leading to simulated values of cooling demand being lower than the predictions and thus increasing the mean bias.

To further investigate the systematic behaviour of prediction errors, Fig. 4 shows the error between predicted and simulated values against the hourly thermal demand obtained from simulations. This graph highlights that while errors are generally centred around zero, the model tends to underpredict for large values of cooling demand (large negative values) and heating demand (large positive values). The bias becomes more pronounced in these regions, which is consistent with the deviations observed in the scatter plot (Fig. 2).

Time series data for a single, randomly selected detached house were taken and shown in Fig. 5 to better help visualise the model’s predictions. From these results, it can be observed the XGBoost model performs well. However, it struggles with peak cooling demand, which is consistent with the discussion on Figs. 2 to 4.

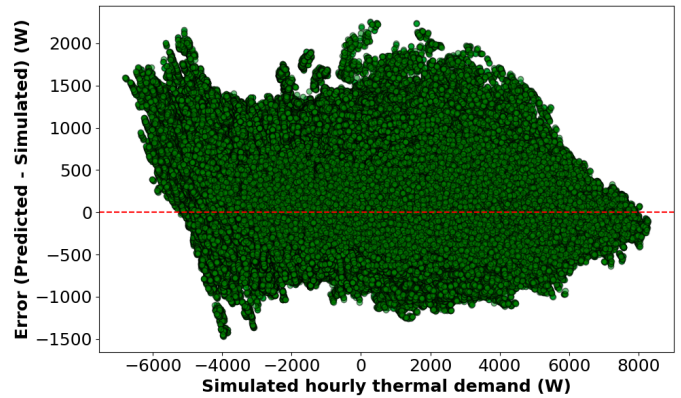


Fig. 4. Bias plot of XGBoost-predicted versus simulated thermal demand for a detached house in Plymouth.

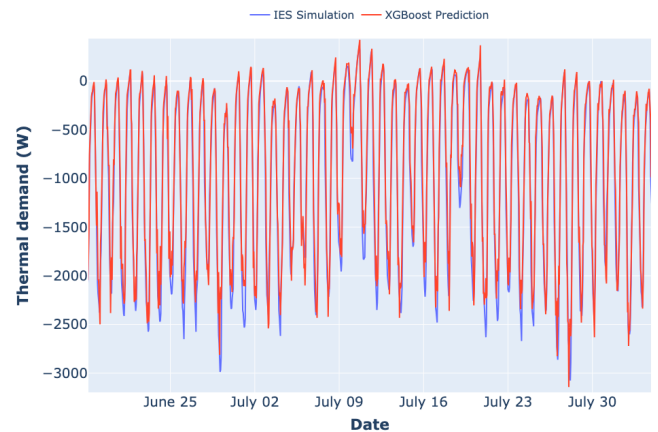
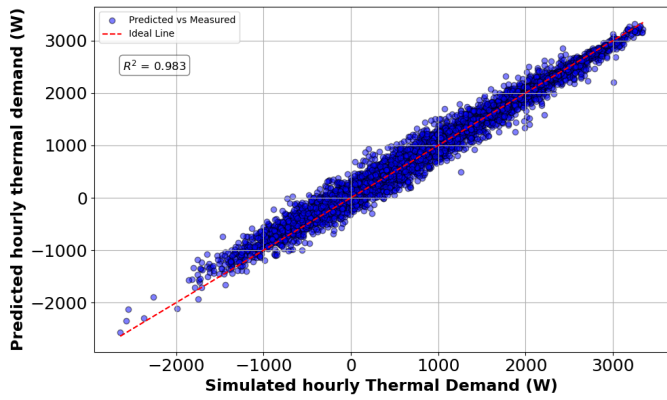


Fig. 5. Hourly time series comparison of predicted and simulated thermal demand for a randomly selected detached dwelling in Plymouth.

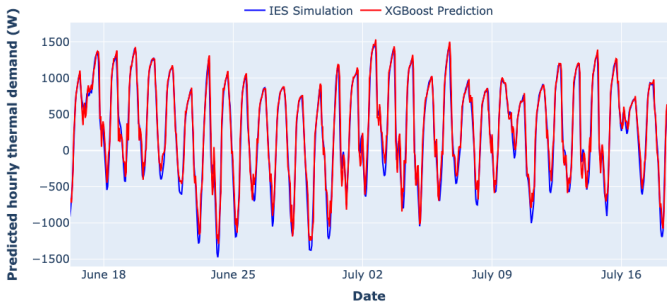
### 3.2 Detached house in Leeds

As stated in Section 2.2, a detached house in Leeds was also modelled in IES VE. This dwelling was assigned randomised U-values not previously shown to the XGBoost model and given a north-east orientation, whereas the model had only been trained on north, east, south and west orientations. The weather file and the dwelling data were fed to the XGBoost model and the results compared to the simulated data. A scatter plot and the time series data are shown in Fig. 6.

Similar to the house in Plymouth (see Fig. 2), the scatter plot for the detached house in Leeds (Fig. 6a) shows that the model underpredicts peak heating and cooling demand. This is corroborated in Fig. 6b where the peak cooling demand is consistently underpredicted. However, the model in general consistently predicts thermal demand with an extremely high degree of accuracy. This is shown with the calculated score  $R^2 = 0.983$ .



(a)



(b)

Fig. 6. Comparison of model predictions versus simulated results for a detached house in Leeds: (a) scatter plot; (b) one-month time series.

For the house in Leeds, the XGBoost prediction model simulated a full year of thermal demand in just 1.3 seconds. In addition, this only required basic dwelling and weather information. In sharp contrast, physics-based modelling (not considering setup and calibration) typically requires almost a minute to run per dwelling, and depending on complexity and user experience, can take up to half a day to prepare and execute.

### 3.3 On the contributions and limitations of the work

The presented modelling approach to forecast building cooling demand considerably reduces computational time compared to traditional physics-based modelling, enabling fast, scalable predictions suitable for electricity network planning, demand management and long-term scenario analysis. Such a tool may be of value to a range of stakeholders, including electricity network operators and system planners, policy makers and regulators, and building developers and engineers.

By using only basic dwelling characteristics alongside readily available weather data, the forecasting model achieves high predictive accuracy and generalises effectively to unseen dwellings with different locations and unique properties. To the best of the authors' knowledge, such a tool to predict the thermal demand of

residential dwellings is not available in the open literature. At the same time, the pool of thermal demand profiles constitutes a well-founded resource enabling a myriad of case studies.

However, a few limitations of the work should be noted, which may constitute the basis for future research. The model was trained on simulated data under fixed assumptions, including no internal heat gains due to occupancy or appliance utilisation, and a constant indoor temperature set point. These assumptions may reduce accuracy for dwellings with variable occupancy or advanced heating, ventilation and air conditioning systems.

Additionally, the pool of thermal demand profiles used in the paper does not include all possible dwelling types and extreme climatic conditions such as heatwaves have not been considered. Expanding the range of dwelling types and modelling approaches to further improve prediction accuracy, particularly under extreme temperatures, would further enhance the predictive model.

## 4. CONCLUSIONS

This paper investigated the combination of a computationally intensive physics-based modelling approach with a machine learning-based algorithm to enable faster predictions of building thermal demand. This requires only basic dwelling and weather data.

The work provided accurate predictions when compared against the results for the same dwellings that the model was trained on but with a different location. A score  $R^2 = 0.987$  was achieved for a house in Plymouth, providing high confidence on the prediction. A similar high accuracy was achieved for a new dwelling with a different location (Leeds), orientation (north-east) and unique U-values. In this case, a score  $R^2 = 0.983$  was obtained.

Results show that once the XGBoost model has been trained, it is possible to provide a fast and effective prediction of the hourly thermal demand for any typical UK dwelling. For instance, the yearly thermal demand for the house in Leeds was obtained in 1.3 seconds, whereas physics-based approaches may take up to a minute to run (excluding modelling time). This represents a major reduction in computational time, making the approach valuable for network operators forecasting and managing peak thermal demand across multiple areas. This can also support long-term scenario analysis to identify areas of the electricity grid most at risk from rising cooling demand.

## ACKNOWLEDGEMENT

The work presented in this paper was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK Research and Innovation, through the project 'Flexibility from Cooling and Storage (Flex-Cool-Store)' under grant EP/V042505/1.

## REFERENCES

- [1] Lowe JA, Bernie D, Bett P, Bricheno L, Brown S, Calvert D, et al. UKCP18 science overview report. Met Office Hadley Centre: Exeter, UK 2018:1–73.
- [2] Corcoran L, Saikia P, Ugalde-Loo CE, Abeysekera M. An effective methodology to quantify cooling demand in the UK housing stock. *Applied Energy* 2025;380:125002.  
<https://doi.org/10.1016/j.apenergy.2024.125002>.
- [3] Department for Business, Energy & Industrial Strategy. Energy Follow Up Survey: thermal comfort, damp and ventilation. 2021.
- [4] Department for Business, Energy & Industrial Strategy. Clean Growth - Transforming Heating. 2018.
- [5] Corcoran L, Ugalde-Loo CE, King L, Demski C, Lowes R. Analysing the Effects of Common Passive Cooling Strategies in UK Homes, Japan: 2024.  
<https://doi.org/10.46855/energy-proceedings-10996>.
- [6] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* 2015;349:255–60.  
<https://doi.org/10.1126/science.aaa8415>.
- [7] Climate One Building. Repository of free climate data for building performance simulation. n.d.  
<https://climate.onebuilding.org/default.html> (accessed March 21, 2024).
- [8] Solargis. Solar resource maps of United Kingdom n.d.  
<https://solargis.com/maps-and-gis-data/download/united-kingdom> (accessed March 21, 2024).
- [9] Rightmove. Rightmove n.d.  
<https://www.rightmove.co.uk/> (accessed April 30, 2024).
- [10] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: Association for Computing Machinery; 2016, p. 785–94.  
<https://doi.org/10.1145/2939672.2939785>.