The Impact of Occupancy on Baseline Building Energy Modelling Performance

by

Md Rezaul Karim Chowdhury

A dissertation submitted in partial fulfilment of the requirements

for the degree of

Research Masters with Training (RMT)

Discipline of Engineering and Energy

College of Science, Health, Engineering and Education

Murdoch University

April 28, 2022

Declaration

I declare that this thesis is my own account of research, except where other sources are acknowledged. This thesis contains as its main content work which has not been previously submitted for a degree at any tertiary educational institution.

Md Rezaul Karim Chowdhury

Signed by: d344b3d0-c868-4f6d-aa01-d4c9c1121358

Acknowledgement

First and foremost, I would like to express my sincere gratitude to Dr. Tania Urmee, my principal supervisor, Dr Jonathan Whale, co-supervisor, Dr Farhad Shahnia, Dr Burak Gunay, co-supervisor for their valuable guidance, encouragement, patience and consistent support throughout my Master's journey at Murdoch University. Their sincere attitude towards research has always inspired me not to overlook details. I wish to extend my special thanks to my external co-supervisor Dr Burak Gunay for his scientific advice and instruction and many insightful discussions and suggestions. I respectfully appreciate all his contributions of time, ideas, and other facilities and logistic supports to make my research experience productive and rewarding.

I am grateful to Murdoch University for the support received through an 'Australian Government Research Training Program Scholarship', thereby giving me the ability to carry out my research without the financial concerns which might otherwise have defeated me.

Last but not the least, my special thanks goes to my wife Zakia Afroz, my son Samin Karim, and my daughter Simeen Karim. This journey would not have been possible without the tremendous support from my family.

Abstract

For promoting energy efficiency practices in the building sector, energy conservation measures are of utmost importance. The energy conservation measures implemented through Energy performance contracting (EPC) are predominantly linked with the performance of baseline building energy models. While occupants are recognized as one of the most important driving factors of energy use in buildings, current research has failed to identify if building occupancy rate can be an influential independent variable to predict baseline energy use in buildings. This research aims to identify the influence of considering occupancy rate as an explanatory variable on the modelling performance of baseline building energy.

Six multivariate machine learning approaches (e.g., linear regression, regression trees, ensembles of trees, Gaussian Process, support vector machine, nonlinear Autoregressive Exogenous model (NARX)) and one univariate (e.g., reformed ten-parameter change point model) inverse modelling approach were employed in the baseline model development process of building heating and cooling energy use and electricity. The specified multivariate baseline modelling approaches were investigated to better isolate the impact of occupancy on building energy performance. NARX outperformed other baseline modelling approaches in terms of model predictive accuracy and data fitting capabilities. On the contrary, the proposed adapted change point model demonstrates the capability of providing operational insight into the case study building.

The hourly fifteen-month worth of energy use data used in baseline models was extracted from the building management system (BMS) server of a real case study building. The prediction period was defined as the most recent six months of the available data representing the COVID lockdown period. The models were trained using the nine-month worth of data that immediately preceded the prediction period. The arrangement of different input parameters selected by a forward

sequential feature selection approach was considered an important step to identify the influence of individual parameters on baseline energy use. The influence of occupancy on the accuracy of baseline models was quantitively evaluated from this analysis. The results show that baseline model performance slightly improves when occupancy data are considered as an explanatory variable. However, occupancy data can significantly influence the performance of a baseline energy use model in an occupant-centric building. The assessment of hourly energy data and associated occupancy data for the case study building indicates the necessity of implementing occupant-centric control strategies to improve its energy performance.

Table of Contents

List of Figures

List of Abbreviations

AHU	Air Handling Unit
AMY	Actual Meteorological Year
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BMS	Building Management System
BPS	Building Performance Simulation
CV(RMSE)	Coefficient of Variation of the Root Mean Square Error
DCV	Demand Controlled Ventilation
ECM	Energy Conservation Measure
EO	Executive Order
EPA	Energy Policy Act
EPC	Energy Performance Contracting
ESCO	Energy Service Company
GHG	Greenhouse Gas
GP	Gaussian Process
GPR	Gaussian Process Regression
IPCC	Intergovernmental Panel on Climate Change
IPMVP	International Performance Measurement and Verification Protocol

M&V	Measurement and Verification
MSE	Mean Square Error
NARX	Nonlinear Autoregressive with External Input
SVM	Support Vector Machine
SVR	Support Vector Regression

Chapter 1

Introduction

1.1. Background

Over the past years, due to climate change, environmental issues have become great concerns and challenges for human beings [1]. This has resulted in the implementation of state-of-the-art actions to monitor and indicate environmental problems as well as devising new solutions to limit environmental emissions [2]. Sustainability practices have become necessary to control environmental emissions in all aspects of human life [3]. In response to climate change, the evaluation of energy demand and energy use in buildings has become increasingly important [4]. In recent reports by IPCC [5, 6] buildings have been recognized as a critical field of action for a number of reasons. Research shows that building operations account for 28% of global energy-related GHG emissions [7]. Since buildings are responsible for a massive amount of current GHG emissions, they also have significant potential to reduce GHG emissions through improved energy efficiency practices. Therefore, to promote energy efficiency practices in the building sector, energy conservation measures are in receipt of the greatest importance, and this has been reflected in the government policy practices of many nations. For example, the U.S. government passed the Energy Policy Act (EPA) of 2005 and Executive Order (EO) 13,423, requiring that 15% of the total number of existing buildings be retrofitted to improve energy efficiency by 2020 compared with the 2003 baseline. Around 30 billion US dollars have been allocated to conduct energy efficiency retrofit of existing buildings and facilities [8].

Incentivized by the policies, providing energy efficiency services through energy service companies (ESCOs) has become a thriving market in the last decade [9]. The energy conservation measures are implemented through Energy Performance Contracting (EPC), which is a recognized financing package provided by ESCOs. The profit (or the payment to ESCOs) of an EPC is predominantly linked with energy cost savings because of the execution of energy conservation measures (ECMs). The energy savings may be determined from the difference between how much energy the building consumed after the retrofit action and how much it would have consumed if no retrofit action was taken place. While the former energy consumption amount can be obtained from utility meters, the latter, referred to as the energy use "baseline", is not measurable but can only be obtained from prediction. The accuracy of baseline energy use can significantly impact energy saving estimations and the calculation of the expected payback period. Additionally, it influences decisions on retrofit measures and the development of the building retrofit market.

The whole process of predicting baseline energy use and assessing energy saving is called "measurement and verification" (M&V) [10]. The mechanism of an M&V approach is to first monitor the energy use of buildings, then develop mathematical or data-driven models trained by observed data, and finally predict baseline energy use based on the developed models. This process seeks to quantify energy savings and uncertainty levels resulting from retrofit practices. For the M&V process, ESCOs follow the International Performance Measurement and Verification Protocol (IPMVP) [11, 12] and American Society of Heating, Refrigerating and Air-Conditioning Engineers' (ASHRAE's) Guideline 14: Measurement of Energy and Demand Savings [13]. The IPMVP offers uncertainty estimation techniques based on measured data and the baseline modelling approach being used to develop the model.

The uncertainty of M&V models holds prime importance during model development since this is interrelated with the accuracy and reliability of the prediction results. It provides the stakeholders (e.g., ESCOs, building owners, facility managers) the information of investment risk for the planned retrofit measures and plays a vital role in their decision making. For example, if there is an opportunity of saving 30% energy after implementing an ECM, but the uncertainty exceeds 30%, it is then very risky to invest in this retrofit project.

The occupancy rate is considered as the key uncertainty factor of M&V models of building energy performance. This is because the occupants in buildings influence energy use in three different ways: (1) sensible and latent heat gains from the occupants, (2) occupants' need of thermal comfort, visual comfort, and indoor air quality, and (3) occupant behavior and interactions with building systems and controls [14-16]. In addition, evidence shows that occupancy rate increases after an energy retrofit because of a number of desirable benefits offered by these buildings, such as lower utility bills, better indoor environment, and higher social reputation [17-20]. Therefore, if the occupancy rate is changed at any stage of the energy retrofit, the baseline of energy use should be adjusted accordingly.

1.2. Rationale

Occupants are recognized as one of the six driving factors of energy use in buildings [21] and as a foundation of uncertainty and operational decision-making process with a significant impact on building performance simulation (BPS) [15]. Apart from occupants being identified as a source of internal heat gains, their interaction with different building systems demonstrates occupants' influence on a building's heating and cooling energy use as well as electrical load [22, 23].

While the research community widely recognizes the positive correlation between occupancy rate and energy use in buildings [14, 24-31], research has not yet reached the phase that attempts to identify the impact of occupancy rate on baseline modelling performance. The contemporary research is yet to reveal whether building occupancy rate can be an influential independent variable to predict baseline energy use. Also, it could be worthwhile to reveal if a typical building where intelligent occupancy-based sensor technologies are not in place performs the same as before in terms of energy use when the occupancy pattern abruptly changes.

1.3. Research Gaps and Research Question

A good number of past studies, e.g., [32-40] concentrated on developing building energy baseline models using different modelling approaches such as change point, Gaussian Process, Gaussian Mixture, Artificial Neural Network, Regression Trees, Linear regression, and Degree day-based regression. Also, some special types of baseline modelling approaches such as the mean-week model, day-time-temperature model, LBNL (Lawrence Berkeley National Laboratory) model, proprietary model were used by some studies [34, 35, 39]. Variations were observed in the frequency of data such as hourly, daily, weekly, monthly, and the explanatory variables (e.g., weather data, time of day, and day of the week used by the studies). The occupancy data were rarely considered by any past studies. Liang, et al. [41] proposed an approach to quantitatively evaluate how the accuracy of energy baseline models is improved by including the occupancy factor. Three types of explanatory variables, e.g., outdoor air temperature, time of the week, and occupancy count data, were used in this study to perform this quantitative analysis. On the other hand, Heo and Zavala [38] used four types of explanatory variables e.g., outdoor air temperature,

AHU supply air temperature, outdoor relative humidity and occupancy data to predict baseline chilled-water energy.

Despite the use of occupancy data in the baseline energy predictions in a limited scope in the recent past, the influence of independent explanatory variables including and excluding occupancy data and the position of occupancy data in the sequential selection process of explanatory variables, are yet to be realised. Therefore, an extensive analysis is required taking into account a wide variety of weather data together with time-of-day information and occupancy data. According to ASHRAE Guideline 14 [13], the use of more detailed energy use data (such as daily or hourly) may decrease the uncertainty in the estimated energy savings. However, this necessitates the use of more independent variables to model the energy use.

Taking into account the identified research gaps and the potential areas for further improvements, this research aims to address the following research question:

- Can a baseline energy model where occupancy data were used as an explanatory variable better capture the changes in energy use baseline?

1.4. Aim and Objective of the Research

To address the aforementioned research questions, this research proposes a methodology to quantitatively evaluate the influence of occupancy rate on the baseline building energy. In this instance, occupancy data is considered as an explanatory variable for the baseline building energy models. Also, this research aims to investigate if a typical office building performs the same in the COVID period as before in terms of energy use when the occupancy pattern abruptly changes.

In order to fulfill the aim of this research, the following objective needs to be addressed:

- Investigate different baseline energy modelling approaches to better isolate the impact of occupancy on building energy performance.

1.5. A brief Overview of Research Methodology

The case study building selected for this research is the six floor Robertson Hall at Carleton University. Since this building is occupied by administrative staff only, its occupancy level resembles to a typical "non-academic" office, which is intended for the generalizability of the findings of this study. The building was selected because of the access to necessary data, e.g., energy consumption, wifi data, and a stable occupancy pattern in order to fulfil the research objectives of this research study.

During baseline model development, five different types of timeseries data were used. These data were extracted from multiple sources. Subsequently, data cleaning and pre-processing tasks were performed before using them in the models (see Section 3.3.1.2). For instance, the hourly meter data (e.g., electricity, heating, and cooling energy use) were extracted from the Schneider meter network. The Wi-Fi device count data representing occupancy count were gathered from the Cisco CMS IT network. Note that Cisco CMS sends daily device count reports, which was automatically saved to the researcher's cloud storage and appended to previous data. For the same period, hourly weather data were extracted from the actual meteorological year (AMY) weather files [42] based on the location of the case study building. These datasets and data pre-processing procedures are detailed in Section 3.3.1.

1.6. Research Significance

The methodology developed and applied in this study comprises the first step in establishing a general approach to evaluate the predictive accuracy of whole building level baseline models. It is worth noting that the predictive performance of baseline models is considered as the most critical component of uncertainty in the M&V practices of whole building level energy saving calculations. It provides the stakeholders (e.g., ESCOs, building owners, facility managers) the information of investment risk, which is important in decision making for executing energy efficiency retrofit measures.

At present, due to the presence of advanced sensor technologies in commercial buildings 15-min interval or hourly occupancy data has become available. These data offer a new-fangled opportunity for the building practitioners to rigorously investigate the impact of occupancy on the accuracy of baseline prediction. The results obtained can assist the building practitioners to understand the influence of occupancy on energy use, improve baseline energy predictive performance by considering the occupancy factor, lessen uncertainties of M&V practices and facilitate financial settlement of energy efficiency retrofit.

1.7. Structure of the Thesis

This thesis has been prepared as per the guideline defined by Murdoch University's Graduate Research Office.

Chapter 1 includes a general background, rationale, research gaps and research questions, research aim and objectives, and significance of this research.

Chapter 2 comprises a review of related literature. This Chapter provides a general background of Measurement and Verification (M&V) practices, different baseline modelling approaches, statistical performance metrics used to quantify the baseline modelling performance. Also, a survey of baseline building energy modelling applications is presented in this Chapter.

Chapter 3 presents an overview of the case study building. A wide variety of baseline modelling approaches considered in this study are detailed. Also, the methodology that was followed to develop and evaluate the models are described.

Chapter 4 discusses modelling results and presents a comparative analysis of modelling approaches in the context of the selection of explanatory variables. This Chapter emphasizes the research findings concerning the influence of occupancy level on building energy use.

Chapter 5 summarizes the findings of this research and draws a conclusion for the entire research study. This Chapter also addresses the limitations of this study and provides future directions accordingly.

Chapter 2

Review of Literature

2.1. General Background

2.1.1. The measurement and verification process (M&V) of building energy use

The whole process of predicting baseline and assessing energy saving is referred to as "measurement and verification" (M&V) [11-13]. The terms M&V denote two components [43]:

(1) measuring (or estimating) actual savings; and

(2) verifying the proper installation and the measure's potential to generate savings.

The second component involves accurately defining the baseline conditions and (b) installing proper equipment/systems which have the potential to generate the predicted savings and performance according to specification. The general approach to verifying baseline and post-installation conditions involves inspections, spot measurement tests, or commissioning activities. Commissioning is the process of documenting and verifying the performance of energy systems that reflects the operational consistency and reliability of the systems according to the design intent.

There are a set of standards that establish the guidelines to comply with different types of projects and contribute to M&V practices by participating in specific developments depending on project type. The most recognized standards applicable to building industries and construction facilities are IPMVP [11, 12] and ASHRAE's Guideline 14 [13]. These standards are discussed in detail below.

International performance measurement and Verification Protocol (IPMVP)

This protocol maintained by the non-profit Efficiency Valuation Organization (EVO) is used in more than 60 countries [44]. IPMVP defines common terminology and the key steps in implementing a robust M&V process. It includes guidance on current best practice retrofit options and verifying the results through quantifying energy savings and uncertainty levels resulting from retrofit practices. This process compares before and after energy consumption or demand on a consistent basis, using the following general M&V equation:

Energy savings = (Baseline Period Energy Use – Reporting Period Energy Use) ± Adjustments (Eq. 1)

IPMVP offers four distinct M&V options to cover a wide range of projects:

- Option A: Retrofit Isolation (Key Parameter Measurement),
- Option B: Retrofit Isolation (All Parameter Measurement),
- Option C: Whole Facility, and
- Option D: Calibrated Simulation

M&V practitioners select the best option for the individual projects taking into account project budget, the number of independent variables to be monitored, frequency of measurement and reporting, length of the baseline and reporting periods, and sample size, if all equipment is not measured. Options A and B separate the retrofit with a project boundary that covers the affected equipment [45]. Option C is a whole-building approach and applicable in cases where the savings are greater than 10% of the total site energy consumption. Option D consists of a calibrated simulation of the energy systems. This approach is beneficial in situations with no baseline data. These options are elaborated below:

Option (A): Energy savings are estimated by field measurement of the key performance parameter(s), which define the energy use of a particular system after the retrofit action. Typical applications may include a lighting retrofit, where the power draw can be measured periodically, and hours of operation can be estimated. In this example, lighting hours of operation are not selected for the field measurement; instead, it is estimated based on historical data, manufacturer's specifications, or engineering judgement.

Option (B): Energy savings are calculated by field measurement of all performance parameters, which define the energy use of a particular system after the retrofit action. Typical applications may include a lighting retrofit where both power draw and hours of operation are monitored and recorded.

Option (C): This option considers energy uses in the whole facility or sub-facility to calculate energy savings. In this option, different baseline modeling techniques (such as regression/inverse modelling) are used to estimate baseline energy uses considering a number of explanatory variables. Under this option, a number of steps are suggested, such as pre-retrofit data process, post-retrofit data analysis, and data quality control procedure. Typical examples may include energy measurement of a facility where several retrofit measures were implemented or where the retrofit measure is expected to affect overall building performance.

Option (D): Computer simulation software is used to predict energy uses in a facility. The simulation model is calibrated to predict energy uses or energy demand that reasonably matches actual energy consumption utility data from the real building. Typical examples may include energy measurement of a facility where several retrofit measures were implemented, but no historical energy data is available.

Despite holistic measures, this option can provide an accurate estimation of the energy savings at the beginning of the project, which reduces the investment risk of energy service companies. Additionally, simulation models can contribute to the development of cost efficient energy conservation measures. According to Ruiz, et al. [46], accurate simulation modeling and calibration with measured energy data are the major challenges associated with option D. The simulation results often do not tie with the real energy consumptions. Therefore, option D is rarely carried out within the M&V process, and a holistic energy savings performance contract is uncommon due to the high risk for both parties.

ASHRAE guideline 14

This Guideline specifies the criteria for three particular building approaches. These are the whole building approach, retrofit isolation approach, and calibrated simulation. The criteria for these approaches are detailed below:

Whole Building Approach: The whole building approach, also called main meter approach, includes procedures that verify the performance of the retrofits for those projects where whole building pre- and post-retrofit data are available to determine the energy savings. This approach is recognized as the most appropriate one when the total building performance is to be calculated rather than the performance of specific retrofits. Examples of the whole building approach include the day adjusted model, two, three, four, or five-parameter model, change point model, multi-variate model.

Retrofit Isolation Approach: The retrofit isolation approach is intended for retrofits where the end-use capacity, demand, or power level can be measured during the baseline period, and the energy use of the building equipment or subsystem can be measured post-installation for a short-term period or continuously over time.

Calibrated Simulation: This approach is an appropriate method to consider when one or more of the following conditions are present:

— Either pre-retrofit or post-retrofit whole-building metered electrical data are not available;

— Savings cannot be readily determined using before-after measurements;

— Measures interact with other building systems and retrofit isolation methods are not readily feasible;

— Only whole-building energy use data are available but savings from individual retrofits are desired;

— Baseline adjustment needs.

2.1.2. Baseline modelling approaches

While baseline modelling approaches are primarily separated into physics-based and data-driven or inverse, the latter is widely used in M&V practices and ongoing commissioning of building performance [32, 38, 40, 47-51].

Physics-based modelling

The application of physics-based models for predicting baseline energy use has been limited by its shortcomings. The development of physics-based models necessitates detailed building construction information that can be challenging to acquire [52]. Moreover, the model development process involves high engineering costs. Furthermore, these models experience model calibration before being used for simulating energy retrofits. The model development and calibration process is time consuming as well as labour intensive, even with the application of modern calibration techniques [53]. Additionally, traditional deterministic methods that are not based on data experience the problem related to their adaptability, as the results obtained are

usually valid for a specific building under analysis [54, 55]. On the contrary, the major advantage of this modelling approach is that it can be applied to systems in the design phase prior to construction [56].

Data-driven or inverse modelling

These baseline modelling approaches received popularity in recent years due to their capability to capture a large amount of information for evaluating the applied retrofit measures and predicting the energy savings potential of new retrofit actions from the measured data. These types of models derive a relationship between a set of inputs (e.g., weather parameters) and outputs (e.g., energy consumption) without explicit or detailed knowledge of its physical behaviour [57]. In practice, data-driven models trained by actual building energy consumption data and associated weather and other explanatory parameter data can provide a reliable estimation of energy use and have been widely adopted for energy savings analysis of retrofit, M&V, and ongoing commissioning of building performance [32, 38].

2.1.3. Baseline model performance metrics

The statistical performance metrics that are considered in the baseline model evaluation process are collectively referred to as 'goodness-of-fit' metrics [34]. To determine how well a mathematical model describes the variability in measured data, the ASHRAE Guideline 14 [13] suggests using three types of performance indices - Coefficient of variation of the standard deviation (CVSTD), coefficient of variation of the root mean square error (CVRMSE), and normalized mean bias error (nMBE).

The CV(RMSE), which is also known as normalized root squared error (nRMSE) is one of the most recognised baseline model performance metrics and referenced in ASHRAE Guideline 14

[13]. This metric quantifies the typical size of the error relative to the mean of the data and is expressed by Eq. (1).

$$CV(RMSE) = \left(\frac{RMSE}{\bar{y}}\right) \times 100 \quad (1)$$

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

where, $y_i$is the measured energy use, $\hat{y}_i$is the predicted energy use, $\bar{y}$ is the mean of the measured energy use, N is the number of data points.

For instance, a CV(RMSE) value of 30% means RMSE value is 0.30 of the mean of the measured data. This performance metric has been widely used in past studies, e.g., [32, 40, 41, 48, 58], for evaluating the predictive accuracy of baseline models and quantify the uncertainty in the model.

The NMBE is referred to as the ratio of the difference between actual energy usage and model predicted energy usage to the degrees of freedom and average energy usage by the facility and is expressed by Eq. (2).

$$nMBE = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)}{N \times \bar{y}} \times 100 \quad (2)$$

Similar to the value of CV-RMSE, a lower NMBE value indicates a better goodness-of-fit for a regression-based baseline model. According to ASHRAE Guideline 14 [13], the nMBE value of a baseline energy model using monthly and hourly data should be equal to or less than 5% and 10%, respectively. This metric was used in a number of past studies, e.g., [32, 40, 48, 58] to evaluate the baseline model performance.

The coefficient of determination, R^2, expressed by Eq. (3), quantifies the proportion of variation in the dependent variable that is explained by a regression model. The R^2 value can be between

zero and one. An R^2 value equal to one indicates a perfect fit between the measured data and the regression model. This metric was used in past studies, e.g., [32, 40, 48] to measure the data fitting capability of models.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (3)$$

2.2. Review of baseline building energy modelling applications

To date, a good number of literature, e.g., [9, 32-40, 48, 50] concentrated on predicting baseline energy use using different modelling approaches, e.g., change point, Gaussian Process, Gaussian Mixture, Artificial Neural Network, Regression Trees, Linear regression, and Degree day-based regression. Also, some particular types of modelling approaches such as the mean-week model, day-time-temperature model, LBNL (Lawrence Berkeley National Laboratory) model, proprietary model, cluster inverse model were used by some studies [34, 35, 39, 48]. While M&V practices largely rely on traditional relatively simple regression modelling approaches, with the advent of advanced metering infrastructure in commercial buildings and cutting-edge energy analytics methods, state-of-the-art baseline modelling approaches are getting popular. Consequently, the research community is taking the challenge of utilizing high-frequency interval data, e.g., daily and hourly data, in the baseline models. On the contrary, Mathieu, et al. [36] used 15-min interval data in a regression-based model for predicting baseline electric load in buildings. To capture the building load shapes that share some features such as morning start-up, morning ramp-up, evening setback, near-base load, evening shoulder, peak load and can vary from one 15-min interval to the next, that high-frequency data were used in the model. The criteria for evaluating a baseline model for hourly and monthly data are well-defined in ASHRAE Guideline 14.

The survey of independent explanatory variables shows that different weather parameters, (e.g., temperature (T), relative humidity (RH), wind speed (WS), wind direction (WD), solar radiation (SR)), the information of time of day, and day of week, and occupancy sensing data were used in the models to characterize baseline energy use (e.g., total electrical or thermal load) in buildings. A summary of surveyed related studies on baseline building energy models is presented in Table 1. Table 1 shows that outdoor temperature is the most common explanatory variable used in all surveyed existing studies, followed by the day of week, and time of day. On the other hand, other weather parameters such as RH, WS, WD and SR were rarely used in different studies. Also, occupancy data was used in only two surveyed studies. Therefore, it could be worthwhile considering all explanatory variables in the baseline energy model and evaluating the performance of these explanatory variables on the model performance. Sorting out an optimum number of input parameters for the individual baseline models creates an untapped opportunity to perform a comparative study among these modelling techniques.

Table 1. Summary of surveyed studies on baseline building energy models

Research studies	Baseline model	Frequency of data	Respondent	Weather parameters					Time of day	Day of week	Occupancy level	Other
				T	RH	WS	WD	SR				
Zhang, et al. [32]	• Change point • Gaussian Process • Gaussian Mixture • Artificial neural network	• Hourly • Daily	Hot water energy use	√								
Burak Gunay, et al. [33]	• Change point • Regression tree • Artificial neural network	Hourly	• Heating load intensity • Cooling load intensity	√		√		√	√			
Afroz, et al. [59]	• Change point • Support Vector Machine (SVM) • Linear Regression • Gaussian Process • Regression trees • Ensembles of trees • Neural Network • HCC/CDD based linear regression • Naïve approach	Hourly	• Heating load intensity • Cooling load intensity	√	√	√	√	√	√			
Granderson and Price [34]	• Change point • Mean-week model • Day-time-temperature model • Proprietary model • LBNL (Lawrence Berkeley National Laboratory) model	• Daily • Weekly • Monthly	Electricity consumption	√					√	√		
Liang, et al. [35]	• MW (mean-week) model • LBNL model • LBNL model including occupancy variable	Hourly	Electricity consumption	√					√	√	√	
Walter, et al. [9]	• Linear regression	Hourly	Electric load	√					√	√		
Mathieu, et al. [36]	Linear regression	15-min	Electricity consumption	√					√	√		
Golden, et al. [37]	• Change point • Degree day	Daily	Electricity consumption	√								
Heo and Zavala [38]	• Gaussian process	Daily	Chilled-water energy	√	√						√	AHU supply air temperature
Touzani, et al. [39]	• Time-of-week-and-temperature model	• Hourly • Daily	Electricity consumption	√						√		

	• Bayesian additive regression trees — Hourly — Daily • Daily linear model								
Carpenter, et al. [40]	• Change point • Gaussian process	• Monthly	• Electricity consumption • Natural gas consumption	√					
Srivastav, et al. [50]	• Gaussian Mixture • Multi-linear regression	Daily	• Electricity consumption • Cooling energy consumption	√	√	√			
Ko, et al. [48]	• Change point • Cluster inverse model	Daily	• Electricity consumption • Gas consumption	√			√		Holiday schedule

This research considers different baseline modelling approaches, and their performances are evaluated to better isolate the impact of occupancy on building energy performance (Section 4.7). This study proposes an adapted change point model that comes with the capability of providing building operational insights. Also, different multi-variate baseline modelling approaches (such as Neural Network, Gaussian Process, Ensembles of Trees, linear regression, regression trees) were used in this study. A range of performance evaluation techniques (such as optimal input parameters, optimum network size, hyperparameter tuning) was applied to these models. Evaluating the performance of a range of modelling techniques that have been found promising in the existing studies, employing an adapted change point model and making use of a range of performance evaluation techniques added originality to this research. The evaluation results point out the necessity of an accurate modelling practice in real-world applications (Section 4.4 - Section 4.6). The description of individual models and independent explanatory variables used in the study is presented in the following Chapter.

Chapter 3

An Overview of the Case Study Building, Corresponding Dataset and Methodological Approaches

3.1. The case study building

The Robertson Hall, Carleton University, is considered as a case study building for this research and was built in 1969 as a multi-storey building comprising of six floors with a gross floor area of 9027.48 m^2 (Figure 1). This building represents an institutional building that houses the University archives, admissions services, and administration facilities. It operates from 9 am to 5 pm, Monday through Friday except on government holidays. The latitude and longitude of this building are 45° 22' 59.16" N and -75° 41' 51.36" W, respectively.

In 2011, the Robertson Hall underwent a M&V process using the Retrofit Isolation (RI) method: key parameters measured (Option A in IMPVP[1]) with the intent of obtaining perceived benefits such as increased energy savings, effective monitoring of CO_2 emissions, executing effective planning, etc. [60]. The energy consumed by building subsystems was isolated and analyzed independently from other facility systems. As per building energy audit report 2011 [61], utility costs are distributed as follows: 76% for electricity costs; 17% for steam costs and 13% for water costs (Figure 2). With regard to energy usage, 71% of energy use is for electricity and 29% for heating.

[1] A definition of IMPVP standard for the M&V practices can be found in Chapter 2.

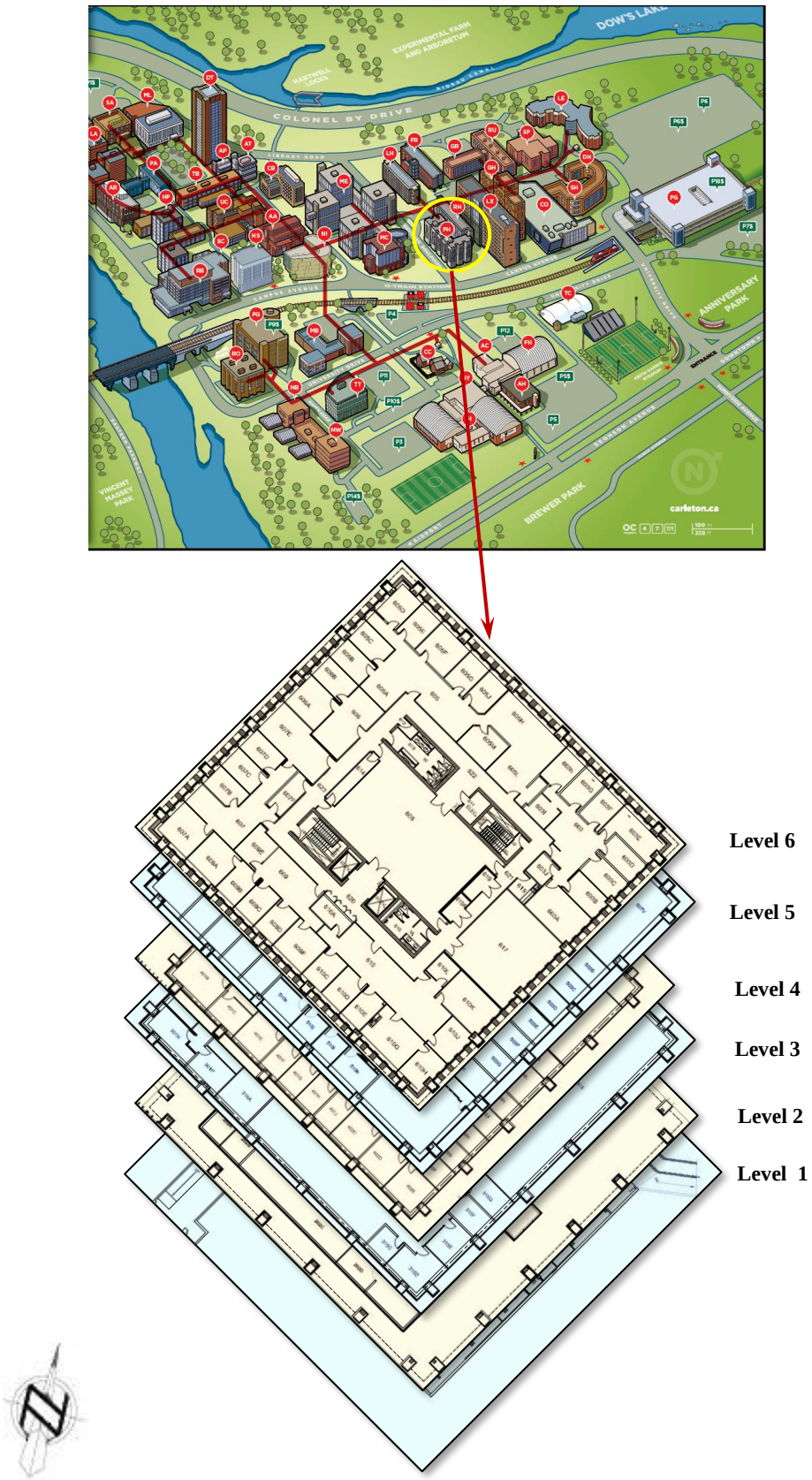


Figure 1. The floor plans of the case study building

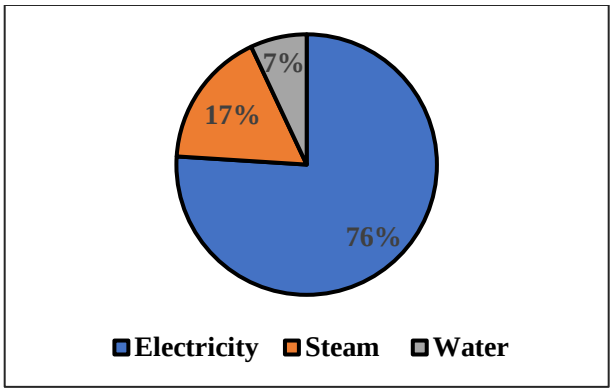


Figure 2. Building utility cost breakdown

Examples of saving measures include temperature setback, ventilation reduction, boiler plant replacements and lighting controls. Subsystem energy utility usage before and after implementing Cost-Saving Measures (CSMs) was obtained by analysing physical data collected using various data acquisition tools and procedures including the building automation system data acquisition feature, portable data acquisition equipment, automatic metering infrastructure, and maintenance logs.

A utility baseline model was developed for the selected base year. The baseline model accounted for variations in utility billing days, occupancy, and weather. However, it was recommended to have further adjustments in future for changes in building operational and equipment performance.

This study looks at the changes in building operational and equipment performance due to COVID-19, where the baseline period spans from April to December 2019. The baseline model was used to predict building energy use if COVID had not occurred. The predicted data were compared with the measured data for the pre-COVID (January 1 – March 23, 2020) and COVID (March 24 – July 4, 2020) periods. The electricity, heating, and cooling meter data used in the baseline model were extracted from a Schneider meter network. The Wi-Fi device count data were gathered from a Cisco CMS IT network. Note that Cisco CMS sends daily device count reports, which were

automatically saved to the researcher's cloud storage and appended to previous data. Details on data collection process are provided in Section 3.3.1.1.

3.2. Baseline modelling approaches

The traditional change point is recognized as the most widely used baseline model because of its simplicity and a great deal of success with monthly data [37, 40, 48]. However, the traditional change point model has been found inappropriate for analysing the daily energy consumption in an existing study [48]. On the contrary, because of the availability of high-frequency energy data and advanced data analytics techniques, at present, comparatively more sophisticated modelling techniques are taking the place of traditional baseline models. These sophisticated modelling approaches are getting popular because of their ability to better capture the building performance and this, in turn, provides high predictive accuracy. In different literature (Table 1), a variety of advanced modelling approaches such as Neural Network, Gaussian Process, Gaussian Mixture, Ensembles of Trees, linear regression, regression trees were used to estimate hourly, daily, and monthly energy usage and their predictive performances were evaluated in some cases. However, implementation of these advanced models requires high-level software skills, creating a challenge for the building practitioners.

Therefore, considering the current challenges and the issue of employing traditional baseline models with high-frequency data (such as hourly or daily), this research proposes an extensive range of baseline modelling approaches. The performance of these modelling approaches was evaluated to better isolate the impact of occupancy on building energy performance. A range of aspects (such as optimal input parameters, optimum network size, model's predictive accuracy taking into account occupancy data) as detailed in the following sections were considered in this study. (Figure 3). Also, an adapted change-point model is proposed here that outweighs the

obstacles of traditional change-point models. A brief description of these modelling approaches is presented in the following sub-section:

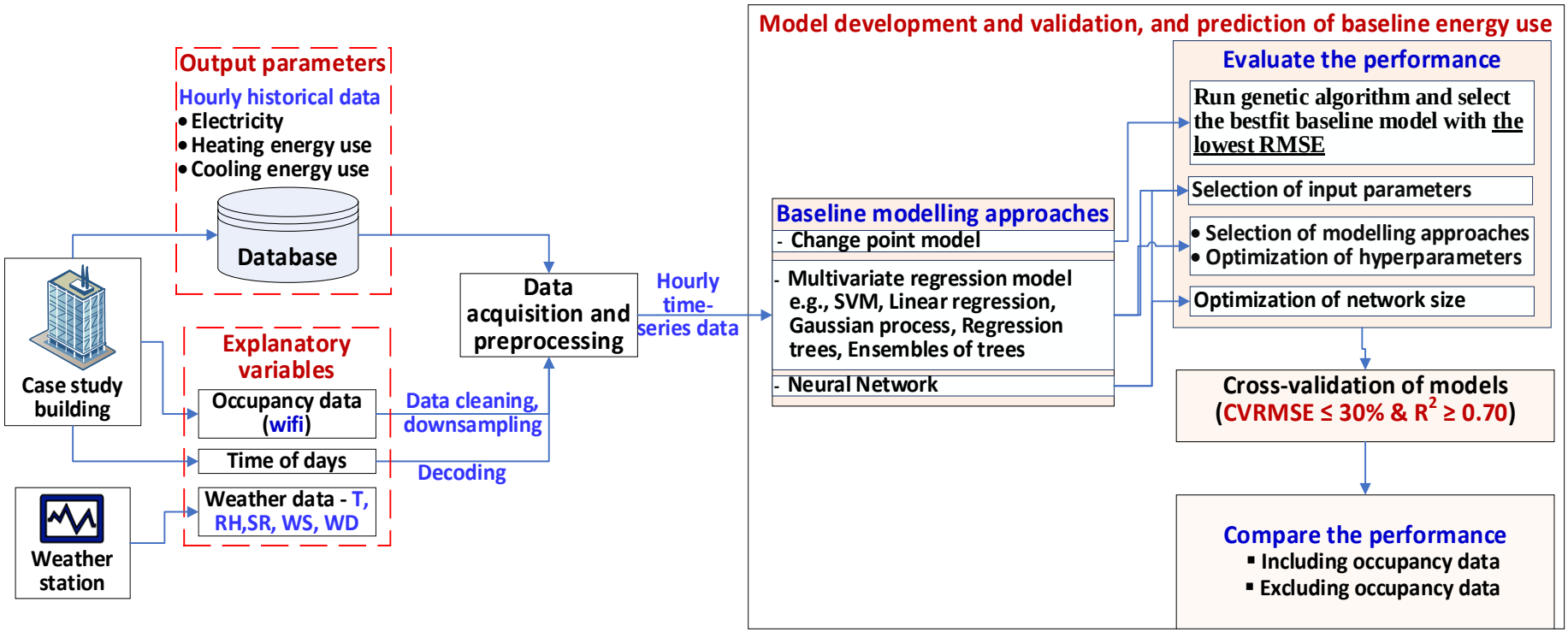


Figure 3. An overview of the methodology

3.2.1. Change-point model

In the three-parameter change point model, the changepoint temperature, and a slope of the line are fitted to the points above or below the change-point. Figure 4 shows a commonly used three-parameter, change point model applied to heating, cooling, and electricity. This model is also known as a "variable-based degree-day model" [62]. Please note Figure 4(c) represents a common instance of change point model for electricity consumption where cooling systems i.e., chillers share a major part of electricity and heating systems predominantly depend on natural gas.

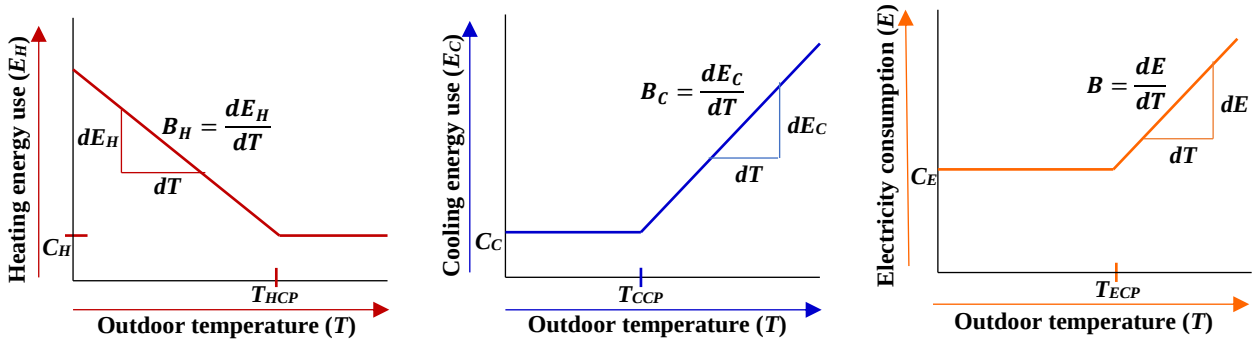


Figure 4. Three-parameter change-point model (a) Heating, (b) Cooling, (c) Electricity

In functional form, the change point models for heating, cooling, and electricity can be expressed by

Heating energy use: $E_H = C_H + B_H (T - T_{HCP})^-$ (1)
Cooling energy use: $E_C = C_C + B_C (T - T_{CCP})^+$ (2)
Electricity consumption: $E = C_E + B(T - T_{ECP})^+$ (3)

where C_H, C_C and C_E indicate the y-intercept values at the change point of the models for heating, cooling, and electricity, respectively. B_H, B_C, and B indicate the slope of the models for heating, cooling, and electricity, respectively. T_{HCP}, T_{CCP} and T_{ECP} indicate the change point temperatures of the models for heating, cooling, and electricity, respectively.

The superscript plus (+) indicates that only positive values of the parenthetical expression are considered, while the minus (-) annotation indicates the opposite [62].

This study presents an adapted change point model that simultaneously solves for ten parameters [63]. Note that this model was adapted from [63]. Figure 5 presents schematics for change point models for four operating conditions and can be expressed by the following regression equations:

Heating energy use during occupied mode, $E_{1,H} = x_{6,H} + x_{5,H}\,(T - x_{7,H})^{-}$ (4)

Heating energy use during unoccupied mode, $E_{2,H} = x_{9,H} + x_{8,H}\,(T - x_{10,H})^{-}$ (5)

Cooling energy use during occupied mode, $E_{1,C} = x_{6,C} + x_{5,C}\,(T - x_{7,C})^{+}$ (6)

Cooling energy use during unoccupied mode, $E_{2,C} = x_{9,C} + x_{8,C}\,(T - x_{10,C})^{+}$ (7)

Electricity use during occupied mode, $E_1 = x_{6,E} + x_{5,E}\,(T - x_{7,E})^{+}$ (8)

Electricity use during unoccupied mode, $E_2 = x_{9,E} + x_{8,E}\,(T - x_{10,E})^{+}$ (9)

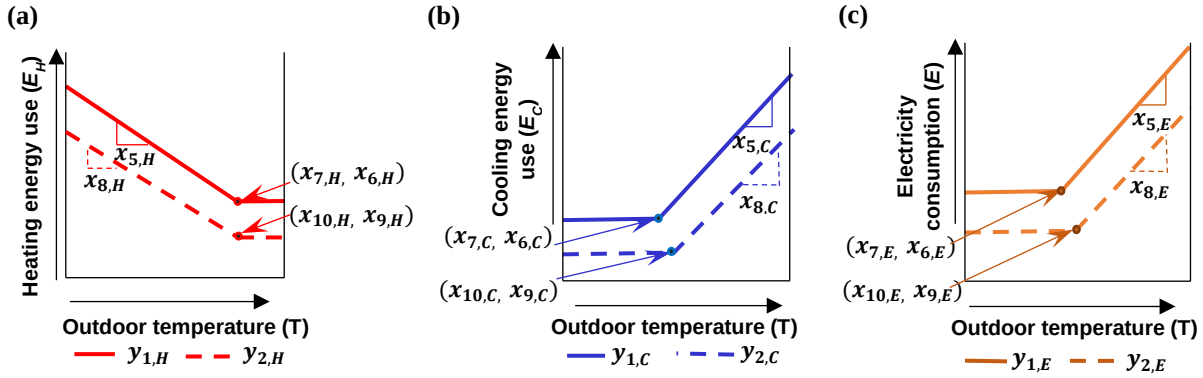


Figure 5. A schematic illustrating the proposed change point models for (a) Heating, (b) Cooling, (c) Electricity (adopted from Ref. [63])

The parameters ($x_{1,E}$ to $x_{10,E}, x_{1,H}$ to $x_{10,H}$, and $x_{1,C}$ to $x_{10,C}$) defined by the proposed adapted change point models represent ten pieces of important information about the operating status of the electricity, heating and cooling systems within the building during occupied and unoccupied modes (Table 2). The first four parameters ($x_{1,E}$ to $x_{4,E}, x_{1,H}$ to $x_{4,H}$, and $x_{1,C}$ to $x_{4,C}$) of the change point models provide information about the duration of building occupied and unoccupied modes for electricity consumption, heating and cooling energy uses. The parameters

$x_{5,E}, x_{8,E}, x_{5,H}, x_{8,H}, x_{5,C}$ and $x_{8,C}$, representing the slopes of the models, indicate the rate of change in the electricity, heating and cooling energy uses in response to changes in outdoor temperature, *T*. The change point temperatures $x_{7,E}$, and $x_{10,E}$ for electricity indicate the minimum temperature above which the electricity profile during occupied and unoccupied modes maintains a linear relationship with the outdoor temperature. The parameters $x_{7,H}$ and $x_{10,H}$ represent the change point temperatures for heating and indicate the maximum temperature below which the heating energy use profile during occupied and unoccupied modes maintains a linear relationship with the outdoor temperature. Similarly, the change point temperatures $x_{7,C}$, and $x_{10,C}$ for cooling indicate the minimum temperature above which the cooling energy use profile during occupied and unoccupied modes maintains a linear relationship with the outdoor temperature. The y-intercept values $x_{6,E}, x_{9,E}, x_{6,H}, x_{9,H}, x_{6,C}$ and $x_{9,C}$ indicate the minimum expected electricity consumption, heating and cooling energy uses during occupied and unoccupied modes independent of the outdoor temperature.

Table 2. The list of ten parameters defined by the change models for electricity, heating and cooling energy uses

Parameter name	***Denotation***
$\mathbf{x_{1,E}, x_{1,H}, x_{1,C}}$	The scheduled start time of occupied mode
$\mathbf{x_{2,E}, x_{2,H}, x_{2,C}}$	The scheduled stop time of occupied mode
$\mathbf{x_{3,E}, x_{3,H}, x_{3,C}}$	The operational status on Saturday
$\mathbf{x_{4,E}, x_{4,H}, x_{4,C}}$	The operational status on Sunday
$\mathbf{x_{5,E}, x_{5,H}, x_{5,C}}$	The slope of the model for occupied mode
$\mathbf{x_{6,E}, x_{6,H}, x_{6,C}}$	The y-intercept value at the change point of the model for occupied mode
$\mathbf{x_{7,E}, x_{7,H}, x_{7,C}}$	The change point temperature of the model for occupied mode
$\mathbf{x_{8,E}, x_{8,H}, x_{8,C}}$	The slope of the model for unoccupied mode
$\mathbf{x_{9,E}, x_{9,H}, x_{9,C}}$	The y-intercept value at the change point of the model for unoccupied mode
$\mathbf{x_{10,E}, x_{10,H}, x_{10,C}}$	The change point temperature of the model for unoccupied mode

Note: E, H and C represent information on the operating status of electricity, heating and cooling systems, respectively

3.2.2. Support vector machine (SVM)

SVM is regarded as a robust learning algorithm for solving non-linear problems [8]. This algorithm is used to find an optimal hyperplane that separates the classes with a maximum margin [34]. If the SVM is trained to predict a time series or real numbers, it is called support vector regression (SVR). SVR uses the same principles as SVM [64]. In SVRs, nonlinear mapping of input data in a higher-dimensional feature space is done with kernel functions. These kernel functions can be polynomials, sigmoidal functions (neural net activation function), and radial basis functions (Gaussian distributions). For each input parameter vector (X) and its corresponding output vector (Y), SVR relates the inputs and outputs using Eq. (10) [65]

$$Y = W.\varphi(X) + b \tag{10}$$

where $\varphi(X)$ function non-linearly maps X to a higher dimensional feature space; W represents the weight vector, and b represents the bias, which are dependent on the selected kernel function. The kernel function quantifies the similarity of two observations [66].

3.2.3. Neural Network

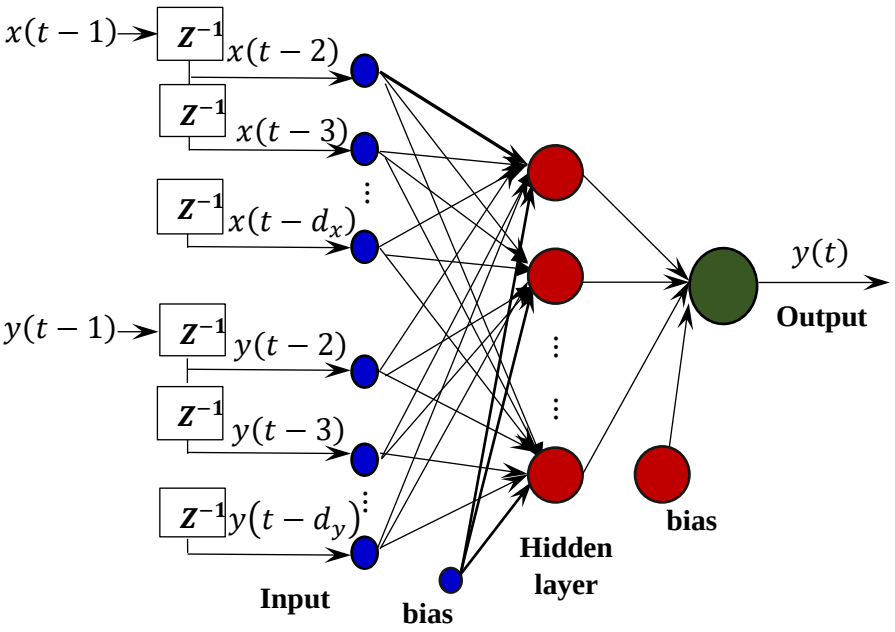
NARX neural network is regarded as a dynamic recurrent neural network that encloses several layers with feedback connections [67]. This neural network allows a delay line on the inputs, and the outputs feed back to the input by another delay line. Therefore, this network has two inputs: one is an external input, and the other is a feedback connection from the network output. The mathematical expression for the output of the NARX network in the training process can be represented by the following equation:

$$y(t) = f\left(y(t-1), y(t-2), \dots, y\left(t-d_y\right), x(t-1), x(t-2) \dots, x(t-d_x)\right) \tag{11}$$

where $y(t)$ the next value of the dependent output signal is regressed on previous values of the output signal and previous values of an independent (exogenous) input signal and f is a nonlinear

function approximated by a Multi-Layer Perceptron. The d_y and d_x terms represent the number of time delays for output and input time series, respectively, while $x(t)$ and $y(t)$ characterize the input and output of the model at time step t respectively.

(a)



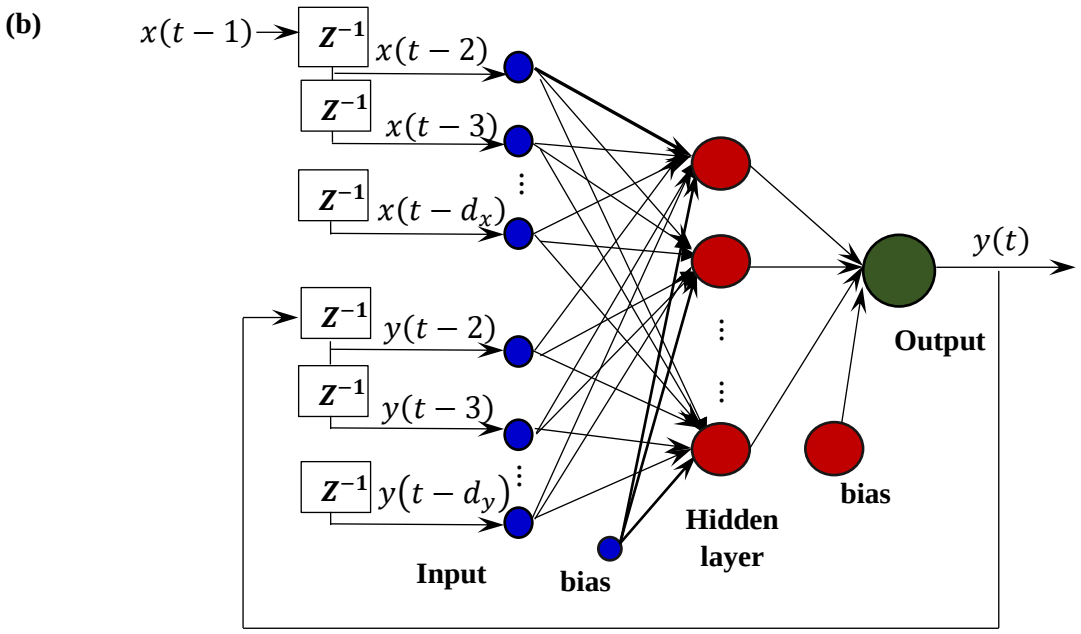


Figure 6. The structure of the NARX Neural Network (a) Open loop configuration, (b) Closed loop configuration (adapted from Ref. [63])

Because of the availability of output data, which refers to building heating and cooling energy use and electricity data during the training phase[2], the baseline models were trained in Series-Parallel (SP) or open-loop configuration (Figure 6). The model trained in SP configuration provides better prediction performance than in Parallel (P) or closed-loop configuration, since the input to the feedforward network is more accurate. Also, the resulting network has a pure feedforward architecture, and static backpropagation can be used for training. Levenberg-Marquardt, with the capability of high computational speed and the best convergence to a minimum of MSE for function approximation problems [68, 69] was selected as the learning algorithm to train the model.

During the training phase, the dataset (input and target data defining the desired output y(t)) was randomly divided into two sets: 70% for training, 30% for validation. Validation was performed

[2] Detailed information about the datasets can be found in Section 3.2.1.

to measure network generalization and to halt training when generalization stops improving. The Matlab '*preparets*' function was used to prepare the data. This function uses the network structure to determine how to divide and shift the data appropriately.

A MATLAB script was generated using the MATLAB neural net time series toolbox to reproduce the results with necessary modifications to the code. By modifying the MATLAB script, the trained network was converted to a closed-loop configuration and used to predict baseline energy use for the testing period. As part of the standard NARX architecture, estimated outputs are fed back and included in the output regressor in the closed-loop configuration.

3.2.4. Gaussian Process

The Gaussian Process (GP) defines a prior over function, which can be converted into a posterior over the function's values at a finite, but arbitrary set of points, $x_1, \ldots, x_N$. This model assumes that $p(f(x_1), \ldots, f(x_N))$ is jointly Gaussian, with some mean $\mu(x)$ and covariance $\sum(x)$ given by $\sum_{ij} = k(x_i, x_j)$, where k is a positive definite kernel function and depends on the Q-dimensional input variables $x_i$and x_j.

In a Gaussian Process Regression (GPR) model, the covariance function plays a vital role in the predictive mean and variance. Covariance functions contain presumptions about the function to be learnt and determines the correlation in the response as a function of the distance between the predictor values. As a result, the choice of covariance function may have profound impacts on the performance of a GPR model. A wide range of covariance functions e.g., exponential, squared exponential, Matern 5/2, rational quadratic was considered during hyperparameter tuning.

A GPR model can be defined as

$$f(x) \sim GPR[\mu(x), k(\theta)|x|] \quad (12)$$

3.2.5. Ensembles of Trees

These models combine results from many weak learners into one high-quality ensemble model [70]. The models' complexity and predictive accuracy largely depend on the hyperparameter values - leaf size and a number of learners.

This study used a least-squares boosting (LSBoost) learning algorithm for this regression model. At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners ($y_n - \eta f(x_n)$) grown previously, where y_n is the observed response, $f(x_n)$ is the aggregated prediction from all weak learners grown so far for observation, x_n and η is the learning rate. The ensemble fits to minimize mean-squared error.

3.2.6. Linear Regression

The linear regression process, as shown in Figure 7, fits the data to a straight line, producing a model that can be used to predict future data.

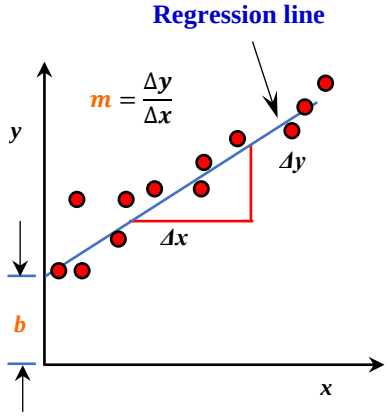


Figure 7. A schematic representation of linear regression model (adapted from [60])

To fit a linear regression model to linear or quadratic data, the following equation is used:

$$y = f(x) = mx + b \tag{13}$$

Where, $f(\,)$ is a linear function and m and b denote the slope and y-intercept value of the model, respectively. For multiple linear regression, Equation (14) can be written as

$$y = f(x_1, x_2, x_3, \ldots) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \cdots + \beta_i * x_i \quad (14)$$

where y, x_i, β_i, β_0 indicate the response feature, predictor features, regression coefficients, and intercept (bias) term, respectively.

3.2.7. Regression Trees

A decision tree is a non-parametric approach that identifies different ways of splitting a data set based on conditions until the information gain is zero. Construction of a tree usually inherits a top-down approach where a variable is chosen at each step, which 'best' splits the set of data [66]. The 'best' split is characteristic to the algorithm used. In order to predict responses, decisions are trailed from the root node to the leaf nodes [66]. Generally, the model development process of regression tree involves two steps:

1. The set of possible values of the predictors, also known as predictor space (*X*) is divided into *J* distinct and non-overlapping regions and are given by R_1, R_2, …, R_J. The regions are constructed in a way that it minimizes the residual sum of squares and can be expressed by Eq. (15):

$$\sum_{j=1}^{j} \sum_{i \in R_j} (y_i - \bar{y}_{R_j})^2 \quad (15)$$

2. For every observation that falls into a region R_j, prediction is made that indicates the mean of response in the training set in that particular region.

3.3. Modelling procedure

A basic workflow, presented in **Error! Reference source not found.**, was followed to develop multivariate baseline models and use the model for predicting baseline energy use during Pre-COVID and COVID lockdown periods spanning from January 1 to July 4, 2020. Note that during the COVID lockdown period, there was limited access to the case study building for the occupants. The modelling steps are detailed in the following subsections:

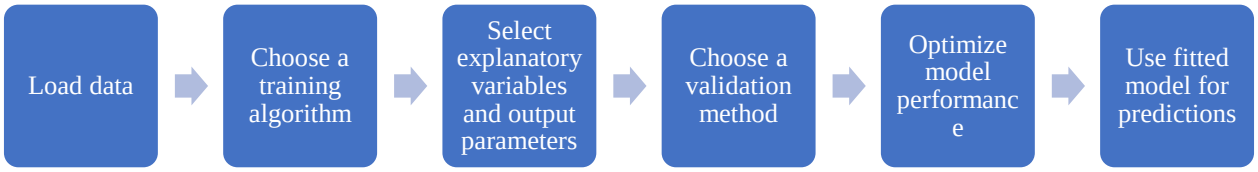


Figure 8. A basic workflow for baseline model development and prediction of baseline energy use

3.3.1. Description of datasets and data pre-processing

For the case study building, the hourly time series data as defined below and detailed in Section 3.1 were split into a training period and prediction period. The prediction period was defined as the most recent six months of the available data (e.g., January 1 to July 4, 2020) representing the pre-COVID and COVID lockdown periods. The models were trained using the nine-month worth of data (e.g., April 1 to December 31, 2019) that immediately preceded the prediction period. As illustrated in Figure 3, a dataset of seven explanatory variables (i.e., five weather variables, e.g., temperature (T), relative humidity (RH), wind speed (WS), wind direction (WD), and solar radiation (SR), time of day, occupancy data) and three output parameters (i.e., building heating and cooling energy uses and electricity consumption) were used in the feature selection and baseline model development process.

3.3.1.1. Data collection

Energy data

The hourly fifteen-month worth of energy use data (e.g., April 1, 2019 to July 4, 2020) for heating, cooling, and electricity were extracted from the building management system (BMS) server of the case study building using Schneider meter network.

Weather data

For the same period, hourly weather data were extracted from the actual meteorological year (AMY) weather files [42] based on the location of the case study building. Note that each weather-related explanatory variable (i.e., T, RH, WS, WD, and SR) denotes a variable for the multivariate baseline models of building heating and cooling energy use and electricity consumption.

Time of day data

Also, the information of time of day (HH:MM) was considered as an explanatory input parameter for the baseline models by generating a sinusoid function from the raw data, *x*. Figure 9 shows that for $x = -\pi\ radians$ which is equivalent to 0:00 as per 24 hrs clock y is equal to 0. Likewise, for $x = -1.64\ radians$ which is equivalent to 6:00 as per 24 hrs clock y is equal to -1.

Generate the sine function over the domain $-\pi \leq x \leq \pi$

```
x=-pi:0.25:pi;
y=sin(x);
```

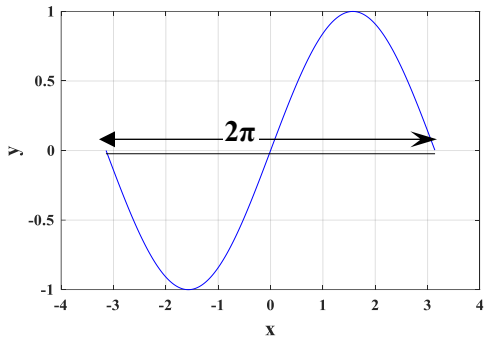


Figure 9. Plot for the sine function

Occupancy-sensing data

The locally distributed Wi-Fi beacons were used to locate a smart device, i.e., a smartphone by requesting connections, or the ID of a device's connected Wi-Fi network was logged over time. Thus, the occupancy entry and departure events were detected based on the continuity of smart devices' transmitted packets received by building's Wi-Fi network's access points (APs).

Wi-Fi enabled device counts were used as a proxy for occupancy. To approximate the number of occupants, the peripheral Wi-Fi devices (Wi-Fi printers, etc.) that remained on overnight were subtracted. Also, necessary calculations were performed as per past study [71] to receive the occupancy data from Wi-Fi enabled device count data. Note that the issue of average occupants carrying 1.2 Wi-Fi devices per person was considered in this study. Past studies such as Hobson, et al. [71] demonstrate the applicability of Wi-Fi data as the means of occupancy counts. The Wi-Fi data were used to gather 15 min interval occupancy data. Subsequently, down-sampling was performed by a factor of four to obtain hourly occupancy data. Note that whole-building-level occupancy data was used in the baseline models as an explanatory variable.

This research attempted to identify the influence of occupancy level on building energy use. Each model's predictive performance was evaluated, considering occupancy data as an explanatory variable and compared predictions with the scenarios where occupancy data was not considered.

3.3.1.2. Data cleaning and pre-processing

The raw data collected from multiple sources, such as Schneider meter network, (AMY) weather files [42], Cisco CMS IT network, passed through a data pre-processing procedure. For example, the information of time of day was converted to a sinusoid function, the 15-min interval Wi-Fi data was converted to hourly data by down-sampling. Thus, a dataset of seven explanatory variables (i.e., five weather variables, e.g., T, RH, WS, WD, and SR, time of day, occupancy data) and three output parameters (i.e., building heating and cooling energy uses and electricity consumption) was created for using them in the model development process. Before starting a detailed modelling process, the raw dataset was screened, and data cleaning was performed (using Matlab's '*filloutliers'* function) to replace the outlier values with a median value over a sliding window of length 24 hours, representing one day of data. Please note that according to ASHRAE guideline 14 [72], a maximum of 25% of measured data can be eliminated from the baseline performance period as part of the data cleaning process before using them in the model development process.

3.3.2. Selection of input parameters

For the individual multivariate baseline models, an optimum number of input parameters were selected using the sequential forward feature selection method [73-75]. Please note the term "feature" refers to independent input parameters. The MATLAB Machine Learning Toolbox was used in the feature selection process. As illustrated in Figure 10, this method starts with an empty set and adds one feature in the first step, which gives the highest value for the objective function. From the second step onwards, the remaining features were added separately to the current subset, and the new subset was evaluated. In each iteration, one feature that best improved the model performance was added. The process was repeated until all the features that were primarily selected

were investigated. If at any stage, model performance did not improve after adding a feature to the existing subset, the iteration process stopped. The performance of the trained model was assessed at each stage based on performance indicators: predictive accuracy and model fitting capability. These performance indicators were quantified by the performance metrics: coefficient of variation of the root mean square error (CV(RMSE)), and coefficient of determination (R-Squared).

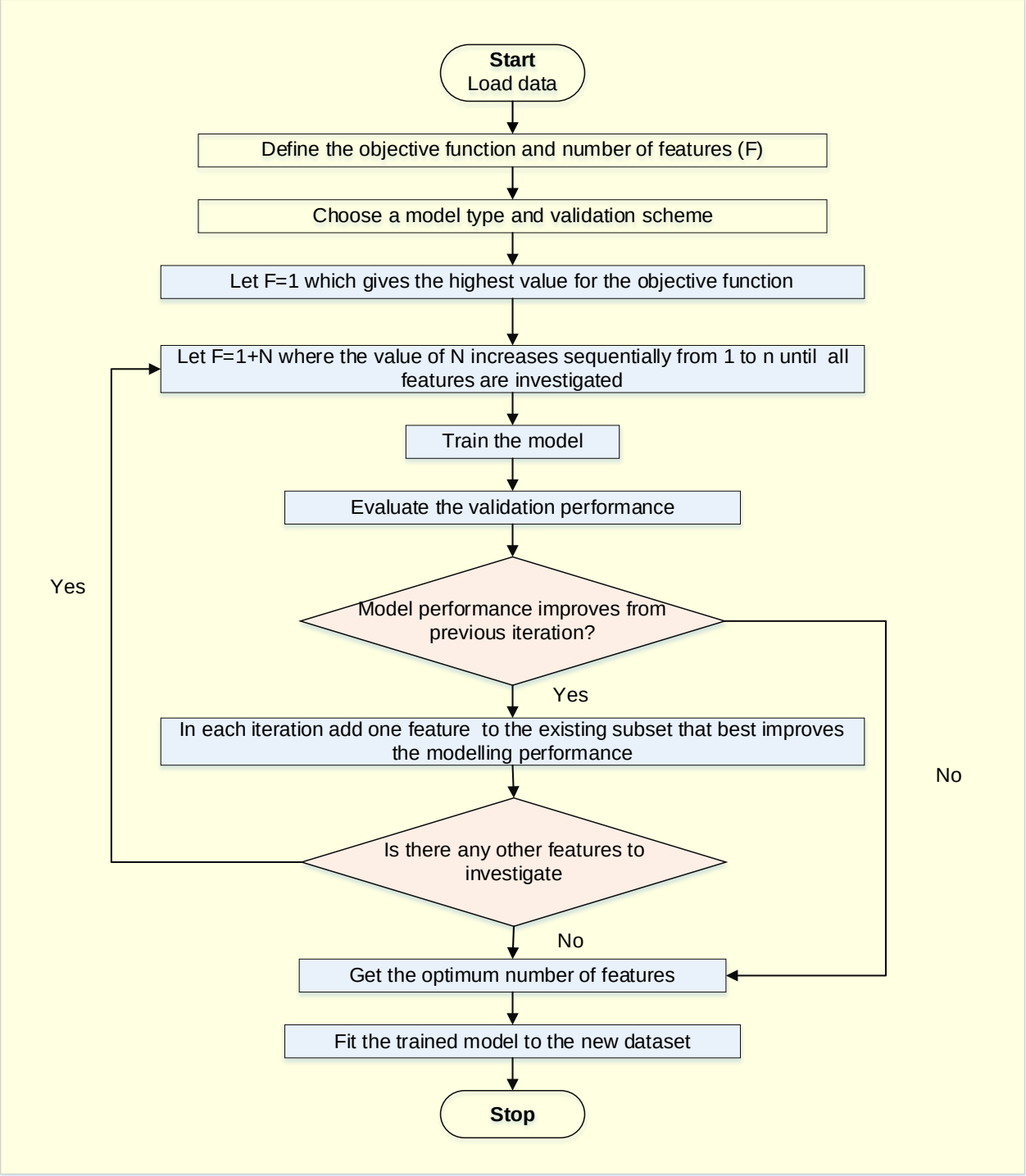


Figure 10. Feature selection process (adopted from [76])

3.3.3. Selection of modelling approaches

Highly flexible models tend to overfit data by modelling minor variations that can create noise. On the contrary, simple models come with the capability of easy interpretation but may have lower accuracy. Therefore, selecting the most suitable baseline modelling approach necessitates a trade-off.

Primarily, a wide range of multivariate regression modelling approaches, e.g., support vector machine, linear regression, regression trees, Gaussian process, ensembles of trees, and NARX neural network were considered for baseline model development. Using the optimum number of input parameters, their performances were compared in terms of predictive accuracy and model fitting capability and quantified by the statistical performance metrics: CV(RMSE), and R-Squared. From this analysis, the three best-performing modelling approaches were selected for the subsequent analysis. The validation of baseline models was performed using a five-fold cross-validation technique; a method known to protect against overfitting by partitioning the data set into five-folds and estimating accuracy on each fold. This method consists of splitting a dataset into five folds or disjoint subsets of the same size; then, iteratively, some are used to learn the model, while the others are utilized to assess its performance [77]. For each fold, the model was trained using the out-of-fold observations and model performance was assessed using in-fold data. Finally, the average test error was calculated overall folds.

3.3.4. Optimization of model performance

Hyperparameter[3] tuning can be an important step of modelling if the model is prone to overfitting. For example, to tune an SVM model, a set of box constraints and kernel scales were chosen; the

[3] A hyperparameter acts as an internal parameter of a regression function, such as the box constraint of a support vector machine.

model was cross-validated for each pair of values, and then their performances were compared using five-fold cross-validated RMSE estimates. Please note this study used Matlab Statistics and Machine Learning Toolbox to perform automatic hyperparameter tuning through Bayesian optimization. Bayesian optimization maintains a Gaussian process model of the objective function and uses objective function evaluations to train the model. The optimization minimizes the model loss based on the selected validation options.

The special process of tuning the number of iterations for such an algorithm is called "Early Stopping". Early Stopping performs model optimisation by monitoring the model's performance on a separate validation data set and stopping the training procedure once the performance on the validation data stops improving beyond a certain number of iterations. It avoids overfitting by attempting to automatically select the inflection point where performance on the validation dataset starts to decrease while performance on the training dataset continues to improve as the model starts to overfit. Early stopping can be based either on an out of bag sample set ("OOB") or cross-validation ("cv").

3.3.5. Evaluation and comparative analysis

Cross-validation was applied to facilitate the quantification of the baseline model predictive accuracy. The accuracy of 21 baseline models was quantified and evaluated by the statistical metric CV(RMSE) (coefficient of variation of the root mean square error), and R-Squared.

As per ASHRAE Guideline 14 [72], a CV(RMSE) of 30% or below indicates a good model fit for hourly data with acceptable predictive capabilities and is given by:

$$CV(RMSE) = \left(\frac{RMSE}{\bar{y}}\right) \times 100 \quad (16)$$

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

where, $y_i$is the measured energy use, $\hat{y}_i$is the predicted energy use, $\bar{y}$ is the mean of the measured energy use, N is the number of data points.

The coefficient of determination (R^2) is the measure of how well the independent variables explain variation in the dependent variable. R^2 value ranges from 0 to 1. R^2 value equal to one indicates a "perfect fit" of the regression line to the data and can be expressed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (17)$$

The rule-of-thumb for an acceptable model using monthly energy data is $R^2 > 0.75$ [78, 79]. For the case study the minimum limit was set to 0.70 for the hourly data. It is recommended to select the best model out of different regression models depending on the value of R^2 and CV(RMSE) [72].

The performance of multivariate regression models selected from the various approaches in section 3.2 and the univariate adapted change point models for electricity, heating and cooling energy uses was compared quantitively and the results are presented in Chapter 4.

In the analysis, the performance of three multivariate regression models was evaluated for two different scenarios – including and without including occupancy data as an explanatory variable to examine the influence of occupancy on the accuracy of baseline models.

The intent of the comparative analysis presented in the next Chapter is to understand the influence of occupancy on the building energy performance. Also, a univariate adapted change point model is proposed in this study to provide an insight into a simple baseline model.

Chapter 4

Results and Discussion

This Section presents a comparative analysis of all modelling approaches considered in this research for predicting baseline heating and cooling energy uses and baseline electricity use in the case study building. The models' performance was evaluated in terms of predictive accuracy, data fitting capability and operational insights (e.g., building operational schedule and performance, energy use profiles during occupied and unoccupied modes). The influence of occupancy level on building energy use was investigated. Each individual model's predictive performance was evaluated, considering occupancy data as an explanatory variable, and compared with the modelling scenario where occupancy data was not considered.

A comparative analysis of multi-variate regression modelling approaches is presented in the context of the selection of explanatory variables for the models. The performance of the models was evaluated at each stage of the input parameter selection process based on performance indicators: predictive accuracy and data fitting capability. The performance of multi-variate baseline models was assessed using an optimum number of input parameters, and best performing models were employed on the testing dataset representing the prediction period.

4.1. Assessment of related weather data

A range of hourly weather data gathered from the local weather station were divided into two segments – training and testing or prediction periods. Starting from January 1 to July 4, 2020, the six months prediction period characterizes the pre-COVID and COVID lockdown stages of Canada. The pre-COVID and COVID periods are defined in Figure 12. The training period represents a nine-month worth of data spanning from April 1 to December 31, 2019. Figure 11

presents the distributions of the weather data used in this study. The symbols μ and σ stand for the mean and the standard deviation, respectively. Recall that the training and testing periods do not encompass the same months of the year. Therefore, variations were observed to the distributions of individual weather data, e.g., outdoor temperature, relative humidity, wind speed, wind direction, solar irradiance, including the μ and σ values. Note that the distribution plots of the weather data coloured in blue and red represent the training and testing periods, respectively.

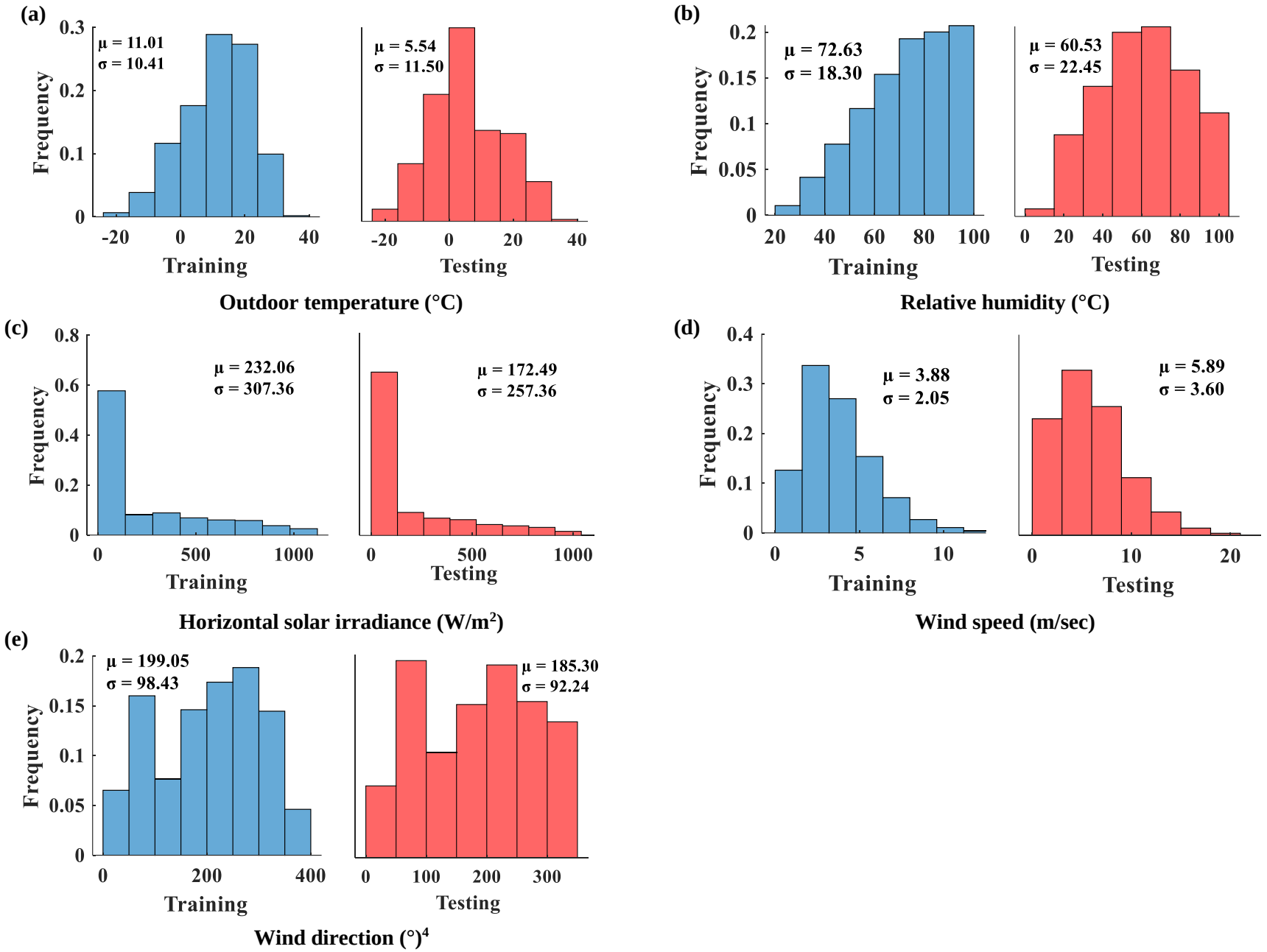


Figure 11.Weather conditions during training (Apr 1 – Dec 31, 2019) and prediction periods (Jan 1 – Jul 4, 2020)

[4] Wind direction is typically reported in degrees and describes the direction from which the wind originates. A direction of 0 degrees is due North on a compass.

4.2. Assessment of hourly energy data and associated occupancy data

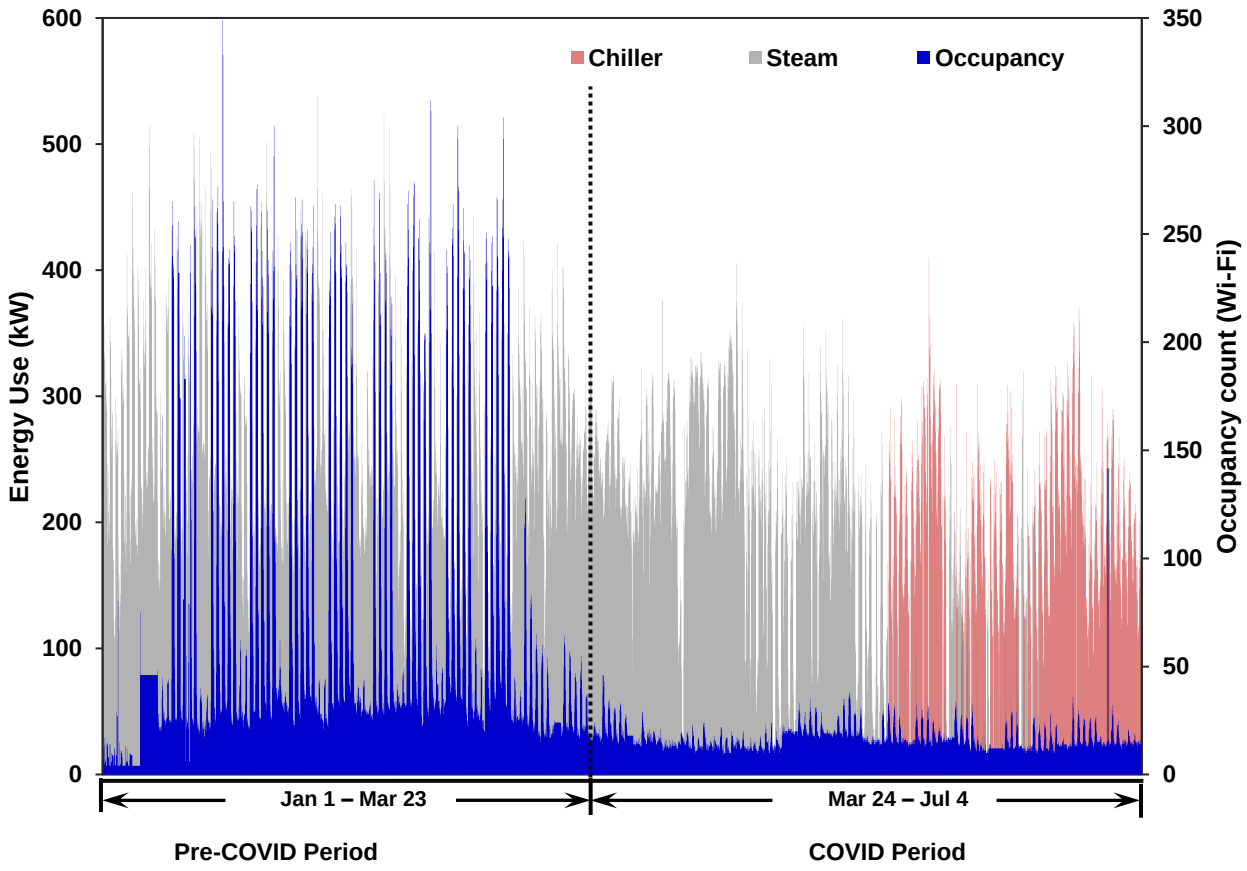
Figure 12 illustrates the relative trend between energy use and occupancy level for the pre-COVID and COVID periods. The hourly heating and cooling energy use data clearly show different time spans, with heating energy use data occasionally overlapping the cooling energy use data during the cooling season (see Figure 12(a)). Occupancy level visibly dropped during the COVID period, and it did not fluctuate much throughout the day. However, heating and cooling energy use data did not follow this changing occupancy pattern during the COVID period. This is partly because an occupant-centric control is not maintained by this building, and as such, energy used by the case study building does not reflect the rapid drop in occupancy level in the pandemic. This issue will be more apparent in the following sections by developing baseline models for energy use, including and excluding occupancy data as an explanatory variable for the models. Note that examples of occupant-centric controls include demand control ventilation (DCV) at the AHU-level, zone-level occupant count-based modulation of minimum airflow setpoint of the VAV terminals, an adaptation of zone mode of operation based on motion detectors. Existing literature [80, 81] shows that in buildings with occupancy sensing technologies, energy use follows a close relationship with occupancy.

Figure 12(b) shows a steady electricity consumption pattern from January to May with a certain degree of electricity fluctuations throughout the day depending on time. The electricity consumption abruptly moved up by around 300 kW during the cooling season, indicating the share of electricity consumed by the functioning chiller of the case study building.

Figure 13(a) and Figure 13(b) present two typical days' electricity and occupancy patterns representing the pre-COVID and COVID situations. Compared with the COVID situation, in the pre-COVID situation, electricity consumption was high and largely followed the occupancy

changing pattern. The occupancy curve characterizing the pre-COVID situation represents the dual-peak feature during 24 hours as described in ASHRAE 90.1 standard [82]. However, on that particular day of the COVID period, a steady electricity consumption profile was observed, and this consumption was quite high compared to a very low occupancy level during this time. (see Figure 13(b)).

(a)



(b)

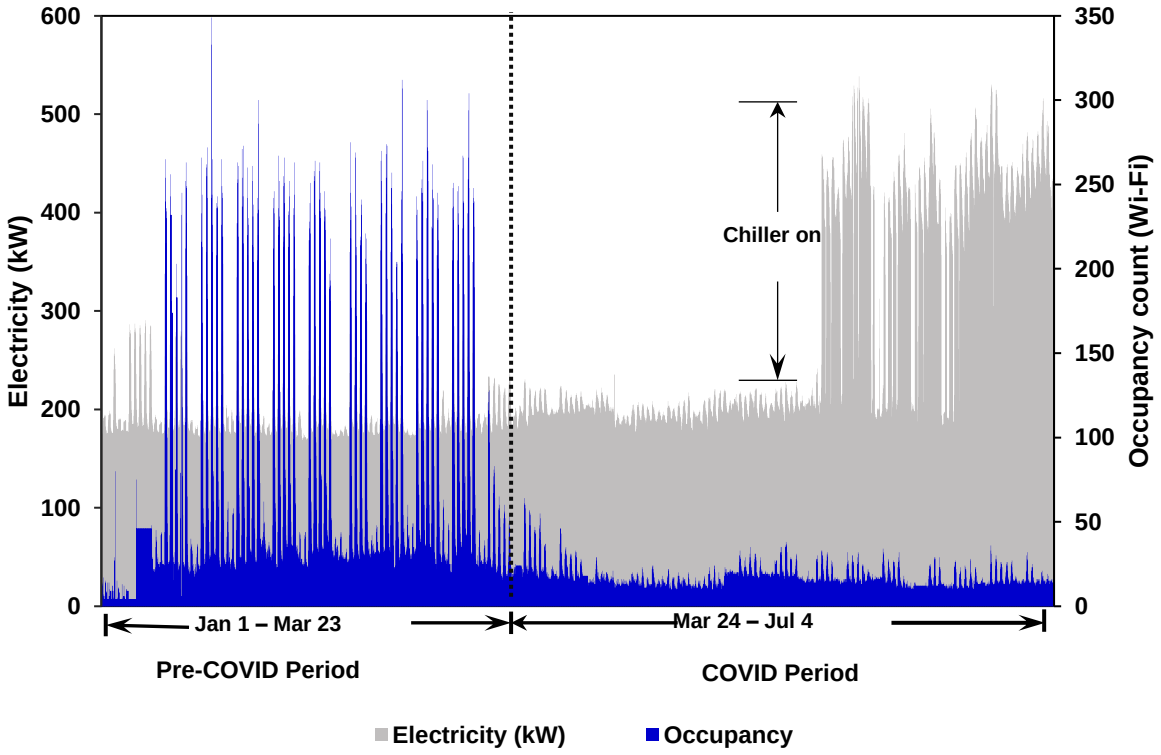
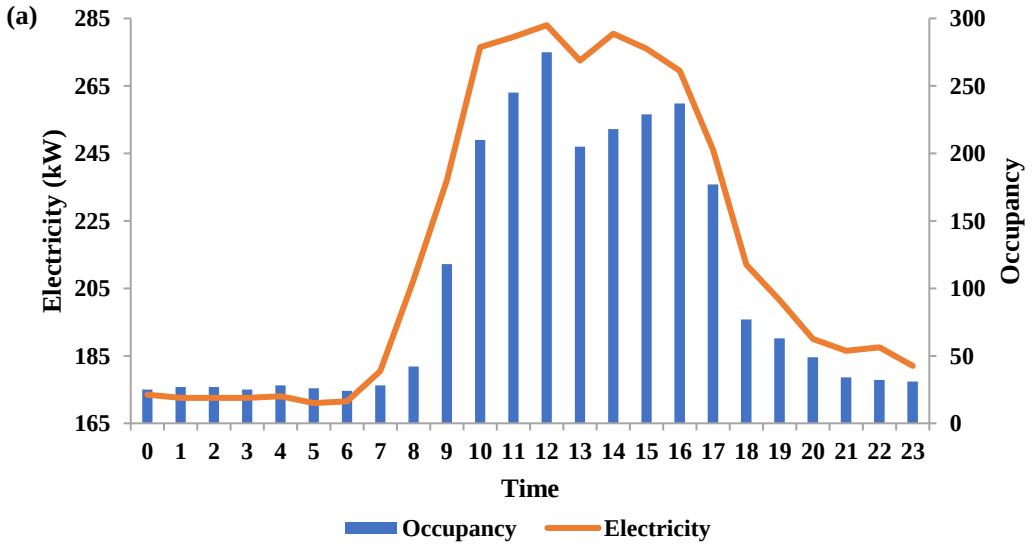


Figure 12. The relative trend between energy use and occupancy level (a) Heating and cooling energy use, (b) Electricity



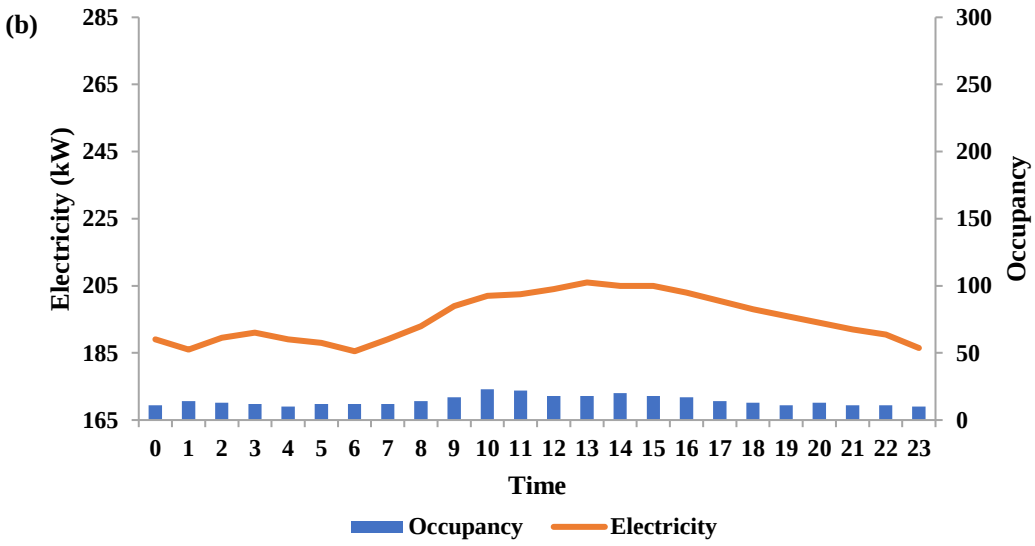


Figure 13. A typical day's electricity and occupancy pattern (a) Pre-COVID situation, (b) COVID Situation

4.3. Adapted change point model

Hourly data spanning from April 1 to December 31, 2019, were used to train and validate the adapted change point models. Fitting the trained validated models for cooling energy use, heating energy use, and electricity consumption to the data of prediction period (January 1 to June 30, 2020), the baseline energy use for the Pre-COVID and COVID were estimated. Note that the only regressor used in these models was the outdoor temperature. For the case study building the $x_{1-10,C}$, $x_{1-10,H}$ *and* $x_{1-10,E}$ values as detailed in Chapter 3 Section 3.2.1. were determined from the baseline models for two operating modes (e.g., Occupied and Unoccupied) of the case study building (Table 3).

Table 3. The change point temperature, y-intercept, and slope parameter values for two modes of operation determined from the baseline models for cooling energy use, electricity, and heating energy use

	Parameters									
Cooling	$x_{1,C}$	$x_{2,C}$	$x_{3,C}$	$x_{4,C}$	$x_{5,C}$	$x_{6,C}$	$x_{7,C}$	$x_{8,C}$	$x_{9,C}$	$x_{10,C}$
	2	23	1	0	24	12	11	17	18	16
Electricity	$x_{1,E}$	$x_{2,E}$	$x_{3,E}$	$x_{4,E}$	$x_{5,E}$	$x_{6,E}$	$x_{7,E}$	$x_{8,E}$	$x_{9,E}$	$x_{10,E}$
	6	20	0	0	23	205	13	15	191	15
Heating	$x_{1,H}$	$x_{2,H}$	$x_{3,H}$	$x_{4,H}$	$x_{5,H}$	$x_{6,H}$	$x_{7,H}$	$x_{8,H}$	$x_{9,H}$	$x_{10,H}$

7	18	0	0	10	75	20	7	42	12

As per the parameter information identified from the adapted change point model and presented in Table 3, for cooling, the case study building was scheduled to be in occupied mode for 21 hrs (2 am to 11 pm). For electricity and heating, the schedules for the occupied mode were found to be 6 am to 8 pm, and 7 am to 6 pm, respectively. The operating schedules of the case study building for the occupied mode of energy use are presented in Figure 14.

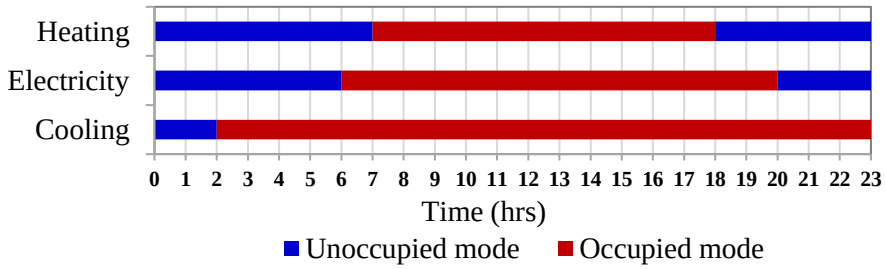


Figure 14. Schedule for the occupied mode for heating, electricity, and cooling energy uses

According to the information revealed from the change point model, during the cooling season, the case study building's cooling system was operational on Saturday. Note that the parameter values 1 and 0 for $x_{3,C}$ and $x_{4,C}$ indicate the building's cooling system was in operating and non-operating status, respectively. For the electricity, the case study building was in non-operating status on Saturday and Sunday (Table 3). Similarly, during the heating season, the case study building's heating system was in non-operating status on weekends.

The change point temperatures of the models refer to the points at which the model switches from weather dependent to non-weather dependent behaviour. According to the parameter values of $x_{7,C}$ and $x_{10,C}$, the case study building required cooling in the building occupied and unoccupied modes when the outdoor temperature was equal to and above 11°C and 16°C, respectively. The case study building required heating in the occupied and unoccupied modes when the outdoor temperature was below 20°C and 12°C, respectively (Table 3 for the parameter values $x_{7,H}$ and $x_{10,H}$). This

indicates that the case study building required both heating and cooling in the occupied mode when the outdoor temperature was between 11°C and 20°C. On the other hand, the building did not require any heating and cooling energy in the unoccupied mode when the outdoor temperature was between 12°C and 16°C. The electrical energy use tends to vastly increase in the occupied and unoccupied modes when the outdoor temperature was equal to and above 13°C and 15°C, respectively (Table 3 for the parameter values $x_{7,E}$ and $x_{10,E}$). A possible reason behind this could be the use of chillers in the cooling season. Note that these findings reflect heating and cooling operating schedules in the occupied and unoccupied modes of the building corresponding to prediction period (January 1 to June 30, 2020).

According to the parameter values of $x_{5,C}$ and $x_{8,C}$, for every one-degree Celsius rise in the outdoor temperatures above the change point temperature, the cooling energy use intensity[5] tends to increase by 2.66 W/m^2 and 1.88 W/m^2 in the occupied and unoccupied modes, respectively. Similarly, with every one-degree Celsius rise in the outdoor temperatures above the change point temperature, the electrical energy intensity increases by 2.55 W/m^2 and 1.66 W/m^2 in the occupied and unoccupied modes, respectively. On the contrary, for every one-degree Celsius rise in the outdoor temperatures before it reaches the change point temperature, the heating energy use intensity drops by approximately 1.11 W/m^2 and 0.78 W/m^2 in the occupied and unoccupied modes, respectively.

The y-intercept values $x_{6,E}, x_{9,E}, x_{6,H}, x_{9,H}, x_{6,C}$ and $x_{9,C}$ indicate the minimum expected electrical, cooling and heating energy use of the case study building in the occupied and unoccupied modes independent of the outdoor temperature. For the case study building the minimum electrical energy

[5] The increase or decease in energy use intensity was calculated by dividing slope values by the gross floor area of the case study building.

use intensities were around 22.71W/m^2 and 21.16 W/m^2 in the occupied and unoccupied modes, respectively independent of the outdoor temperature.

4.4. Selection of input parameters and its impact on the predictive performance of the models

As detailed in Chapter 3, seven input parameters e.g., temperature (T), relative humidity (RH), wind speed (WS), wind direction (WD), solar radiation (SR), time of day (HH:MM), occupancy sensing data (Wi-Fi data) were considered in the feature selection process. Note that in the analysis, each input parameter refers to individual features. For the individual regression models, features were selected using the sequential forward feature selection approach. Two separate analysis was performed including and excluding the Wi-Fi data, and their predictive performances were compared in terms of CV(RMSE), and R-squared values. Note that Wi-Fi data were used in the model as the proxy for occupancy.

In the feature selection process, features were added successively to the existing subset depending on model performance. The explicit feature selection process of the regression models for cooling energy use is illustrated in Appendix A Table 1 A to Table 6 A. Table 4 presents the comparative predictive performance of the models in each iteration. The features selected in this process indicate the best combination of input parameters for the individual models. The results indicate that while a similar feature selection process was followed with all models, the same sequence and combination of input parameters did not show the best performance in terms of statistical performance metrics – CV(RMSE), and R-Squared values. For instance, while with Ensembles of Trees the ultimate combination of input parameters was found to be T+RH+Wi-Fi+WS+SR+HH:MM+WD, with SVM, Gaussian process, linear regression, regression trees, and

NARX different sequence and combination of input parameters were observed. The predictive performance of individual models slightly improves when Wi-Fi data was considered.

Table 4. Comparative predictive performance of multi-variate baseline models for cooling energy use

Regression models		Predictors	Performance metrics	
			CVRMSE	**R-Squared**
Ensembles of Trees	Including Wi-Fi data	T	22.61	0.82
		T+RH	20.15	0.86
		T+RH+Wi-Fi	19.70	0.87
		T+RH+Wi-Fi +WS	19.47	0.87
		T+RH+Wi-Fi +WS+SR	19.28	0.87
		T+RH+Wi-Fi +WS+SR+ HH:MM	19.23	0.87
		T+RH+Wi-Fi+WS+SR+HH:MM+WD	**19.22**	**0.87**
	Excluding Wi-Fi data	T	22.61	0.82
		T+RH	20.15	0.86
		T+RH+WS	19.81	0.86
		T+RH+WS +HH:MM	19.63	0.87
		T+RH+WS +HH:MM +WD	19.61	0.87
		T+RH+WS +HH:MM +WD+SR	**19.60**	**0.87**
SVM	Including Wi-Fi data	T	25.61	0.77
		T+RH	23.76	0.80
		T+RH+WS	23.47	0.81
		T+RH+WS+WD	23.39	0.81
		T+RH+WS+WD+Wi-Fi	23.31	0.81
		T+RH+WS+WD+Wi-Fi+SR	23.14	0.81
		T+RH+WS+WD+WiFi+SR+HH:MM	**23.12**	**0.82**
	Excluding Wi-Fi data	T	25.61	0.77
		T+RH	23.76	0.80
		T+RH+WS	23.47	0.81
		T+RH+WS +WD	23.39	0.81
		T+RH+WS+WD+SR	23.36	0.81
		T+RH+WS+WD+SR+HH:MM	**23.33**	**0.81**
Regression tree	Including Wi-Fi data	T	23.72	0.81
		T+RH	21.67	0.84
		T+RH+HH:MM	19.98	0.85
		T+RH+HH:MM+Wi-Fi	19.62	0.85
		T+RH+HH:MM+Wi-Fi +SR	**19.45**	**0.85**
	Excluding Wi-Fi data	T	23.72	0.81
		T+RH	21.67	0.84
		T+RH+HH:MM	**19.98**	**0.85**
Gaussian Process	Including Wi-Fi data	T	22.37	0.83
		T+RH	19.98	0.86
		T+RH +WD	19.47	0.87
		T+RH +WD + Wi-Fi	19.07	0.87
		T+RH +WD + Wi-Fi +HH:MM	18.71	0.88
		T+RH +WD + Wi-Fi +HH:MM +WS	**18.39**	**0.88**
	Excluding Wi-Fi data	T	22.37	0.83
		T+RH	19.98	0.86

		T+RH +WD	19.47	0.87
		T+RH +WD +WS	19.12	0.87
		T+RH +WD +WS +HH:MM	**18.82**	**0.88**
Linear regression	Including Wi-Fi data	T	37.40	0.53
		T+RH	32.28	0.58
		T+RH +WD	30.31	0.60
		T+RH +WD + Wi-Fi	29.73	0.60
		T+RH +WD + Wi-Fi +SR	**29.47**	**0.61**
	Excluding Wi-Fi data	T	37.40	0.53
		T+RH	32.28	0.58
		T+RH +WD	30.31	0.60
		T+RH +WD+WS	29.88	0.60
		T+RH +WD+WS+SR	**29.63**	**0.60**
NARX Neural Network	Including Wi-Fi data	T	17.27	0.92
		T+RH	14.78	0.94
		T+RH+SR	14.32	0.94
		T+RH+SR+Wi-Fi	**13.73**	**0.95**
	Excluding Wi-Fi data	T	17.27	0.92
		T+RH	14.78	0.94
		T+RH+SR	14.32	0.94
		T+RH+SR+HH:MM	**14.09**	**0.94**

Table 5 and Table 6 show the ultimate combination of input parameters for the individual models of cooling energy use selected through the feature selection process, including and excluding Wi-Fi data. When Wi-Fi data is considered, we can see quite different results compared to their peers in terms of the combination of input parameters. In all cases, the T and RH indicating outdoor temperature and relative humidity were found to be the common useful explanatory variables for all multi-variate regression models. A possible explanation behind differences in the input parameter selection process with different baseline modelling approaches is that they differ functionally from each other. Detailed information about these models can be found in [83-87].

Table 5. The comparison of sequential feature selection process to predict cooling energy use including Wi-Fi data

Regression models	Combination of input parameters						
	T	RH	SR	WS	WD	Wi-Fi	HH:MM
Ensembles of Trees	●	●	●	●	●	●	●
Support Vector Machine	●	●	●	●	●	●	●
Regression tree	●	●	●			●	●
Gaussian Process	●	●		●	●	●	●

Linear regression	●	●	●		●	●	
NARX Neural Network	●	●	●			●	

Table 6. The comparison of sequential feature selection process to predict cooling energy use excluding Wi-Fi data

Regression models	Combination of input parameters					
	T	RH	SR	WS	WD	HH:MM
Ensembles of Trees	●	●	●	●	●	●
Support Vector Machine	●	●	●	●	●	●
Regression tree	●	●				●
Gaussian Process	●	●		●	●	●
Linear regression	●	●	●	●	●	
NARX Neural Network	●	●	●			●

Similarly, for the prediction to baseline electricity and heating energy use, the same order and combination of input parameters did not show the best performance with different multi-variate regression models in terms of statistical performance metrics – CV(RMSE) and R-Squared values (Table 7 to Table 12). The detailed feature selection process of the regression models for electricity and heating energy use can be found in Appendix A, Table 7 A to Table 13 A. The results indicate that compared to baseline cooling energy use and electricity, the baseline models for heating energy use provide relatively low performance. It appears that the nine-month worth of data that were used to train the models do not represent the heating season as much as they represent the cooling season. However, further studies are required to enquire into the effect of seasonal variations and the size of the training dataset on the model's predictive performance. The combination of input parameters selected through the feature selection process was quite different when Wi-Fi data were considered compared to their peers where Wi-Fi data were not considered (Table 8 and Table 9; Table 11 and Table 12). In both cases, solar radiation was found to be a useful explanatory variable along with outdoor temperature and relative humidity to predict heating energy use.

Table 7. Comparative predictive performance of multi-variate baseline models for electricity

Regression models		Predictors	Performance metrics	
			CVRMSE	R-Squared
Ensembles of Trees	Including Wi-Fi data	T	17.07	0.80
		T+Wi-Fi	16.09	0.82
		T+Wi-Fi +RH	14.19	0.86
		T+Wi-Fi +RH +SR	13.78	0.87
		T+Wi-Fi +RH +SR +WS	13.58	0.87
		T+Wi-Fi +RH +SR +WS +WD	13.54	0.87
		T+Wi-Fi+RH+SR+WS+WD+HH:MM	13.41	0.88
	Excluding Wi-Fi data	T	17.07	0.8
		T+HH:MM	15.97	0.82
		T+HH:MM +RH	15.26	0.84
		T+HH:MM +RH +WS	15.06	0.84
		T+HH:MM +RH +WS +WD	15.03	0.84
		T+HH:MM+RH+WS+WD+SR	14.94	0.85
Support Vector Machine	Including Wi-Fi data	T	20.13	0.75
		T+ RH	18.97	0.77
		T+ RH+WS	18.35	0.79
		T+ RH+WS+Wi-Fi	17.67	0.79
		T+ RH+WS+Wi-Fi+WD	17.09	0.79
	Excluding Wi-Fi data	T	20.13	0.75
		T+ RH	18.97	0.77
		T+ RH+WS	18.35	0.79
		T+ RH+WS+WD	17.85	0.79
		T+ RH+WS+WD+HH:MM	17.56	0.79
Regression tree	Including Wi-Fi data	T	17.76	0.80
		T+Wi-Fi	16.35	0.82
		T+Wi-Fi +RH	14.54	0.85
		T+Wi-Fi +RH+WS	14.21	0.86
		T+Wi-Fi +RH+WS+SR	13.95	0.87
		T+Wi-Fi +RH+WS+SR+WD	13.67	0.87
		T+Wi-Fi +RH+WS+SR+WD+HH:MM	13.51	0.88
	Excluding Wi-Fi data	T	17.76	0.80
		T+RH	16.27	0.82
		T+RH+SR	14.89	0.85
		T+RH+SR+HH:MM	14.63	0.86
		T+RH+SR+HH:MM+WS	14.46	0.86
Gaussian Process	Including Wi-Fi data	T	16.47	0.81
		T+Wi-Fi	14.51	0.85
		T+Wi-Fi +RH	13.31	0.88
		T+Wi-Fi+RH +SR	12.70	0.89
		T+Wi-Fi+RH+SR +WS	12.54	0.89
		T+Wi-Fi +RH +SR +WS +WD	12.26	0.9
	Excluding Wi-Fi data	T	16.47	0.81
		T+HH:MM	15.22	0.84
		T+HH:MM +RH	14.31	0.86
Linear regression	Including Wi-Fi data	T	32.89	0.61
		T+RH	30.21	0.63

		T+RH+Wi-Fi	29.13	0.63
		T+RH+Wi-Fi+WS	28.83	0.63
		T+RH+Wi-Fi+WS+HH:MM	28.12	0.64
		T+RH+Wi-Fi+WS+HH:MM+SR	27.91	0.64
	Excluding Wi-Fi data	T	32.89	0.61
		T+ RH	30.21	0.63
		T+ RH+HH:MM	29.32	0.63
		T+ RH+HH:MM+WS	29.05	0.63
		T+ RH+HH:MM+WS+WD	28.77	0.63
		T+ RH+HH:MM+WS+WD+SR	28.13	0.64
NARX Neural Network	Including Wi-Fi data	T	14.34	0.88
		T+Wi-Fi	12.67	0.91
		T+Wi-Fi +RH	11.73	0.92
		T+Wi-Fi+RH +WS	11.36	0.92
		T+Wi-Fi+RH+WS+SR	11.07	0.92
		T+Wi-Fi +RH +WS+SR+HH:MM	10.78	0.93
	Excluding Wi-Fi data	T	14.34	0.88
		T+ RH	12.73	0.90
		T+RH +WS	12.37	0.90

Table 8. The comparison of sequential feature selection process to predict electricity including Wi-Fi data

Regression models	Combination of input parameters						
	T	RH	SR	WS	WD	Wi-Fi	HH:MM
Ensembles of Trees	●	●	●	●	●	●	●
Support Vector Machine	●	●		●	●	●	
Regression tree	●	●	●	●	●	●	●
Gaussian Process	●	●	●	●	●	●	
Linear regression	●	●	●	●		●	●
NARX Neural Network	●	●	●	●		●	●

Table 9. The comparison of sequential feature selection process to predict electricity excluding Wi-Fi data

Regression models	Combination of input parameters					
	T	RH	SR	WS	WD	HH:MM
Ensembles of Trees	●	●	●	●	●	●
Support Vector Machine	●	●		●	●	●
Regression tree	●	●	●	●		●
Gaussian Process	●	●				●

Linear regression	●	●	●	●	●	●
NARX Neural Network	●	●		●		

Table 10. Comparative predictive performance of multi-variate baseline models for heating energy use

Regression models		Predictors	Performance metrics	
			CVRMSE	**R-Squared**
Ensembles of Trees	Including Wi-Fi data	T	30.32	0.69
		T+HH:MM	26.06	0.77
		T+HH:MM +RH	25.81	0.78
		T+HH:MM +RH +SR	25.55	0.78
		T+HH:MM +RH +SR +Wi-Fi	25.49	0.78
		T+HH:MM+RH +SR+Wi-Fi +WS	25.40	0.78
	Excluding Wi-Fi data	T	30.32	0.69
		T+HH:MM	26.06	0.77
		T+HH:MM +RH	25.81	0.78
		T+HH:MM +RH +SR	25.55	0.78
		T+HH:MM +RH +SR +WS	25.49	0.78
Support Vector Machine	Including Wi-Fi data	T	33.73	0.64
		T+RH	29.82	0.71
		T+RH+SR	28.79	0.71
		T+RH+SR+WS	28.56	0.72
		T+RH+SR+WS+Wi-Fi	28.27	0.72
		T+RH+SR+WS+Wi-Fi+WD	28.03	0.72
	Excluding Wi-Fi data	T	33.73	0.64
		T+RH	29.82	0.71
		T+RH+SR	28.79	0.71
		T+RH+SR+WS	28.56	0.72
		T+RH+SR+WS+HH:MM	28.35	0.72
Regression tree	Including Wi-Fi data	T	30.71	0.68
		T+RH	26.17	0.77
		T+RH+HH:MM	25.78	0.78
		T+RH+HH:MM+SR	25.61	0.78
		T+RH+HH:MM+SR+Wi-Fi	25.52	0.78
		T+RH+HH:MM+SR+Wi-Fi+WS	25.47	0.78
	Excluding Wi-Fi data	T	30.71	0.68
		T+RH	26.17	0.77
		T+RH+HH:MM	25.78	0.78
		T+RH+HH:MM+SR	25.61	0.78
		T+RH+HH:MM+SR+WD	25.55	0.78
Gaussian Process	Including Wi-Fi data	T	29.99	0.70
		T+HH:MM	25.59	0.78
		T+HH:MM +SR	24.68	0.79
		T+HH:MM +SR +RH	24.15	0.80
		T+HH:MM +SR +RH +Wi-Fi	24.04	0.80
	Excluding Wi-Fi data	T	29.99	0.70
		T+HH:MM	25.59	0.78
		T+HH:MM+SR	24.68	0.79
		T+HH:MM+SR+RH	24.15	0.80

		T+HH:MM+SR+RH +WD	24.11	0.80
Linear regression	Including Wi-Fi data	T	52.63	0.45
		T+RH	46.82	0.52
		T+RH+SR	43.85	0.54
		T+RH+SR+Wi-Fi	42.47	0.55
		T+RH+SR+Wi-Fi+WS	42.03	0.55
	Excluding Wi-Fi data	T	52.63	0.45
		T+RH	46.82	0.52
		T+RH+SR	43.85	0.54
		T+RH+SR+WS	42.87	0.55
		T+RH+SR+WS+HH:MM	42.56	0.55
NARX Neural Network	Including Wi-Fi data	T	24.27	0.82
		T+RH	21.23	0.86
		T+RH+Wi-Fi	19.73	0.87
		T+RH+Wi-Fi+WS	19.54	0.87
		T+RH+Wi-Fi+WS+HH:MM	19.17	0.87
		T+RH+Wi-Fi+WS+HH:MM+SR	18.89	0.88
	Excluding Wi-Fi data	T	24.27	0.82
		T+RH	21.23	0.86
		T+RH+SR	19.94	0.87
		T+RH+SR+WS	19.72	0.87
		T+RH+SR+WS+HH:MM	19.29	0.87

Table 11. The comparison of sequential feature selection process to predict heating energy use including Wi-Fi data

Regression models	Combination of input parameters						
	T	RH	SR	WS	WD	Wi-Fi	HH:MM
Ensembles of Trees	●	●	●	●		●	●
Support Vector Machine	●	●	●	●	●	●	
Regression tree	●	●	●	●		●	●
Gaussian Process	●	●	●			●	
Linear regression	●	●	●	●		●	
NARX Neural Network	●	●	●	●		●	●

Table 12. The comparison of sequential feature selection process to predict heating energy use excluding Wi-Fi data

Regression models	Combination of input parameters					
	T	RH	SR	WS	WD	HH:MM
Ensembles of Trees	●	●	●	●		●
Support Vector Machine	●	●	●	●		●
Regression tree	●	●	●		●	●
Gaussian Process	●	●	●		●	●
Linear regression	●	●	●	●		●
NARX Neural Network	●	●	●	●		●

4.5. Models' predictive performance

The predictive performance of multi-variate regression models for cooling energy use, electricity, and heating energy use were quantified using an optimum number of input parameters. Figure 15, Figure 16, and Figure 17 present a comparative performance of the models for two different situations - including and excluding Wi-Fi data. The results indicate that for cooling energy use, and electricity CV(RMSE) values of all considered multi-variate regression models satisfy the acceptable limit of 30% specified by ASHRAE Guideline 14 [72] for hourly data. On the other hand, except for linear regression, all regression models for cooling energy use, and electricity satisfy the minimum limit of 0.70 for R^2 value, set as per [78, 79] (Chapter 3 Section 3.3.5). The baseline models for heating energy use developed with linear regression do not satisfy the acceptable limits for CV(RMSE) and R^2 values (Figure 17). However, in all cases NARX performs the best in terms of predictive performance quantified by the values of CV(RMSE) and R-squared, followed by Gaussian Process and Ensembles of Trees. Note that computational complexity was not considered in this analysis as the computational runtime of the considered models was found less than or close to one-minute.

From the predictive performance modelling results, it appears that Wi-Fi data do not have much influence on the energy use calculations. This suggests that the case study building's energy use is not occupancy controlled. It seems to be the building AHUs do not hold occupant centric control features. Also, there could be a lack of intelligent sensor technologies such as PIR motion detectors and occupancy sensors with the BMS system.

Based on the performance results of the considered models, three modelling approaches – NARX, Gaussian Process, and Ensembles of Trees were selected and employed in the subsequent studies.

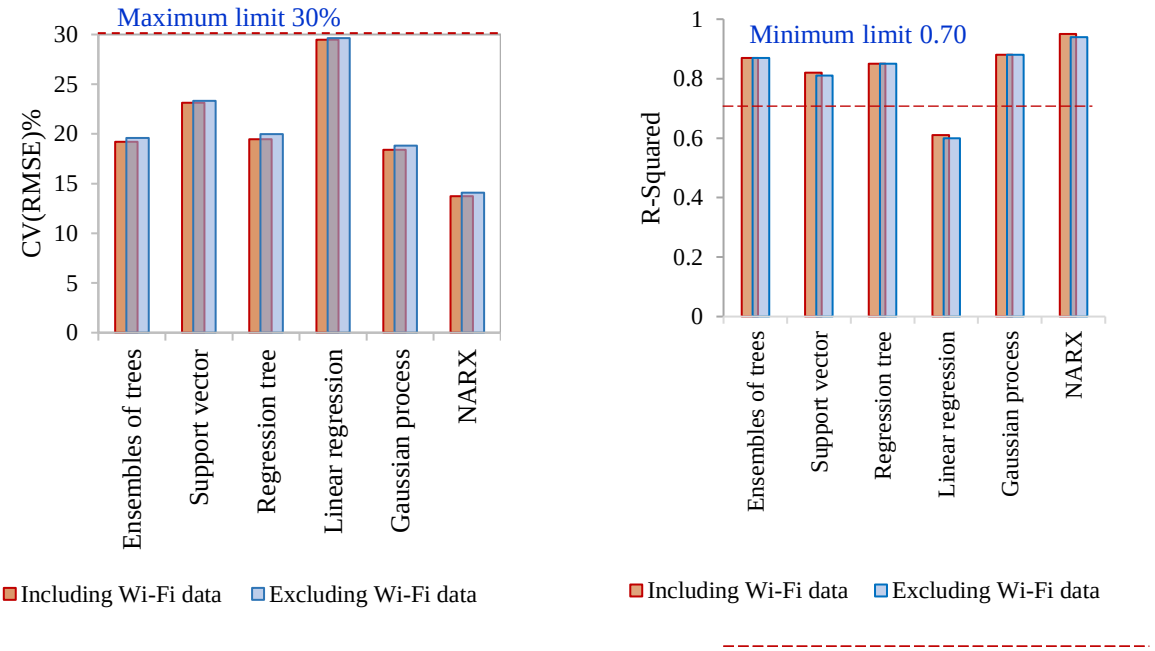


Figure 15. The comparative predictive performance results of individual multi-variate regression models for cooling energy use including and excluding Wi-Fi data (a) CV(RMSE), (b) R-squared

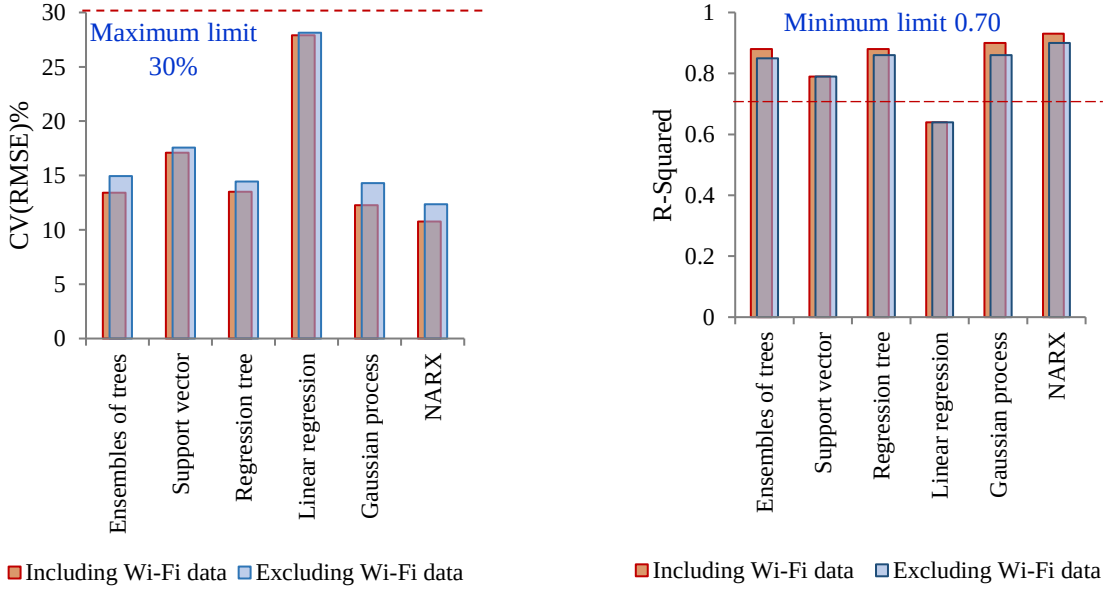


Figure 16. The comparative predictive performance results of individual multi-variate regression models for electricity including and excluding Wi-Fi data (a) CV(RMSE), (b) R-squared

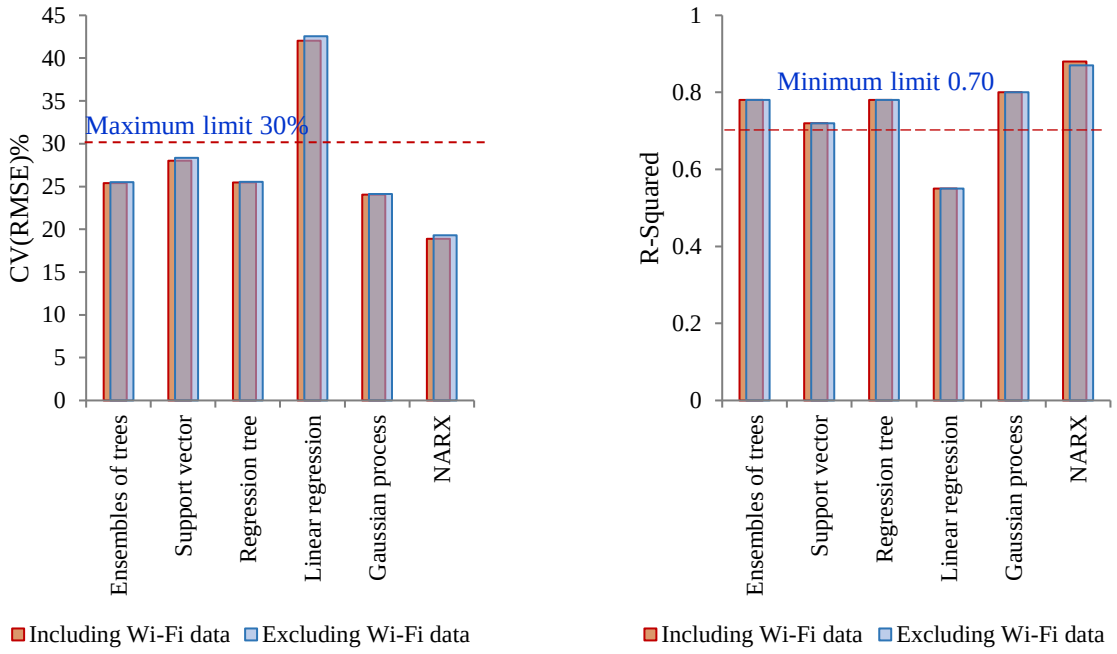


Figure 17. The comparative predictive performance results of individual multi-variate regression models for heating energy use including and excluding Wi-Fi data (a) CV(RMSE), (b) R-squared

4.6. Optimization of model performance

As an important step to avoid overfitting and optimize model performance, hyperparameters were automatically tuned for the selected models using Bayesian optimization. Bayesian optimization locates a point that minimizes an objective function. The searched point represents a set of hyperparameter values, and the objective function is the mean squared error (MSE). It uses the acquisition function to determine the next set of hyperparameter values for the subsequent iteration process. The hyperparameter search ranges and tuned values used to optimize the performance of the regression models for cooling energy use, heating energy use and electricity is presented in Appendix B Table 1 B.

A Minimum MSE Plot is presented in Figure B1 of Appendix B for cooling energy use model with the Ensembles of Trees (excluding Wi-Fi data). This figure shows that at each iteration, the Bayesian optimization tries a different combination of hyperparameter values and updates the plot

with the minimum validation mean squared error (MSE) observed up to that iteration, indicated in dark blue. At the end of the optimization process, the optimizer selects the set of optimized hyperparameters, indicated by a red square.

4.7. Evaluation and comparative analysis

The trained optimized models were employed on the testing datasets to predict baseline energy use for cooling, heating, and electricity. Note that the training results statistics were calculated on the validation datasets.

The accuracy of 21 baseline models was quantified on the training and testing datasets and evaluated by the statistical metric CV(RMSE), and R-Squared values. In Table 13 their performances are compared across four models for two different situations – including and without including occupancy data as an explanatory variable. Note that the adapted change point model being a univariate regression model, only one explanatory variable – the outdoor temperature was used in the models for heating energy use, electricity, and cooling energy use.

In all cases, NARX provided the best performance. On average, all studied models provided the highest predictive accuracies for the electricity followed by cooling energy use and heating energy use. A possible reason for that could be that the whole nine month-worth of data used to train the models represents electricity uses at different times of the day. On the contrary, during the heating season, cooling energy use tends to be zero. Similarly, during the cooling season, there could be minimal heating energy use because of the hot water systems. Due to the seasonal variations, the entire nine months' heating or cooling energy use data do not represent the heating or cooling seasons.

Table 13. Predictive performance of the models for cooling energy use, electricity, and heating energy use on different datasets

Regression model	Dataset		CV(RMSE)	R-Squared
Cooling Energy Use				
Adapted change point model	-	Training	21.04	0.78
		Testing	27.75	0.71
Ensembles of Trees	Including Wi-Fi data	Training	17.87	0.88
		Testing	21.35	0.79
	Excluding Wi-fi data	Training	18.21	0.88
		Testing	22.19	0.76
Gaussian Process	Including Wi-Fi data	Training	17.13	0.89
		Testing	21.73	0.79
	Excluding Wi-Fi data	Training	17.87	0.89
		Testing	22.63	0.76
NARX Neural Network	Including Wi-Fi data	Training	11.37	0.96
		Testing	15.26	0.89
	Excluding Wi-Fi data	Training	12.09	0.95
		Testing	16.76	0.87
Electricity				
Adapted change point model	-	Training	16.18	0.79
		Testing	21.53	0.73
Ensembles of Trees	Including Wi-Fi data	Training	12.56	0.89
		Testing	18.65	0.82
	Excluding Wi-fi data	Training	13.12	0.87
		Testing	20.73	0.81
Gaussian Process	Including Wi-Fi data	Training	11.95	0.89
		Testing	17.73	0.82
	Excluding Wi-Fi data	Training	12.32	0.87
		Testing	19.32	0.78
NARX Neural Network	Including Wi-Fi data	Training	10.52	0.93
		Testing	15.26	0.85
	Excluding Wi-Fi data	Training	11.78	0.92
		Testing	17.76	0.83
Heating Energy Use				
	-	Training	25.43	0.75

Adapted change point model		Testing	30.02	0.67
Ensembles of Trees	Including Wi-Fi data	Training	23.75	0.79
		Testing	27.61	0.72
	Excluding Wi-fi data	Training	23.87	0.79
		Testing	28.93	0.71
Gaussian Process	Including Wi-Fi data	Training	22.74	0.82
		Testing	27.53	0.74
	Excluding Wi-Fi data	Training	23.07	0.81
		Testing	29.29	0.72
NARX Neural Network	Including Wi-Fi data	Training	17.54	0.89
		Testing	22.74	0.82
	Excluding Wi-Fi data	Training	18.53	0.88
		Testing	24.25	0.81

As an example, the baseline model predictive performance results with Gaussian Process are detailed below. Figure 18 illustrates the relationships between predicted and measured cooling energy use data for the testing period with Gaussian Process. The R-Square value shows how successful the fit is in explaining the variation of the data. The cooling energy use model, including Wi-Fi data as an explanatory variable, provides a slightly better fit compared to that of excluding Wi-Fi data. Note that only cooling season data of the testing period are presented in Figure 18. In the figure the measured values of zero indicate that chillers were not in the operating mode that time. During the cooling season the chiller system could be in the idle condition in the unoccupied mode of the case study building.

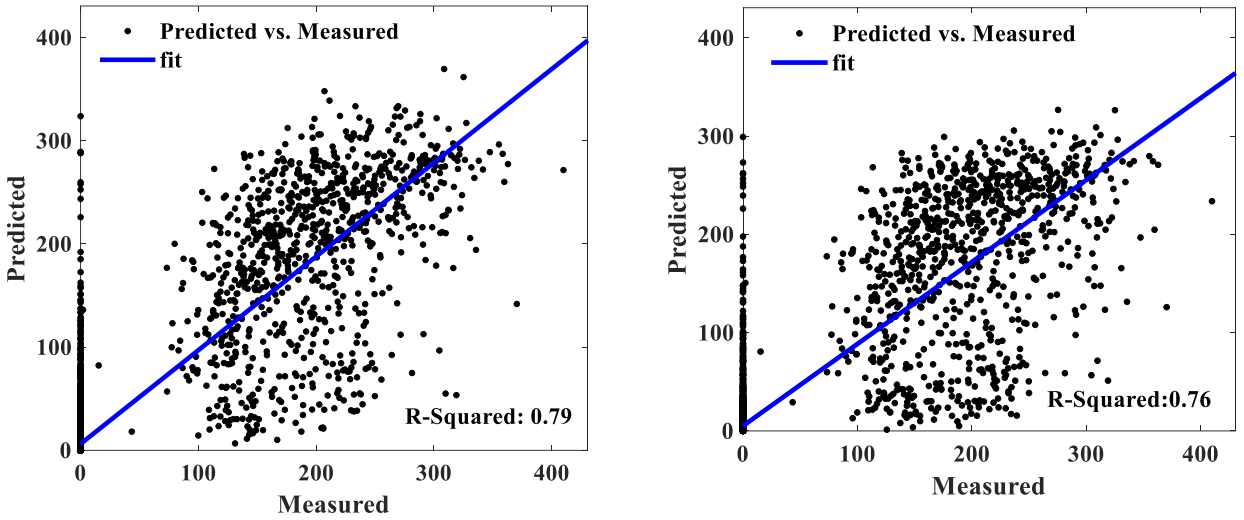


Figure 18. Predicted (using Gaussian Process) vs. measured cooling energy use data during testing period (a) Including Wi-Fi data (b) Excluding Wi-Fi data

Similarly, the baseline model for electricity using Gaussian Process shows a slightly better fit on the testing dataset when Wi-Fi data were considered in the model (Figure 19). Note that when the chiller system came into operation, electricity consumption jumped to a higher level. This gives an explanation for not getting any measured data between 280kW and 370kW.

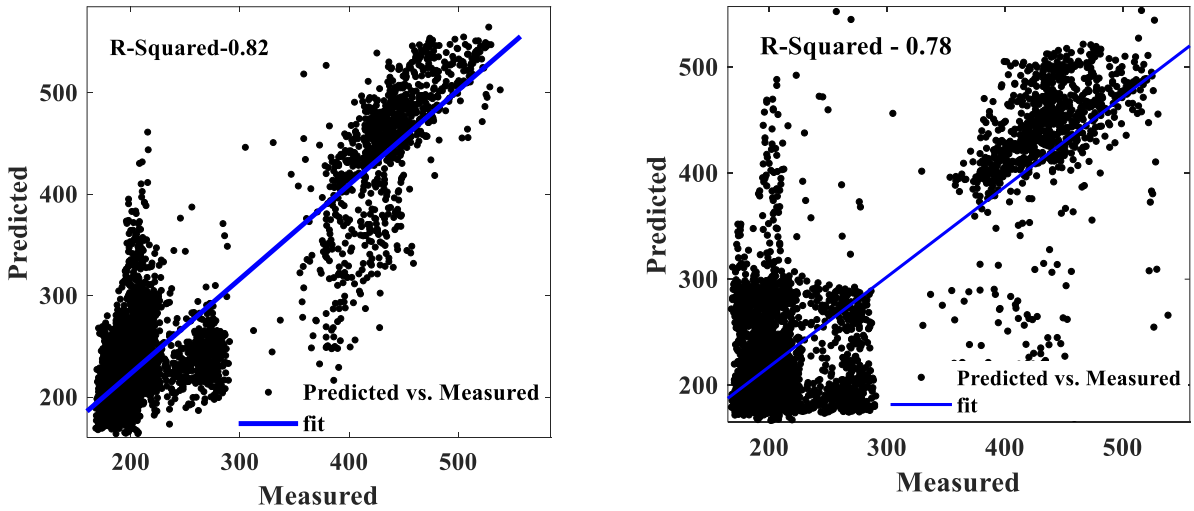


Figure 19. Predicted (using Gaussian Process) vs. measured electricity data during the testing period (a) Including Wi-Fi data (b) Excluding Wi-Fi data

Figure 20 presents a time-series plot for hourly measured data versus predicted data for cooling energy use. This figure provides a visual inspection of the measured data against their estimated peers with the Gaussian process. Two separate models were developed, including and excluding the Wi-Fi data as an explanatory variable and compared on the testing data of energy use. For the ease of visualization, a particular area of the plot is zoomed in. The model with Wi-Fi data provides quite similar prediction results to the model without Wi-Fi data. As detailed in Table 13, the model with Wi-Fi data provides slightly better predictive performance.

As shown in Figure 21 and Figure 22, electricity and heating energy use during the COVID period is better captured by the model with Wi-Fi data compared with the modelling scenarios where occupancy data was not considered. This situation can be better visualized from the zoomed-in plot.

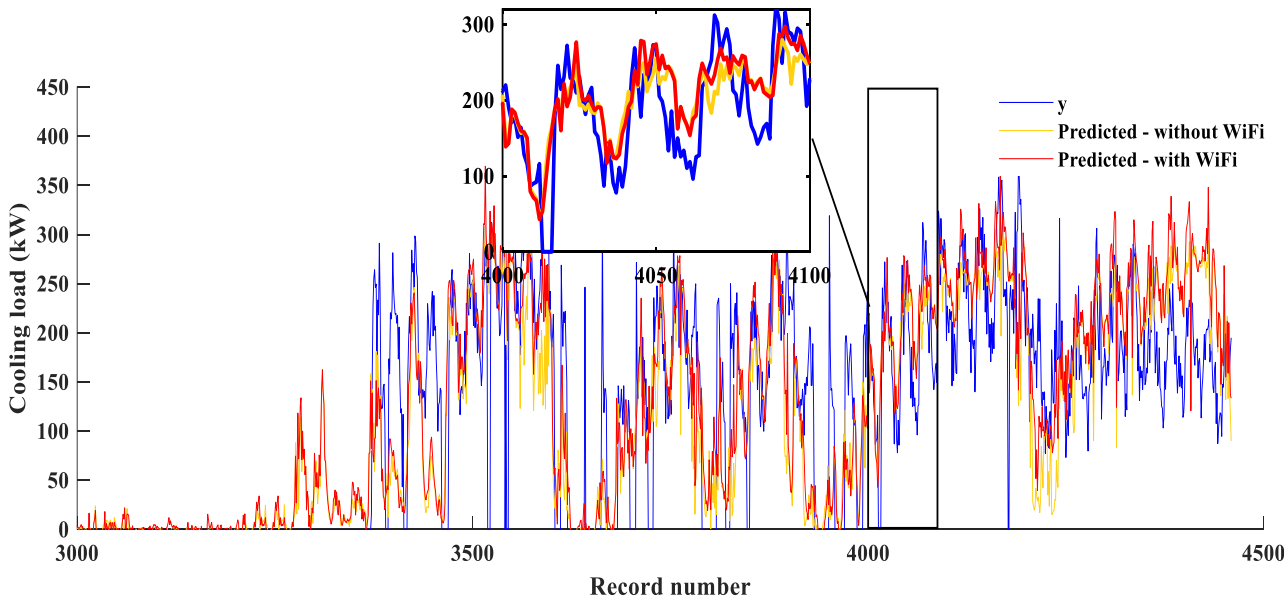


Figure 20. Predicted vs. measured hourly data for cooling energy use using Gaussian Process model

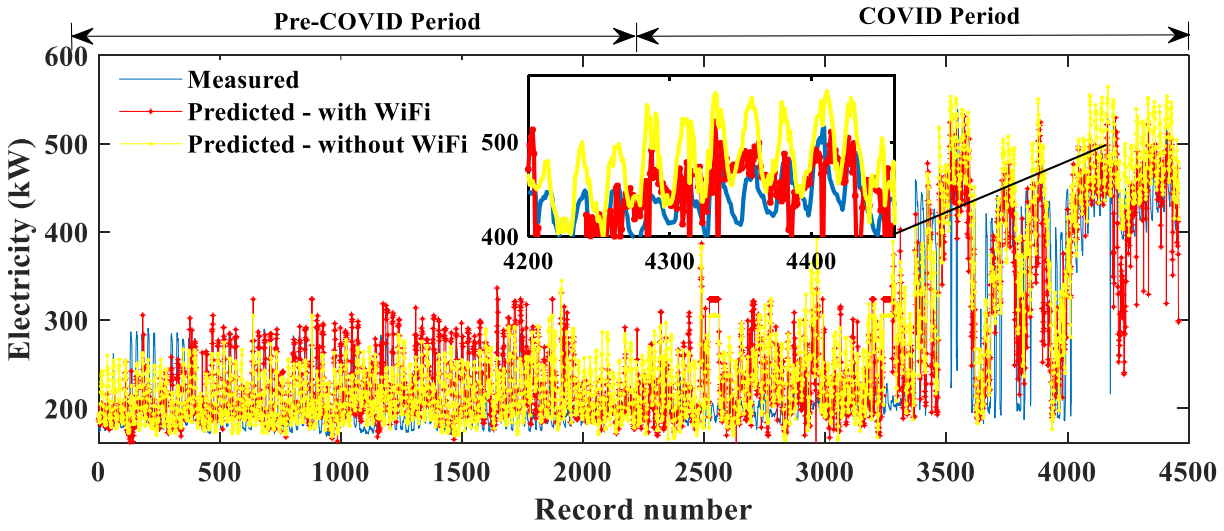


Figure 21. Predicted vs. measured hourly data for electricity using Gaussian Process model

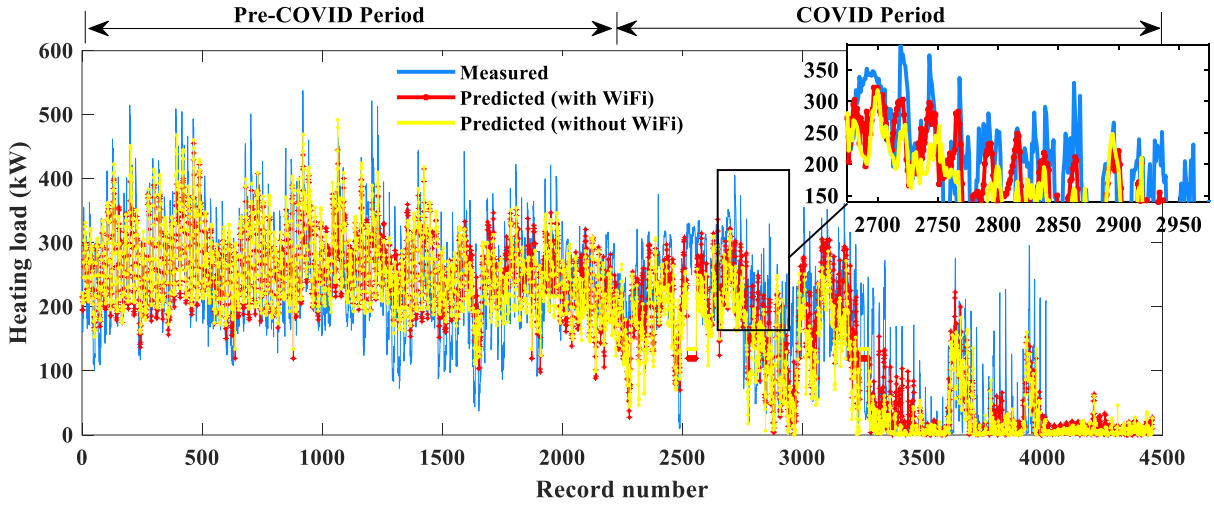


Figure 22. Predicted vs. measured hourly data for heating energy use using Gaussian Process

While the past studies [21-23] demonstrate the influence of occupancy level on the energy consumption in buildings, the presented baseline modelling results do not establish any substantial connection between occupancy level and energy use in the case study building. This indicates the absence of technologies that improve the adaptability of buildings to variable occupancy. A building featured with the intelligent building operation and control systems such as DCV, occupancy-centric zone-level ventilation, setbacks following zone-level occupancy, occupancy

presence-based lighting control would have seen a significant drop in energy use compared with a traditional building. Therefore, as a future enquiry, this analysis could be extended for a building with smart building technologies.

Chapter 5

Conclusions and Future Directions

5.1. Conclusions

For promoting energy efficiency practices in the building sector, energy conservation measures (ECMs) receive the highest importance nowadays. The decision on ECMs encounters several challenges due to the uncertainty over expected energy savings, lack of information and benchmarks about the actual performance of the building and its systems after the design phase, and a risk of poor execution of retrofit measures and resulting comfort consequences. A baseline energy model that acts as a reference point for a facility manager or a building operator assists in determining energy and cost-saving potential, energy system fault diagnostics, and acquiring physical insight into the operating patterns. The uncertainty of accurate prediction of baseline energy use that a building would have consumed if no ECMs had taken place, largely depends on the accuracy of the baseline model. Thus, energy conservation measures implemented through energy performance contracting (EPC) are predominantly linked with the performance of baseline building energy models. Even though baseline models were used in this study to evaluate the influence of occupancy rate on the baseline energy use, the outcome of the comparative study of modelling performance will assist the building practitioners in selecting suitable baseline models in the real-world applications.

The presented modelling results highlight the capabilities of baseline modelling approaches on multiple aspects such as predictive performance, ease of applicability, and ability to reveal building operational insight. Note that an insufficient number of inputs affects the prediction performance of a model. On the other hand, involvement of large numbers of input parameters comprising

excess or redundant input parameters creates unnecessary complexity, increases the probability of overfitting the network and decreases the computation speed during the execution of the model. The execution of sequential forward feature selection process in this study ensured an optimum number of input parameters for the individual multivariate baseline models. Thus, this demonstrates multivariate models' ease of applicability to the real system. On the other hand, high R-Squared values indicate good data fitting capacity of the models. Also, the models fulfilled the condition of ASHRAE Guideline 14 [72] which is $CV(RMSE) \leq 30\%$. This indicates a good model fit for hourly data with acceptable predictive capabilities. The presented adapted change point model as discussed in Section 4.3 comes with the capability of providing operational insight into the case study building such as identification of operating status and hours during the heating and cooling season.

The arrangement of different input parameters selected by a forward sequential feature selection approach was found to be an important step to identify the influence of individual parameters on baseline energy use. Differences were observed in each iteration in the input parameter selection process of electricity, heating, and cooling energy use baseline models. Hence, the same sequence and combination of input parameters did not show the best performance in terms of statistical performance metrics – CV(RMSE) and R-Squared values. For example, when two different regression modelling approaches, such as *Gaussian process* and *Ensembles of Trees*, were employed in the baseline model development process of building electricity or heating or cooling energy uses, the same combination of input parameters was not proved to be the best in terms of predictive accuracy. The dry bulb temperature followed by relative humidity was found to be the common useful parameters for all developed baseline models.

The results show that baseline model performance slightly improves when occupancy data are considered as an explanatory variable. However, based on the assessment of hourly energy and associated occupancy data it seems to be occupancy data can significantly influence the performance of a baseline energy use model in an occupant-centric building. In this instance, further studies are required to demonstrate the necessity of implementing occupant-centric building control strategies to improve its energy performance. A visual inspection of the measured data for electricity, heating, and cooling energy uses against their estimated peers with the *Gaussian process* shows that during the COVID period building energy use is better captured by the model with Wi-Fi data compared with the modelling scenarios where occupancy data was not considered.

The assessment of hourly energy and associated occupancy data indicates separate time spans for heating and cooling energy use. However, during the cooling season in the COVID period, heating energy was used occasionally. A possible reason behind this could be the use of domestic water heating systems in the cooling season. Occupancy level visibly dropped in the COVID period, and it remained pretty stable throughout the day. Note that in the COVID period, occupants had restricted access to the case study building. However, heating and cooling energy use did not seem to follow the occupancy pattern during this period. This could be partly because an occupant-centric control such as DCV, occupancy-centric zone-level ventilation, setbacks following zone-level occupancy, occupancy presence-based lighting control was not applied in this building, and as such, energy use of the case study building did not reflect the rapid fall in occupancy level in the pandemic period.

The results show that electricity consumption was steady periodic over daily and weekly cycles during the pre-COVID period. However, during the COVID period electricity consumption was

steady constant. During the cooling season, the electricity consumption abruptly shifted up, indicating the share of electricity consumed by the functioning chiller of the case study building. Evaluation of two typical days' electricity and occupancy patterns selected from the pre-COVID and COVID periods shows that occupancy level was much higher in the pre-COVID period and electricity profile largely followed the changing pattern of occupancy level. In the COVID period, a steady electricity consumption profile was observed, and this consumption was quite high compared to a very low occupancy level during this time.

NARX outperformed other baseline modelling approaches in terms of model predictive accuracy and model data fitting capabilities. On average, all studied models provided higher predictive accuracies for the electricity followed by cooling energy use and heating energy use. This could be due to the effect of seasonal variations. Note that the full nine months' worth of data used to train and validate the models for heating and cooling energy uses do not represent the heating or cooling season; rather the hourly heating and cooling energy use data clearly show different time spans. Therefore, further studies are required to enquire into the effect of seasonal variations and the size of the training dataset on the model's predictive performance.

Given the limited data time frame and considering a single case study building, a comprehensive conclusion cannot be drawn. The *Gaussian Process* and *Ensembles of Trees* performed almost similar with respect to the performance metrics such as CV(RMSE), and R^2 considered in the evaluation. All multivariate regression models tend to outperform the adapted change point model. However, the adapted change point model comes with the capability of providing operational insight into the case study building. According to ASHRAE Guideline 14 [72], a *change point model* that fits well between energy consumption and outdoor temperature can be considered one of the best baseline modelling approaches in practice. Research shows that change-point linear

models have physical significance to the actual heat loss/gain mechanisms that drive most buildings' energy use [72].

The modified change point model demonstrates its adaptability for the M&V applications. This model can be considered simplistic of all considered modelling approaches because of not requiring any advanced level analysis such as regressor selection, hyperparameter tuning, and adjusting the network size (suitable for *NARX*) to optimize the model performance. Conversely, the implementation of state-of-the-art machine learning baseline modelling approaches requires advanced level data analytics skills. Therefore, these modelling approaches could be principally applicable for a building system with an advanced metering interface [63].

The adapted change point models for heating, electricity, and cooling provided some important information related to the operating status and performance of the case study building. The baseline model for cooling energy use showed the longest operating hour (2 am to 11 pm) for cooling in the occupied mode compared with heating and electricity. Also, during the cooling season, the case study building's cooling system was found to be operational on Saturday. On the other hand, the case study building's heating system during the heating season and electricity were in non-operating status on weekends.

As per the parameter information identified from the adapted change point model, the case study building required both heating and cooling in the occupied mode when the outdoor temperature was between 11°C and 20°C. According to Afroz, et al. [63] and Gunay, et al. [88], the status of internal heat gain, large temperature variations within the adjacent zones, or inappropriate control programming could be responsible for simultaneous heating and cooling in a building. On the other hand, the building did not require any heating and cooling energy in the unoccupied mode when the outdoor temperature was between 12°C and 16°C. This indicates the presence of economizer

mode in the case study building for a certain outdoor temperature range. Due to the share of electricity consumed by the building's chiller system, the electrical energy use tends to vastly increase in the occupied and unoccupied modes when the outdoor temperature was equal to and above 13°C and 15°C, respectively.

As per the slope parameter values detected from the change point model, for every one-degree Celsius rise in the outdoor temperatures above the change point temperature, the cooling energy use intensity tends to increase by 2.66 W/m^2 and 1.88 W/m^2 in the occupied and unoccupied modes, respectively. Similarly, with every one-degree Celsius rise in the outdoor temperatures above the change point temperature, the electrical energy intensity increases by 2.55 W/m^2 and 1.66 W/m^2 in the occupied and unoccupied modes, respectively. On the contrary, for every one-degree Celsius rise in the outdoor temperatures before it reaches the change point temperature, the heating energy use intensity drops by approximately 1.11 W/m^2 and 0.78 W/m^2 in the occupied and unoccupied modes, respectively.

As per the y-intercept values, the minimum electrical energy use intensities were around 22.71 W/m^2 and 21.16 W/m^2 in the occupied and unoccupied modes, respectively, independent of the outdoor temperature. The minimum heating energy use intensities were found to be 8.31 W/m^2 and 4.65 W/m^2 in the occupied and unoccupied modes, respectively. On the contrary, the cooling energy requirements in the occupied and unoccupied modes were found to be approximately zero in the heating season.

This thesis attempted to address the research question “Can a baseline energy model where occupancy data were used as an explanatory variable better capture the changes in energy use baseline”? Based on the discussions mentioned above, it could be concluded that the influence of occupancy data on the predictive accuracy of baseline models largely depends on the state of

intelligent sensor technologies and zone-level metering infrastructure existing within the building. A building integrating occupant-centric control strategies seems to better follow the occupancy trend in terms of energy use. The results show that occupant-centric control strategies (such as DCV, occupancy-centric zone-level ventilation, setbacks following zone-level occupancy, occupancy presence-based lighting control) are not maintained in the case study building. Also, the assessment of hourly energy data and associated occupancy data shows that heating and cooling energy use data did not follow this changing occupancy pattern during the COVID period. On the other hand, evaluation of electricity consumption of two typical days' electricity and occupancy patterns representing the pre-COVID and COVID situations shows that in the pre-COVID situation, electricity consumption largely followed the occupancy changing pattern. However, on a particular day of the COVID period, a steady electricity consumption profile was observed, and this consumption does not seem to follow the occupancy level during this time. This points out occupants' energy related behaviour issues which ultimately cause misuse of electricity by the lighting system, mechanical ventilation, and major electrical equipment such as printers, photocopiers and kitchen appliances. The energy for heating and cooling dominates the energy for electrical appliances and so the overall effect is that the total energy consumption between pre-COVID and COVID periods did not differ significantly. Thus, although the modelling results slightly improved when occupancy data was considered as an explanatory variable, it was difficult to answer the research question based on this case study and further research is required to draw a a comprehensive conclusion on that.

5.2. Limitations and Future Research Directions

Despite the fact that this research complements existing baseline modelling studies, investigating the influence of occupancy data on the baseline modelling performance adds originality to this

study. Limitations still exist, which gives direction to further studies. Because of not having occupancy count data before the training period used in this study, full one-year data could not be used to train the model. Given the limited data time frame, it was not possible to evaluate the performance of the models using different time frame data and determine the optimum size of training data. Also, a typical institutional building is considered in this study where occupancy-based sensor technologies are not present. Therefore, a comprehensive conclusion could not be drawn from this study. Even so, it is expected that the outcome of this comparative study of the modelling approaches will shed light on the decision-making process of the building practitioners in selecting suitable baseline models in the real-world applications.

Therefore, future research opportunities that can be accomplished in line of the current research are identified:

- The assessment of baseline model predictive performance highlights the necessity of further studies to enquire into the effect of seasonal variations on the energy use in a building. In relation to this, the training and prediction periods could be varied, and model predictions of hourly, daily, and monthly energy use data could be compared to meter data to determine the model predictive accuracy.
- Based on the evaluation of hourly energy and associated occupancy data it seems to be occupancy data can considerably influence the performance of a baseline energy use model in an occupant-centric building. In this instance, further studies are required to realize the need for implementing occupant-centric building control strategies to improve the energy performance in a building. A large scale analysis in relation to automatically benchmarking energy use in buildings could be performed, including buildings with intelligent control

systems (such as DCV, occupancy-centric zone-level ventilation, setbacks following zone-level occupancy, occupancy presence-based lighting control) and traditional buildings.

References

[1] H.-S. Ryu and K.-S. Park, "A study on the LEED energy simulation process using BIM," *Sustainability,* vol. 8, no. 2, p. 138, 2016.

[2] F. Fuerst, "Building momentum: An analysis of investment trends in LEED and Energy Star-certified properties," *Journal of Retail & Leisure Property,* vol. 8, no. 4, pp. 285-297, 2009.

[3] H. M. Alshuwaikhat and I. Abubakar, "An integrated approach to achieving campus sustainability: assessment of the current campus environmental management practices," *Journal of cleaner production,* vol. 16, no. 16, pp. 1777-1785, 2008.

[4] E. Delzendeh, S. Wu, A. Lee, and Y. Zhou, "The impact of occupants' behaviours on building energy analysis: A research review," *Renewable and Sustainable Energy Reviews,* vol. 80, pp. 1061-1071, 2017/12/01/ 2017, doi: https://doi.org/10.1016/j.rser.2017.05.264.

[5] C. A. Ogunbode, R. Doran, and G. Böhm, "Exposure to the IPCC special report on 1.5 C global warming is linked to perceived threat and increased concern about climate change," *Climatic Change,* vol. 158, no. 3, pp. 361-375, 2020.

[6] H. de Coninck *et al.*, "Strengthening and implementing the global response," 2018.

[7] T. Abergel, B. Dean, and J. Dulac, "Towards a zero-emission, efficient, and resilient buildings and construction sector: Global Status Report 2017," *UN Environment and International Energy Agency: Paris, France,* vol. 22, 2017.

[8] E. P. Act, "Public law 109-58: 109th Congress: An act to ensure jobs for our future with secure, affordable, and reliable energy," *Washington, DC: EPA,* 2005.

[9] T. Walter, P. N. Price, and M. D. Sohn, "Uncertainty estimation improves energy measurement and verification procedures," *Applied Energy,* vol. 130, pp. 230-236, 2014.

[10] X. Xia and J. Zhang, "Mathematical description for the measurement and verification of energy efficiency improvement," *Applied Energy,* vol. 111, pp. 247-256, 2013.

[11] I. Committee, "International Performance Measurement and Verification Protocol: Concepts and options for determining energy and water savings, Volume I," National Renewable Energy Lab., Golden, CO (US), 2001.

[12] S. J. Hansen and J. W. Brown, *Investment grade energy audit*. CRC Press, 2003.

[13] ASHRAE, "ASHRAE Guideline 14-2002: measurement of energy and demand savings," *ASHRAE Guide*, vol. 8400, pp. 1-165, 2002.

[14] X. Feng, D. Yan, and T. Hong, "Simulation of occupancy in buildings," *Energy and Buildings*, vol. 87, pp. 348-359, 2015.

[15] D. Yan *et al.*, "Occupant behavior modeling for building performance simulation: Current state and future challenges," *Energy and Buildings*, vol. 107, pp. 264-278, 2015.

[16] T. Hong, S. C. Taylor-Lange, S. D'Oca, D. Yan, and S. P. Corgnati, "Advances in research and applications of energy-related occupant behavior in buildings," *Energy and buildings*, vol. 116, pp. 694-702, 2016.

[17] B. Güçyeter and H. M. Günaydın, "Optimization of an envelope retrofit strategy for an existing office building," *Energy and Buildings*, vol. 55, pp. 647-659, 2012.

[18] T. Dixon, P. McNamara, E. Miller, and L. Buys, "Retrofitting commercial office buildings for sustainability: tenants' perspectives," *Journal of Property Investment & Finance*, 2008.

[19] F. Ardente, M. Beccali, M. Cellura, and M. Mistretta, "Energy and environmental benefits in public buildings as a result of retrofit actions," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 1, pp. 460-470, 2011.

[20] R. Jagarajan, M. N. Abdullah Mohd Asmoni, A. H. Mohammed, M. N. Jaafar, J. Lee Yim Mei, and M. Baba, "Green retrofitting – A review of current status, implementations and challenges," *Renewable and Sustainable Energy Reviews*, vol. 67, pp. 1360-1368, 2017/01/01/ 2017, doi: https://doi.org/10.1016/j.rser.2016.09.091.

[21] H. Yoshino, T. Hong, and N. Nord, "IEA EBC annex 53: Total energy use in buildings—Analysis and evaluation methods," *Energy and Buildings*, vol. 152, pp. 124-136, 2017.

[22] A. Heydarian *et al.*, "What drives our behaviors in buildings? A review on occupant interactions with building systems from the lens of behavioral theories," *Building and Environment*, vol. 179, p. 106928, 2020.

[23] S. Chen, W. Yang, H. Yoshino, M. D. Levine, K. Newhouse, and A. Hinge, "Definition of occupant behavior in residential buildings and its application to behavior analysis in case studies," *Energy and Buildings*, vol. 104, pp. 1-13, 2015.

[24] K. Schakib-Ekbatan, F. Z. Çakıcı, M. Schweiker, and A. Wagner, "Does the occupant behavior match the energy concept of the building? – Analysis of a German naturally

ventilated office building," *Building and Environment,* vol. 84, pp. 142-150, 2015/01/01/ 2015, doi: https://doi.org/10.1016/j.buildenv.2014.10.018.

[25] P. van den Brom, A. Meijer, and H. Visscher, "Performance gaps in energy consumption: household groups and building characteristics," *Building Research & Information,* vol. 46, no. 1, pp. 54-70, 2018.

[26] D. Calì, T. Osterhage, R. Streblow, and D. Müller, "Energy performance gap in refurbished German dwellings: Lesson learned from a field test," *Energy and Buildings,* vol. 127, pp. 1146-1158, 2016.

[27] K. Gram-Hanssen and S. Georg, "Energy performance gaps: Promises, people, practices," ed: Taylor & Francis, 2018.

[28] A. L. Pisello and F. Asdrubali, "Human-based energy retrofits in residential buildings: A cost-effective alternative to traditional physical strategies," *Applied Energy,* vol. 133, pp. 224-235, 2014.

[29] N. Li, G. Calis, and B. Becerik-Gerber, "Measuring and monitoring occupancy with an RFID based system for demand-driven HVAC operations," *Automation in construction,* vol. 24, pp. 89-99, 2012.

[30] A. Chonga, G. Augenbroeb, and D. Yanc, "Occupancy data at different spatial resolutions: building energy performance and model calibration."

[31] S. Hu, D. Yan, E. Azar, and F. Guo, "A systematic review of occupant behavior in building energy policy," *Building and Environment,* vol. 175, p. 106807, 2020.

[32] Y. Zhang, Z. O'Neill, B. Dong, and G. Augenbroe, "Comparisons of inverse modeling approaches for predicting building energy performance," *Building and Environment,* vol. 86, pp. 177-190, 2015.

[33] H. Burak Gunay, W. Shen, G. Newsham, and A. Ashouri, "Detection and interpretation of anomalies in building energy use through inverse modeling," *Science and Technology for the Built Environment,* vol. 25, no. 4, pp. 488-503, 2019.

[34] J. Granderson and P. N. Price, "Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models," *Energy,* vol. 66, pp. 981-990, 2014.

[35] X. Liang, T. Hong, and G. Q. Shen, "Improving the accuracy of energy baseline models for commercial buildings with occupancy data," *Applied Energy*, vol. 179, pp. 247-260, 2016/10/01/ 2016, doi: https://doi.org/10.1016/j.apenergy.2016.06.141.

[36] J. L. Mathieu, P. N. Price, S. Kiliccote, and M. A. Piette, "Quantifying changes in building electricity use, with application to demand response," *IEEE Transactions on Smart Grid*, vol. 2, no. 3, pp. 507-518, 2011.

[37] A. Golden, K. Woodbury, J. Carpenter, and Z. O'Neill, "Change point and degree day baseline regression models in industrial facilities," *Energy and Buildings*, vol. 144, pp. 30-41, 2017/06/01/ 2017, doi: https://doi.org/10.1016/j.enbuild.2017.03.024.

[38] Y. Heo and V. M. Zavala, "Gaussian process modeling for measurement and verification of building energy savings," *Energy and Buildings*, vol. 53, pp. 7-18, 2012/10/01/ 2012, doi: https://doi.org/10.1016/j.enbuild.2012.06.024.

[39] S. Touzani, J. Granderson, D. Jump, and D. Rebello, "Evaluation of methods to assess the uncertainty in estimated energy savings," *Energy and Buildings*, vol. 193, pp. 216-225, 2019.

[40] J. Carpenter, K. A. Woodbury, and Z. O'Neill, "Using change-point and Gaussian process models to create baseline energy models in industrial facilities: A comparison," *Applied Energy*, vol. 213, pp. 415-425, 2018.

[41] X. Liang, T. Hong, G. Q. Shen, and B. C. A. Lawrence Berkeley National Lab, "Improving the accuracy of energy baseline models for commercial buildings with occupancy data," *Applied energy*, vol. 179, no. C, pp. 247-260, 2016, doi: 10.1016/j.apenergy.2016.06.141.

[42] C. Bianchi and A. D. Smith, "Localized Actual Meteorological Year File Creator (LAF): A tool for using locally observed weather data in building energy simulations," *SoftwareX*, vol. 10, p. 100299, 2019.

[43] E. L. Vine and J. A. Sathaye, "The monitoring, Evaluation, reporting, verification, and certification of energy-efficiency projects," *Mitigation and adaptation strategies for global change*, vol. 5, no. 2, pp. 189-216, 2000, doi: 10.1023/A:1009606005358.

[44] J. G. S. F. E. C. S. Earni, "IPMVP's Snapshot on Advanced Measurement & Verification," Efficiency Valuation Organization. [Online]. Available: http://evo-world.org/images/corporate_documents/NRE-NRA_White_Paper_Final_2701.pdf

[45] C. V. Gallagher, K. Leahy, P. O'Donovan, K. Bruton, and D. T. O'Sullivan, "Development and application of a machine learning supported methodology for measurement and verification (M&V) 2.0," *Energy and Buildings,* vol. 167, pp. 8-22, 2018.

[46] G. R. Ruiz, C. F. Bandera, T. G.-A. Temes, and A. S.-O. Gutierrez, "Genetic algorithm for building envelope calibration," *Applied energy,* vol. 168, pp. 691-705, 2016.

[47] P. A. Mathew, L. N. Dunn, M. D. Sohn, A. Mercado, C. Custudio, and T. Walter, "Big-data for building energy performance: Lessons from assembling a very large national database of building energy use," *Applied Energy,* vol. 140, pp. 85-93, 2015.

[48] J.-H. Ko, D.-S. Kong, and J.-H. Huh, "Baseline building energy modeling of cluster inverse model by using daily energy consumption in office buildings," *Energy and Buildings,* vol. 140, pp. 317-323, 2017.

[49] S. Park, S. Ryu, Y. Choi, J. Kim, and H. Kim, "Data-driven baseline estimation of residential buildings for demand response," *Energies,* vol. 8, no. 9, pp. 10239-10259, 2015.

[50] A. Srivastav, A. Tewari, and B. Dong, "Baseline building energy modeling and localized uncertainty quantification using Gaussian mixture models," *Energy and Buildings,* vol. 65, pp. 438-447, 2013/10/01/ 2013, doi: https://doi.org/10.1016/j.enbuild.2013.05.037.

[51] A. Kusiak, M. Li, and Z. Zhang, "A data-driven approach for steam load prediction in buildings," *Applied Energy,* vol. 87, no. 3, pp. 925-933, 2010.

[52] M. Kavgic, A. Mavrogianni, D. Mumovic, A. Summerfield, Z. Stevanovic, and M. Djurovic-Petrovic, "A review of bottom-up building stock models for energy consumption in the residential sector," *Building and Environment,* vol. 45, no. 7, pp. 1683-1697, 2010/07/01/ 2010, doi: https://doi.org/10.1016/j.buildenv.2010.01.021.

[53] Y. Heo, R. Choudhary, and G. Augenbroe, "Calibration of building energy models for retrofit analysis under uncertainty," *Energy and Buildings,* vol. 47, pp. 550-560, 2012.

[54] Z. Afroz, G. Shafiullah, T. Urmee, and G. Higgins, "Modeling techniques used in building HVAC control systems: A review," *Renewable and sustainable energy reviews,* vol. 83, pp. 64-84, 2018.

[55] Z. O'Neill and B. Eisenhower, "Leveraging the analysis of parametric uncertainty for building energy model calibration," in *Building simulation*, 2013, vol. 6, no. 4: Springer, pp. 365-377.

[56] A. H. F. Ashrae and G. Atlanta, "American society of Heating," *Refrigerating and Air-Conditioning Engineers*, vol. 1, 2009.

[57] B. Grillone, S. Danov, A. Sumper, J. Cipriano, and G. Mor, "A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings," *Renewable and Sustainable Energy Reviews*, vol. 131, p. 110027, 2020/10/01/ 2020, doi: https://doi.org/10.1016/j.rser.2020.110027.

[58] Z. O'Neill and C. O'Neill, "Development of a probabilistic graphical model for predicting building energy performance," *Applied energy*, vol. 164, pp. 650-658, 2016.

[59] Z. Afroz, H. B. Gunay, W. O'Brien, G. Newsham, and I. Wilton, "An inquiry into the capabilities of baseline building energy modelling approaches to estimate energy savings," *Energy and Buildings*, vol. 244, p. 111054, 2021.

[60] H. E. Solutions, "Energy and Facility Renewal Report for Carleton University Rebertson Hall," 2012.

[61] H. E. Solutions, "Energy and Facility Renewal Report," April 2011.

[62] A. ASHRAE, "Ashrae guideline 14: measurement of energy and demand savings," *American Society of Heating, Refrigerating and Air-Conditioning Engineers*, vol. 35, pp. 41-63, 2002.

[63] Z. Afroz, H. B. Gunay, W. O'Brien, G. Newsham, and I. Wilton, "An inquiry into the capabilities of baseline building energy modelling approaches to estimate energy savings," *Energy and Buildings*, p. 111054, 2021.

[64] D. Lixing, L. Jinhu, L. Xuemei, and L. Lanlan, "Support vector regression and ant colony optimization for HVAC cooling load prediction," in *2010 International Symposium on Computer, Communication, Control and Automation (3CA)*, 2010, vol. 1: IEEE, pp. 537-541.

[65] S. Seyedzadeh, F. P. Rahimian, I. Glesk, and M. Roper, "Machine learning for estimation of building energy consumption and performance: a review," *Visualization in Engineering*, vol. 6, no. 1, pp. 1-20, 2018.

[66] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[67] S. Haykin, "Neural networks: a comprehensive foundation. 1999," *Mc Millan, New Jersey*, pp. 1-24, 2010.

[68] M. H. Beale, M. T. Hagan, and H. B. Demuth, "Neural network toolbox™ user's guide," in *R2012a, The MathWorks, Inc., 3 Apple Hill Drive Natick, MA 01760-2098,, www. mathworks. com*, 2012: Citeseer.

[69] Z. Afroz, G. Shafiullah, T. Urmee, and G. Higgins, "Prediction of indoor temperature in an institutional building," *Energy Procedia*, vol. 142, pp. 1860-1866, 2017.

[70] M. Sugiyama, *Introduction to statistical machine learning*. Morgan Kaufmann, 2015.

[71] B. W. Hobson, D. Lowcay, H. B. Gunay, A. Ashouri, and G. R. Newsham, "Opportunistic occupancy-count estimation using sensor fusion: A case study," *Building and environment*, vol. 159, p. 106154, 2019.

[72] A. Guideline, "Guideline 14-2014," *Measurement of energy, demand, and water savings*, 2014.

[73] Z. Afroz, T. Urmee, G. M. Shafiullah, and G. Higgins, "Real-time prediction model for indoor temperature in a commercial building," *Applied Energy*, vol. 231, pp. 29-53, 2018/12/01/ 2018, doi: https://doi.org/10.1016/j.apenergy.2018.09.052.

[74] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning from data*: Springer, 1996, pp. 199-206.

[75] A. Marcano-Cedeno, J. Quintanilla-Domínguez, M. Cortina-Januchs, and D. Andina, "Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network," in *IECON 2010-36th annual conference on IEEE industrial electronics society*, 2010: IEEE, pp. 2845-2850.

[76] Z. Afroz, G. Shafiullah, T. Urmee, M. Shoeb, and G. Higgins, "Predictive modelling and optimization of HVAC systems using neural network and particle swarm optimization algorithm," *Building and Environment*, vol. 209, p. 108681, 2022.

[77] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella, "The'K'in K-fold Cross Validation," in *ESANN*, 2012, pp. 441-446.

[78] *Regression for M&V: Reference Guide*, May 2012.

[79] R. C. Sonderegger, "A baseline model for utility bill analysis using both weather and non-weather-related variables," *Transactions-American Society of Heating Refrigerating and Air Conditioning Engineers*, vol. 104, pp. 859-870, 1998.

[80] J. Y. Park *et al.*, "A critical review of field implementations of occupant-centric building controls," *Building and Environment*, vol. 165, p. 106351, 2019.

[81] W. O'Brien and H. B. Gunay, "Do building energy codes adequately reward buildings that adapt to partial occupancy?," *Science and Technology for the Built Environment*, vol. 25, no. 6, pp. 678-691, 2019.

[82] B. Liu, M. Rosenberg, and R. Athalye, "National Impact of ANSI/ASHRAE/IES Standard 90.1-2016," in *2018 Building Performance Analysis Conference and SimBuild*, 2018, pp. 45-52.

[83] P. Sollich, "Can Gaussian process regression be made robust against model mismatch?," in *International Workshop on Deterministic and Statistical Methods in Machine Learning*, 2004: Springer, pp. 199-210.

[84] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802-813, 2008.

[85] P. Bühlmann, "Bagging, boosting and ensemble methods," in *Handbook of Computational Statistics*: Springer, 2012, pp. 985-1022.

[86] M. Kuss, "Gaussian process models for robust regression, classification, and reinforcement learning," echnische Universität Darmstadt Darmstadt, Germany, 2006.

[87] Z. Afroz, G. M. Shafiullah, T. Urmee, and G. Higgins, "Modeling techniques used in building HVAC control systems: A review," *Renewable and Sustainable Energy Reviews*, vol. 83, pp. 64-84, 2018/03/01/ 2018, doi: https://doi.org/10.1016/j.rser.2017.10.044.

[88] H. Gunay, W. Shen, G. Newsham, and A. Ashouri, "Detection and interpretation of anomalies in building energy use through inverse modeling," *Science and Technology for the Built Environment*, vol. 25, no. 4, pp. 488-503, 2019.

Appendix A

Table 1 A. Detailed feature selection results in each iteration step with ensembles of trees (including Wi-Fi data) for cooling energy use

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	**22.61**	**12.53**	**0.82**
	RH	49.16	41.68	0.16
	SR	49.75	43.40	0.14
	WS	53.32	47.68	0.02
	WD	51.90	45.55	0.07
	Wi-Fi	50.91	44.08	0.1
	HH:MM	52.70	47.53	0.04
2	**T+RH**	**20.15**	**10.84**	**0.86**
	T+HH:MM	21.61	11.71	0.84
	T+Wi-Fi	22.00	11.97	0.83
	T+SR	21.50	11.71	0.84
	T+WS	21.93	11.89	0.83
	T+WD	22.34	12.33	0.83
3	**T+RH+Wi-Fi**	**19.70**	**10.49**	**0.87**
	T+RH+HH:MM	19.88	10.64	0.86
	T+RH+SR	19.95	10.68	0.86
	T+RH+WS	19.81	10.62	0.86
	T+RH+WD	20.11	10.86	0.86
4	T+RH+Wi-Fi+SR	19.47	10.41	0.87
	T+RH+Wi-Fi+HH:MM	19.51	10.36	0.87
	T+RH+Wi-Fi +WS	**19.39**	**10.35**	**0.87**
	T+RH+WiFi +WD	19.70	10.50	0.87
5	**T+RH+Wi-Fi +WS+SR**	**19.28**	**10.24**	**0.87**
	T+RH+Wi-Fi +WS+ HH:MM	19.29	10.27	0.87

	T+RH+WiFi +WS+WD	19.36	10.33	0.87
6	**T+RH+Wi-Fi +WS+SR+ HH:MM**	**19.23**	**10.24**	**0.87**
	T+RH+WiFi +WS+SR+WD	19.27	10.25	0.87
7	**T+RH+Wi-Fi+WS+SR+HH:MM+WD**	**19.22**	**10.23**	**0.87**

Table 2 A. Detailed feature selection results in each iteration step with ensembles of trees (excluding Wi-Fi data) for cooling energy use

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	**22.61**	**12.53**	**0.82**
	RH	49.16	41.68	0.16
	SR	49.75	43.40	0.14
	WS	53.32	47.68	0.02
	WD	51.90	45.55	0.07
	HH:MM	52.70	47.53	0.04
2	T+RH	**20.15**	**10.84**	**0.86**
	T+HH:MM	21.61	11.71	0.84
	T+SR	21.50	11.71	0.84
	T+WS	21.93	11.89	0.83
	T+WD	22.34	12.33	0.83
3	T+RH+HH:MM	19.88	10.64	0.86
	T+RH+SR	19.95	10.68	0.86
	T+RH+WS	**19.81**	**10.62**	**0.86**
	T+RH+WD	20.11	10.86	0.86
4	T+RH+WS +SR	19.67	10.52	0.87
	T+RH+WS +HH:MM	**19.63**	**10.51**	**0.87**
	T+RH+WS +WD	19.74	10.60	0.87
5	T+RH+WS +HH:MM +SR	19.64	10.51	0.87
	T+RH+WS +HH:MM +WD	**19.61**	**10.51**	**0.87**
6	**T+RH+WS +HH:MM +WD+SR**	**19.60**	**10.48**	**0.87**

Table 3 A. Detailed feature selection results in each iteration step with support vector machine (including Wi-Fi data) for cooling energy use

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	**25.61**	**18.04**	**0.77**
	RH	56.75	40.92	0.12
	SR	57.40	43.32	0.14
	WS	64.62	43.56	-0.45
	WD	64.63	43.57	-0.45
	Wi-Fi	54.95	38.72	0.2
	HH:MM	59.79	43.43	-0.25
2	T+RH	**23.76**	**17.16**	**0.80**
	T+HH:MM	25.25	18.39	0.78
	T+Wi-Fi	25.59	18.61	0.77
	T+SR	24.77	17.99	0.79
	T+WS	24.97	18.12	0.78
	T+WD	25.56	18.58	0.77
3	T+RH+Wi-Fi	23.75	17.14	0.80
	T+RH+HH:MM	23.69	17.05	0.81
	T+RH+SR	23.71	17.11	0.81
	T+RH+WS	**23.47**	**16.92**	**0.81**
	T+RH+WD	23.74	17.17	0.80
4	T+RH +WS + HH:MM	23.46	16.90	0.81
	T+RH +WS +SR	23.43	16.89	0.81
	T+RH +WS +WD	**23.39**	**16.89**	**0.81**
	T+RH +WS+ Wi-Fi	23.41	16.84	0.81
5	**T+RH +WS +WD + Wi-Fi**	**23.31**	**16.82**	**0.81**
	T+RH +WS +WD + HH:MM	23.38	16.88	0.81
	T+RH +WS +WD + SR	23.36	16.87	0.81

6	T+RH +WS +WD + Wi-Fi +HH:MM	23.27	16.80	0.81
	T+RH +WS+WD + Wi-Fi +SR	**23.14**	**16.62**	**0.81**
7	**T+RH+WS+WD+WiFi+SR+HH:MM**	**23.12**	16.58	0.82

Table 4 A. Detailed feature selection results in each iteration step with support vector machine (excluding Wi-Fi data) for cooling energy use

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	**25.61**	**18.04**	**0.77**
	RH	56.75	40.92	0.12
	SR	57.40	43.32	0.14
	WS	64.62	43.56	-0.45
	WD	64.63	43.57	-0.45
	HH:MM	59.79	43.43	-0.25
2	T+RH	**23.76**	**17.16**	**0.80**
	T+HH:MM	25.25	18.39	0.78
	T+SR	24.77	17.99	0.79
	T+WS	24.97	18.12	0.78
	T+WD	25.56	18.58	0.77
3	T+RH+HH:MM	23.69	17.05	0.81
	T+RH+SR	23.71	17.11	0.81
	T+RH+WS	**23.47**	**16.92**	**0.81**
	T+RH+WD	23.74	17.17	0.80
4	T+RH +WS + HH:MM	23.46	16.90	0.81
	T+RH +WS +SR	23.43	16.89	0.81
	T+RH +WS +WD	**23.39**	**16.89**	**0.81**
5	T+RH +WS +WD + HH:MM	23.38	16.88	0.81
	T+RH +WS +WD + SR	**23.36**	**16.87**	**0.81**
6	**T+RH +WS +WD + SR +HH:MM**	**23.33**	**16.82**	**0.81**

Table 5 A. Detailed feature selection results in each iteration step with Gaussian Process (including Wi-Fi data) for cooling energy use

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	**22.37**	**12.25**	**0.83**
	RH	48.96	40.37	0.17
	SR	49.69	43.61	0.14
	WS	53.32	48.17	0.01
	WD	51.78	45.82	0.07
	Wi-Fi	50.94	44.26	0.1
	HH:MM	52.62	47.77	0.04
2	T+RH	**19.98**	**10.59**	**0.86**
	T+HH:MM	21.46	11.68	0.84
	T+Wi-Fi	21.90	11.86	0.83
	T+SR	21.41	11.66	0.84
	T+WS	21.82	11.93	0.84
	T+WD	22.14	12.20	0.83
3	T+RH+Wi-Fi	19.54	10.14	0.87
	T+RH+HH:MM	19.52	10.38	0.87
	T+RH+SR	19.96	10.57	0.87
	T+RH+WS	19.57	10.54	0.87
	T+RH+WD	**19.47**	**10.48**	**0.87**
4	T+RH +WD + HH:MM	19.56	10.56	0.87
	T+RH +WD +SR	19.62	10.52	0.87
	T+RH +WD +WS	19.12	10.53	0.87
	T+RH +WD + Wi-Fi	**19.07**	**10.56**	**0.87**
5	**T+RH +WD + Wi-Fi +HH:MM**	**18.71**	**10.21**	**0.88**
	T+RH +WD + Wi-Fi +WS	18.76	10.36	0.88
	T+RH +WD + Wi-Fi +SR	18.87	10.13	0.88
6	**T+RH +WD + Wi-Fi +HH:MM +WS**	**18.39**	**10.42**	**0.88**

	T+RH +WD + Wi-Fi +HH:MM +SR	18.86	10.38	0.88
7	T+RH+WD+WiFi+HH:MM+WS+SR	18.43	10.58	0.88

Table 6 A. Detailed feature selection results in each iteration step with Gaussian Process (excluding Wi-Fi data) for cooling energy use

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	**22.37**	**12.25**	**0.83**
	RH	48.96	40.37	0.17
	SR	49.69	43.61	0.14
	WS	53.32	48.17	0.01
	WD	51.78	45.82	0.07
	HH:MM	52.62	47.77	0.04
2	T+RH	**19.98**	**10.59**	**0.86**
	T+HH:MM	21.46	11.68	0.84
	T+SR	21.41	11.66	0.84
	T+WS	21.82	11.93	0.84
	T+WD	22.14	12.20	0.83
3	T+RH+HH:MM	19.52	10.38	0.87
	T+RH+SR	19.96	10.57	0.87
	T+RH+WS	19.57	10.54	0.87
	T+RH+WD	**19.47**	**10.48**	**0.87**
4	T+RH +WD + HH:MM	19.56	10.56	0.87
	T+RH +WD +SR	19.62	10.52	0.87
	T+RH +WD +WS	**19.12**	**10.53**	**0.87**
5	**T+RH +WD +WS +HH:MM**	**18.82**	**10.66**	**0.88**
	T+RH +WD +WS +SR	18.99	10.57	0.88
6	T+RH +WD +WS +HH:MM+SR	19.00	10.83	0.87

Table 7 A. Detailed feature selection results in each iteration step with ensembles of trees (including Wi-Fi data) for electricity

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	17.07	12.30	0.80
	RH	34.10	29.49	0.19
	SR	33.65	29.58	0.22
	WS	38.04	34.21	0.01
	WD	36.96	32.81	0.05
	Wi-Fi	34.76	31.28	0.16
	HH:MM	36.37	33.29	0.08
2	T+RH	16.28	11.53	0.82
	T+HH:MM	15.97	11.27	0.82
	T+Wi-Fi	**15.16**	**10.00**	**0.84**
	T+SR	16.09	11.26	0.82
	T+WS	16.74	12.00	0.81
	T+WD	16.87	12.13	0.8
3	T+Wi-Fi +RH	**14.19**	**9.31**	**0.86**
	T+Wi-Fi +HH:MM	14.69	9.61	0.85
	T+Wi-Fi +SR	14.43	9.44	0.86
	T+Wi-Fi +WS	14.79	9.78	0.85
	T+Wi-Fi +WD	15.08	9.94	0.84
4	**T+Wi-Fi +RH +SR**	**13.78**	**9.01**	**0.87**
	T+Wi-Fi +RH +HH:MM	13.97	9.16	0.86
	T+Wi-Fi +RH +WS	13.92	9.17	0.87
	T+Wi-Fi +RH +WD	14.13	9.28	0.86
5	**T+Wi-Fi +RH +SR +WS**	**13.58**	**8.94**	**0.87**
	T+Wi-Fi +RH +SR + HH:MM	13.77	8.99	0.87
	T+Wi-Fi +RH +SR +WD	13.76	9.03	0.87
6	T+Wi-Fi +RH +SR +WS + HH:MM	13.58	8.94	0.87

	T+Wi-Fi +RH +SR +WS +WD	**13.54**	**8.90**	**0.87**
7	**T+Wi-Fi+RH+SR+WS+WD+HH:MM**	**13.41**	**8.83**	**0.88**

Table 8 A. Detailed feature selection results in each iteration step with ensembles of trees (excluding Wi-Fi data) for electricity

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	17.07	12.30	0.8
	RH	34.10	29.49	0.19
	SR	33.65	29.58	0.22
	WS	38.04	34.21	0.01
	WD	36.96	32.81	0.05
	HH:MM	36.37	33.29	0.08
2	T+RH	16.28	11.53	0.82
	T+HH:MM	**15.97**	**11.27**	**0.82**
	T+SR	16.09	11.26	0.82
	T+WS	16.74	12.00	0.81
	T+WD	16.87	12.13	0.8
3	**T+HH:MM +RH**	**15.26**	**10.72**	**0.84**
	T+HH:MM +SR	15.84	11.12	0.83
	T+HH:MM +WS	15.63	11.05	0.83
	T+HH:MM +WD	15.77	11.14	0.83
4	T+HH:MM +RH +SR	15.25	10.74	0.84
	T+HH:MM +RH +WS	**15.06**	**10.64**	**0.84**
	T+HH:MM +RH +WD	15.24	10.70	0.84
5	T+HH:MM +RH +WS+SR	15.23	10.71	0.84
	T+HH:MM +RH +WS +WD	**15.01**	**10.61**	**0.84**
6	**T+HH:MM+RH+WS+WD+SR**	**14.94**	10.56	0.85

Table 9 A. Detailed feature selection results in each iteration step with Gaussian Process (including Wi-Fi data) for electricity

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	16.47	12.12	0.81
	RH	35.74	32.08	0.11
	SR	33.53	30.37	0.22
	WS	37.88	33.63	0.01
	WD	36.70	33.17	0.07
	Wi-Fi	34.59	31.84	0.17
	HH:MM	36.08	33.79	0.10
2	T+RH	15.51	11.42	0.83
	T+HH:MM	15.22	10.20	0.84
	T+Wi-Fi	**14.51**	**9.06**	**0.85**
	T+SR	15.45	10.45	0.83
	T+WS	16.23	11.97	0.82
	T+WD	16.28	12.01	0.82
3	T+Wi-Fi + HH:MM	13.80	8.86	0.87
	T+Wi-Fi +RH	**13.31**	**8.22**	**0.88**
	T+Wi-Fi +SR	13.62	8.61	0.87
	T+Wi-Fi +WS	14.37	9.53	0.86
	T+Wi-Fi +WD	14.60	9.52	0.85
4	T+Wi-Fi +RH + HH:MM	12.83	8.17	0.89
	T+Wi-Fi +RH +SR	**12.70**	**7.99**	**0.89**
	T+Wi-Fi +RH +WS	13.09	8.60	0.88
	T+Wi-Fi +RH +WD	13.37	8.72	0.88
5	T+Wi-Fi +RH +SR +HH:MM	12.75	8.26	0.89
	T+Wi-Fi +RH +SR +WS	**12.54**	**7.94**	**0.89**
	T+Wi-Fi +RH +SR +WD	12.61	8.20	0.89
6	**T+Wi-Fi +RH +SR +WS +WD**	**12.26**	**7.89**	**0.90**

	T+Wi-Fi +RH +SR +WS +HH:MM	12.52	8.41	0.89
7	T+Wi-Fi+RH+SR+WS+WD+HH:MM	12.37	8.43	0.89

Table 10 A. Detailed feature selection results in each iteration step with Gaussian Process (excluding Wi-Fi data) for electricity

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	16.47	12.12	0.81
	RH	35.74	32.08	0.11
	SR	33.53	30.37	0.22
	WS	37.88	33.63	0.01
	WD	36.70	33.17	0.07
	HH:MM	36.08	33.79	0.1
2	T+RH	15.51	11.42	0.83
	T+HH:MM	**15.22**	**10.20**	**0.84**
	T+SR	15.45	10.45	0.83
	T+WS	16.23	11.97	0.82
	T+WD	16.28	12.01	0.82
3	**T+HH:MM +RH**	**14.31**	**9.67**	**0.86**
	T+HH:MM +SR	15.19	10.54	0.84
	T+HH:MM +WS	15.36	10.87	0.84
	T+HH:MM +WD	15.24	10.64	0.84
4	T+HH:MM +RH +SR	14.30	9.75	0.86
	T+HH:MM +RH +WS	**14.29**	**10.02**	**0.86**
	T+HH:MM +RH +WD	14.50	10.18	0.85

Table 11 A. Detailed feature selection results in each iteration step with Ensembles of trees (including Wi-Fi data) for heating energy

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	30.32	19.55	0.69

	RH	49.91	41.29	0.16
	SR	51.57	43.78	0.10
	WS	53.96	47.44	0.02
	WD	53.62	53.62	0.03
	Wi-Fi	49.33	41.49	0.18
	HH:MM	54.02	48.06	0.02
2	T+RH	29.49	18.96	0.71
	T+HH:MM	**26.06**	**15.44**	**0.77**
	T+Wi-Fi	29.00	18.17	0.72
	T+SR	27.09	16.03	0.75
	T+WS	29.82	19.09	0.70
	T+WD	30.08	19.37	0.69
3	T+HH:MM +Wi-Fi	25.99	15.31	0.77
	T+HH:MM +RH	**25.81**	**14.96**	**0.78**
	T+HH:MM +SR	25.84	15.21	0.78
	T+HH:MM +WS	25.88	15.29	0.77
	T+HH:MM +WD	26.06	15.42	0.77
4	**T+HH:MM +RH +SR**	**25.55**	**15.09**	**0.78**
	T+HH:MM +RH +Wi-Fi	25.77	15.29	0.78
	T+HH:MM +RH +WS	25.68	15.22	0.78
	T+HH:MM +RH +WD	25.86	15.35	0.77
5	T+HH:MM +RH +SR +WS	25.49	15.07	0.78
	T+HH:MM +RH +SR +Wi-Fi	**25.47**	**15.02**	**0.78**
	T+HH:MM +RH +SR +WD	25.60	15.15	0.78
6	T+HH:MM+RH+SR +Wi-Fi +WD	25.51	15.09	0.78
	T+HH:MM+RH+SR+Wi-Fi+WS	**25.40**	**15.00**	**0.78**
7	T+HH:MM+RH+SR+Wi-Fi+WS+WD	25.43	15.05	0.78

Table 12 A. Detailed feature selection results in each iteration step with Ensembles of trees (excluding Wi-Fi data) for heating energy

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	T	30.32	19.55	0.69
	RH	49.91	41.29	0.16
	SR	51.57	43.78	0.1
	WS	53.96	47.44	0.02
	WD	53.62	53.62	0.03
	HH:MM	54.02	48.06	0.02
2	T+RH	29.49	18.96	0.71
	T+HH:MM	**26.06**	**15.44**	**0.77**
	T+SR	27.09	16.03	0.75
	T+WS	29.82	19.09	0.70
	T+WD	30.08	19.37	0.69
3	**T+HH:MM +RH**	**25.81**	**14.96**	**0.78**
	T+HH:MM +SR	25.84	15.21	0.78
	T+HH:MM +WS	25.88	15.29	0.77
	T+HH:MM +WD	26.06	15.42	0.77
4	**T+HH:MM +RH +SR**	**25.55**	**15.09**	**0.78**
	T+HH:MM +RH +WS	25.68	15.22	0.78
	T+HH:MM +RH +WD	25.86	15.35	0.77
5	**T+HH:MM +RH +SR +WS**	**25.49**	**15.07**	**0.78**
	T+HH:MM +RH +SR +WD	25.60	15.15	0.78
6	**T+HH:MM+RH+WS+WD+SR**	**25.51**	15.09	0.78

Table 13 A. Detailed feature selection results in each iteration step with Gaussian Process (including Wi-Fi data) for heating energy

No. of Predictors	Predictors	CVRMSE	MAPE	R-Squared
1	**T**	**29.99**	**19.42**	**0.7**

	RH	52.27	45.04	0.08
	SR	51.78	44.32	0.1
	WS	53.81	47.79	0.02
	WD	53.62	47.39	0.03
	Wi-Fi	49.51	41.88	0.17
	HH:MM	53.91	48.40	0.02
2	T+RH	29.38	19.27	0.71
	T+HH:MM	**25.59**	**15.38**	**0.78**
	T+Wi-Fi	28.84	18.45	0.72
	T+SR	27.38	16.88	0.75
	T+WS	29.58	19.31	0.70
	T+WD	29.90	19.56	0.70
3	T+HH:MM +Wi-Fi	25.41	15.42	0.78
	T+HH:MM +RH	25.14	15.43	0.79
	T+HH:MM +SR	**24.68**	**14.88**	**0.79**
	T+HH:MM +WS	25.55	15.85	0.78
	T+HH:MM +WD	25.59	15.90	0.78
4	**T+HH:MM +SR +RH**	**24.15**	**14.68**	**0.8**
	T+HH:MM +SR +Wi-Fi	24.61	15.00	0.8
	T+HH:MM +SR +WS	24.94	15.57	0.79
	T+HH:MM +SR +WD	24.78	15.54	0.79
5	T+HH:MM +SR +RH +WS	24.10	15.02	0.80
	T+HH:MM +SR +RH +Wi-Fi	**24.04**	**14.37**	**0.80**
	T+HH:MM +SR +RH +WD	24.07	14.46	0.80
6	T+HH:MM +SR +RH +Wi-Fi+WD	24.24	15.33	0.80
	T+HH:MM +SR +RH +Wi-Fi+WS	24.08	15.08	0.80

Appendix B

Table 1 B. Tuned hyperparameter values for the selected algorithms

Regression model	***Hyperparameter***	***Hyperparameter search range***	***Tuned values***					
			Cooling energy use		**Heating energy use**		**Electricity**	
			Model formation					
			Including Wi-Fi data	Excluding Wi-Fi data	Including Wi-Fi data	Excluding Wi-Fi data	Including Wi-Fi data	Excluding Wi-Fi data
Ensembles of Trees	Ensemble method	BAG, LSBoost	Bag	Bag	Bag	Bag	Bag	Bag
	Number of learners	10-500	37	19	68	97	53	27
	Minimum leaf size	1-3038	7	1	9	3	7	5
Gaussian Process	Kernel function	Nonisotropic exponential, Nonisotropic Matern 3/2, Nonisotropic Matern 5/2, Nonisotropic Rational Quadratic, Nonisotropic Squared Exponential, Isotropic Exponential, Isotropic Matern 3/2, Isotropic Matern 5/2, Isotropic Rational Quadratic, Isotropic Squared Exponential	Nonisotropic Squared Exponential	Nonisotropic Rational Quadratic	Isotropic Exponential	Isotropic Rational Quadratic	Nonisotropic exponential	Nonisotropic Matern 5/2

	Kernel scale	1.092-1092	93.28	62.67	78.97	67.54	128.73	28.77
	Sigma	0.0001-1112.13	0.0127	0.0024	0.0091	0.0113	0.0015	0.0096
NARX Neural Network	Number of hidden neurons	5-20	13	12	10	12	13	15
	Number of time delays	1-5	3	2	2	2	2	3

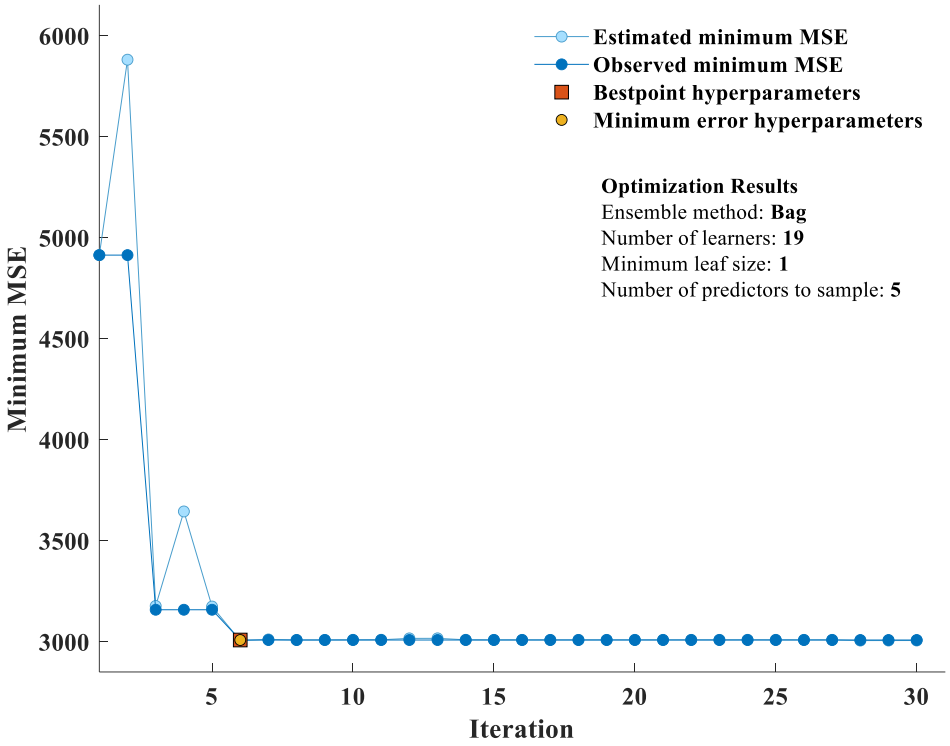


Figure B1. A Minimum MSE Plot for cooling energy use model with the Ensembles of Trees (excluding Wi-Fi data)