MEASUREMENT AND VERIFICATION BUILDING ENERGY PREDICTION (MVBEP): AN INTERPRETABLE DATA-DRIVEN MODEL DEVELOPMENT AND ANALYSIS FRAMEWORK

by

ABDURAHMAN S. ALROBAIE

B.A., King Fahd University of Petroleum and Minerals University, 2020

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Civil, Environmental, and Architectural Engineering

2022

Committee Members:

Moncef Krarti, PhD., Chair

Gregor P. Henze, PhD.

Kyri Baker, PhD

Alrobaie, Abdurahman S. (M.S., Building Energy Systems)

Measurement and Verification Building Energy Prediction (MVBEP): An Interpretable Data-Driven Model Development and Analysis Framework

Thesis directed by Prof. Moncef Krarti, PhD.

The operation of building energy systems including Heating, Ventilation, and Air Conditioning (HVAC), lighting, and equipment accounts for 85% of the global building energy consumption. With several countries pledging to achieve sustainability goals, building retrofit is becoming a crucial pillar in attaining most of the set energy efficiency targets. However, several obstacles remain that prevent retrofitting buildings to be economically feasible. An essential task in retrofitting a building is justifying the cost effectiveness of the installed Energy Conservation Measures (EMC), that is, Measurement and Verification (M&V) of the achieved energy and cost savings.

Due to the rapid development of Advanced Metering Infrastructure (AMI), data-driven approaches are becoming more effective than deterministic methods in developing M&V baseline energy models for existing buildings using historical energy consumption data. This thesis develops and applies a framework for M&V baseline data-driven modeling aimed at estimating energy savings from building retrofits. The framework employs the common data-driven modeling approaches used for building energy prediction, necessary data processing steps, and industry-used evaluation approaches. The considered modeling approaches in the developed framework include linear regression, ensemble models, support vector regression, neural networks, and kernel regression. The developed framework is applied to two case studies with different data sources of: energy consumption data generated via a simulated office building model and mostly measured energy use data collected by Building Data Genome 2 (BDG2), a publicly available dataset.

The framework application to data from over 208 buildings indicated that energy use predictions can be achieved with medians of Coefficient of Variation of Root Mean Squared Error (CV(RMSE)) that are less than 20% and %25 for daily and hourly frequency, respectively. The Median Normalized Mean Bias Error (NMBE) for these predictions has a range of $\pm|3\%|$ for all modeling approaches and data frequencies. No significant impact on prediction accuracies was observed for the building typology and the site climate.

ACKNOWLEDGEMENTS

I would like to thank Dr. Moncef Krarti for making this wonderful journey fruitful and exciting. Thank you for the motivation, assistance, and most importantly patience in guiding me. I would like to thank the people at King Fahd University of Petroleum and Minerals, especially the department of architectural engineering, for their trust in me and I hope to meet your expectations. I thank my family for their unconditional love, support, and for being close despite the great distance between us.

CONTENTS

CHAPTER

TABLES

Table

FIGURES

Figure

CHAPTER 1

INTRODUCTION

Buildings represent the sector that consumes the most energy worldwide with with a share of 35% of global energy demand exceeding the contributions of industrial activities and transportation [1]. It is estimated that 85% of the building energy consumption is attributed to Heating, Ventilation, and Air Conditioning (HVAC), lighting, and plug loads. Moreover, residential buildings account for approximately 63% of the total energy used by the building sector [1]. In terms of electricity consumption, buildings are responsible for 50% of the world's electricity consumption [1]. According to U.S. Energy Information Administration (EIA), projections show that the residential and commercial buildings will increase from 2018 to 2050 by 1.3% annually for the developed countries part of the Organization for Economic Cooperation and Development (OECD), and 2% annually for the non-OECD countries [2]. Several studies have analyzed the historical and current status of energy consumed by buildings and have projected future increases in building-related energy use globally [3] especially in China [4], the European Union [5], and Gulf Cooperation Council (GCC) countries [6]. The high energy consumption by the built environment has been shown to have significant detrimental effects on the environment and the climate. Several governmental agencies and global organizations are adopting initiatives and programs that target the reduction of energy consumption in the building sector. For instance, the U.S. Department of Energy has set a 2030 goal of tripling the 2020 energy efficiency levels for both commercial and residential buildings [7]. Similarly, the UK has developed a net-zero energy strategy for buildings so that by 2050, buildings will be completely decarbonized [8]. The goal included a plan that is driven by decisions to fund several research projects to identify innovative technologies, support owners to improve their buildings' energy efficiency, and subsidize clean and energy

efficiency projects [8]. In addition, China, the largest carbon dioxide emitter, has pledged to reach carbon neutral emissions before 2060 [9].

Such initiatives and pledges can be achieved by a combination of several approaches including enhancing renewable energy sources, setting more stringent energy efficiency regulations, and funding research to develop effective and transformative technologies in the building energy sector. However, improvements in the energy efficiency levels of existing buildings are required to attain the desired goals. Indeed, the average annual rate of replacing existing buildings is low reaching only 1% in the UK [10]. It is argued that the environmental and economic benefits of retrofitting existing buildings outweigh those achieved by replacing them with more efficient new buildings. Hasik et al. [11] performed a Life Cycle Assessment (LCA) of both retrofitted and newly constructed buildings and found that retrofitting results in a reduction ranging between 53% and 75% for over 6 different environmental impact factors compared to new construction. Economic benefits are highest when retrofitting the least energy-efficient buildings when considering aspects such as the creation of employment and reduction of carbon emissions compared to constructing new buildings [12].

Retrofitting existing buildings include renovations of mechanical, structural, and electrical systems with a range of options such as refurbishment, replacement, or addition of new equipment. In the case of energy efficiency retrofits, the replacement and addition of new equipment are usually referred to as Energy Conservation Measures (ECMs). Several ECMs can be considered for existing buildings such as changing HVAC equipment, lighting systems, and envelope features such as glazing types and wall assemblies. The deployment of ECM aims primarily at reducing the energy use and cost of the buildings. The required investments for ECMs are justified based on economic and environmental benefits. However, the implementation of ECMs can face several challenges, especially during assessment and identification as well as installation and verification. In the first period, any missing information and documentation can hinder good assessment of the existing building's energy performance and thus effective identification. Similarly, uncertainty and lack of data can affect the installation and validation process. For the validation analysis, an energy model of the building is typically needed to predict the energy use before the deployment of any ECM with minimum prediction uncertainty. Additionally, this process includes an essential step in justifying the effectiveness of the installed EMCs, that is, Measurement and Verification (M&V) analysis.

Although the energy and cost benefits of retrofitting existing buildings are promising, several challenges remain for accurate M&V analysis to estimate these benefits. Due to the rapid development in Advanced Metering Infrastructure (AMI), data-driven approaches are becoming more effective than deterministic methods in developing baseline energy models for existing buildings using historical energy consumption data. Numerous reported studies have considered data-driven models to predict the energy consumption of existing buildings. Many of the reported data-driven models are based on historical data that is used for training the models and testing their prediction accuracy. However, few of such models have been applied to create a baseline for M&V analysis to determine energy savings achieved by installed ECMs. Additionally, reported data-driven models for M&V applications have been developed and tested only for specific building types and ECMs. Furthermore, models have not been evaluated for multiple buildings and ECMs. With the increasing interest in applying data-driven models for building energy retrofit analysis, there are limited guidelines on the suitability of these models for M&V applications. Therefore, this thesis examines various methods and algorithms that have been applied to develop data-driven models for M&V analysis of building energy retrofits as shown in Chapter 2. Chapter 2 aims to build the required background literature for developing a M&V data-driven model including industry protocols, a bibliometric analysis of interest in data-driven modeling, and the necessary and most commonly used steps when developing a data-driven model for building energy prediction. Chapter 3 explains the conceptualization and creation of the proposed methodology for building a data-driven M&V baseline model. The application of the methodology is demonstrated on selected datasets along with highlighting several aspects of the proposed methodology as presented in Chapter 4. Finally, 5 lists the thesis outcomes and concludes it by discussing future work and developments to the proposed methodology.

CHAPTER 2

LITERATURE REVIEW

The literature presented in this chapter provides an extensive summary of data-driven modeling approaches for building energy prediction that are suitable for M&V applications. The presented literature review describes commonly used data-driven modeling approaches including linear regression, decision trees, ensemble methods, support vector regression, neural networks, and kernel regressions. The advantages and limitations of each data-driven modeling approach and its variants are discussed including their cited applications. Additionally, feature engineering methods used in building energy data-driven modeling are described based on reported case studies to outline commonly used features as well as the selection and processing techniques.

Section 2.1 gives an overview of the process of M&V baselining and discusses the relevant industry protocols. Section 2.2 provides a bibliometric analysis of the research interest in data-driven modeling in buildings, mostly used modeling approaches, and cases building typology. Section 2.3 introduces the mostly used categories in modeling a building's energy baseline for general prediction and M&V with an explanation of each category, a summary of relevant publications, and a discussion that combines theory with publications' conclusions. Section 2.4 explains feature engineering in the context of M&V baseline modeling along with the categories of features and feature engineering techniques. Section 2.5 discusses the mostly used data requirements for building a M&V data-driven modeling and their possible effect on training and prediction. Section 2.6 mentions existing M&V frameworks and their features, used models, and required inputs. Finally, Section 2.7 outlines relevant evaluation approaches and performance metrics often used to assess data-driven models' accuracy and goodness-of-fit.

2.1 Overview of Measurement and Verification Analysis

M&V analysis is a process of quantifying the energy use savings due to the deployment of ECMs when retrofitting existing buildings. A baseline energy model allows the prediction of the energy use of an existing building due to variations in environmental and behavioral factors such as different climatic conditions or changes in occupancy levels before any ECM implementation. The baseline energy model is often used as a benchmark to estimate energy savings due to installing one or several EMCs. Figure 2.1 shows the difference between metered and modeled energy use of an existing building over three periods. The first period corresponds to the pre-retrofit operation of the building with the energy use data being metered using historical data collected from utility bills or a Building Management System (BMS). The baseline energy model is typically developed and tested during this pre-retrofit period. During the retrofit period, ECMs are installed in the building resulting in a gradual reduction in the energy consumption compared to the predictions of the baseline energy model as noted in Figure 2.1. After completing the ECM installation phase, the building typically consumes less energy than during the pre-retrofit period as demonstrated in Figure 2.1 during the post-retrofit period. Indeed, the baseline energy model predicts higher energy consumption than the metered data during the post-retrofit period. The difference between the baseline and the metered energy consumption during the post-retrofit period represents energy savings incurred by the installed ECMs during the post-retrofit period. A building energy model baseline is often established for M&V analysis of retrofitting existing buildings especially for those with energy use historical data that meets certain requirements such as date range, missing values, or reporting frequency. These requirements, however, are not well-defined and vary from one case to another depending on a wide range of factors including the nature of occupancy. The process of constructing a data-driven baseline model requires that a building has been in operation during a sufficiently long period to gather enough data to establish correlation between its energy performance and other independent factors such as weather and occupancy parameters.

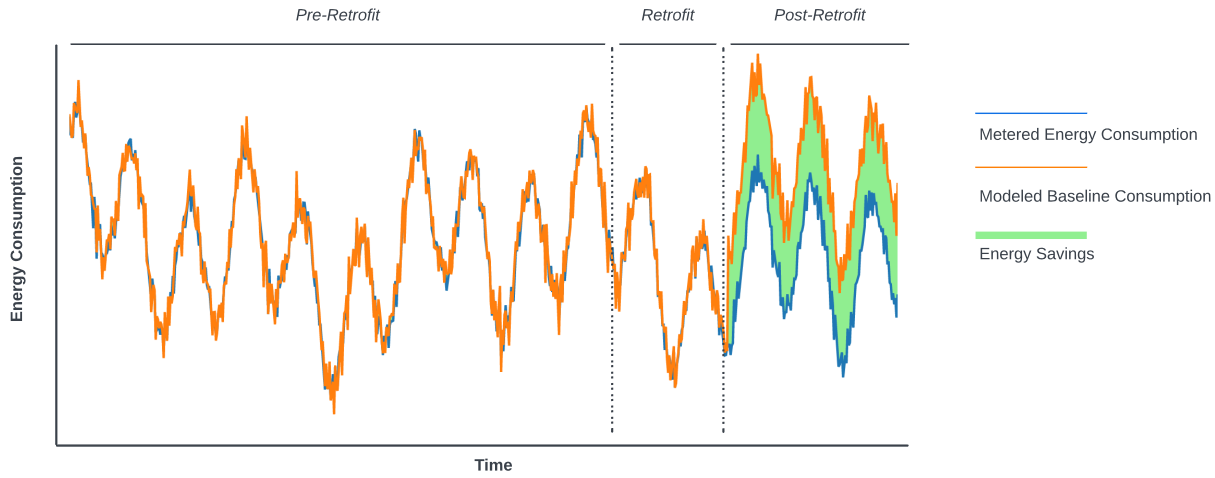


Figure 2.1: Actual energy consumption and modeled building baseline vs time

2.1.1 Measurement and Verification Protocols

Several M&V protocols have been developed to improve consistency and reduce uncertainty in estimating the energy savings attributed to retrofitting existing buildings such as the International Performance Measurement and Verification Protocol (IPMVP) [13] and ASHRAE Guideline 14 [14]. The analysis approaches outlined in these protocols differ depending on the geographical regulatory requirements, the types of ECMs, and building typologies. Additionally, specific frameworks and methodologies have been proposed to achieve the desired objectives from retrofitting projects. For instance, Ma et al. [15] developed a systematic methodology for carrying out retrofitting projects and successfully completing the various phases and analyses including the M&V analysis.

2.1.1.1 International Performance Measurement and Verification Protocol (IPMVP)

The International Performance Measurement and Verification Protocol or IPMVP is one of the most common frameworks in performing M&V analysis of retrofitting existing buildings with four evaluation options as outlined in Table 2.1. The selection of the most appropriate analysis option depends on the boundary of the deployed ECMs. Options A and B are applied when the retrofit is restricted to only one

specific and isolated building energy system. These two options differ depending on the analysis method and the availability of metered data. In particular, option A can be used for a M&V analysis for a lighting system retrofit using only key parameters including power ratings and operation schedules to calculate energy savings. On the other hand, option B is applied for systems whose energy performance can be monitored such as chillers and boilers. In addition, Options C and D can be applied when the retrofit affects the energy performance of the entire building. When metered building energy data can be collected before and after the retrofit periods, Option C is suitable for conducting the M&V analysis. When historical data of metered energy consumption are not available or are unreliable before or after the retrofit, option D is considered using calibrated energy models [13].

Table 2.1: IPMVP M&V options

M&V Options	Boundary	Parameters	Process
Option A	System	Key system parameters with estimation	Simple calculations with the estimated parameters
Option B	System	All system parameters with no estimation	More rigorous calculations with all related parameters
Option C	Whole Building	Whole building energy consumption historical data	whole building baseline modeling with building energy consumption and related parameters
Option D	Whole Building and/or System	Whole building and/or system parameters with energy bills	Calibrated simulation of the boundary using data and modeling tools

2.1.1.2 ASHRAE Guideline 14

ASHRAE has developed Guideline 14 for Measurement of Energy, Demand, and Water Savings to standardize the M&V calculations used to estimate achieved energy demand and water savings from retrofit projects. ASHRAE Guideline 14 utilizes three M&V analysis options that are similar to those specified by the IPMVP including retrofit isolation, whole facility, and whole building calibrated simulation. Instead of having two analysis options for isolated systems, ASHRAE Guideline 14 allows only one method with the flexibility of the parameters that can be used in the calculations. The whole facility option is similar to Option C of the IPMVP using the whole facility metered energy consumption along with independent

variables to establish the building's baseline energy model. The third approach is similar to the IPMVP's option D using a calibrated baseline model to quantify savings from the retrofit. While ASHRAE Guideline 14 shares some features with IPMVP, it does not cover specific details such as energy performance contracting and metering provisions like IPMVP does [14].

2.1.1.3 Advanced Measurement and Verification

Advanced M&V, or usually referred to as M&V 2.0, encompasses detailed analysis approaches using high frequency metered data (i.e., sub-hourly), and end-use loads using Advanced Metering Infrastructures (AMI) [16]. In fact, M&V 2.0 enables metered data to be more effective for building real-time performance assessment, occupant engagement, and resource management using various analysis tools and algorithms. The improvements of both hardware and software over the last decade have resulted in better accuracy in performing various M&V tasks such as developing baseline models, detecting non-routine events, and benchmarking energy consumption. Furthermore, retrieval of metered data at higher frequencies and shorter time intervals facilitates performing data analytics and automating savings quantification for retrofit projects which reduces the time lag between implementation and evaluation phases [17].

2.1.2 Baseline Modeling

Three approaches are commonly used to establish building baseline models: deterministic (also referred to as direct or white-box), data-driven (also referred to as indirect or back-box), and hybrid (also referred to as grey-box) methods. All approaches reach the same objective in M&V (i.e. constructing a baseline) with different inputs and processes. The comparison between such means for a single case is time-consuming and rarely performed as each approach has sub-approaches which alone will take more time and effort. Therefore, this subsection aims to provide a concise comparison between them.

Deterministic modeling relies on physics-based tools to predict energy consumption of buildings due to their thermal interactions with the outdoor environment. Such interactions are often represented using heat and mass balance equations that are solved using a set of algorithms that are the basis for a deterministic building energy modeling tool. There is a wide range of commercially available and open-source deterministic

modeling tools that can be utilized for developing building energy models including EnergyPlus [18], TRNSYS [19], DOE-2 [20], DesignBuilder [21], Matlab/Simulink [22], and Modelica/Dymola buildings library [23]. Most of these deterministic modeling tools require comprehensive input data about the building features such as envelope thermal properties, mechanical equipment efficiency, and operation schedules. Ke et al. [24] developed a deterministic (also referred to as white-box) baseline energy model using eQUEST software (based on DOE-2 simulation engine) for an existing office building with a Mean Bias Error (MBE) of 0.37%. The building energy model includes over 50 input variables indicating the types and operation characteristics of chillers, indoor air-conditioning units, and cooling towers performance in addition to several variables describing other building systems such as the envelope elements and lighting fixtures. The study has demonstrated high levels of interpretability in understanding the specific interactions between energy end-uses of various building systems and occupancy behaviors that deterministic building energy modeling can offer. However, the interpretability as well as the high prediction accuracy of the deterministic models come at a significant computing times and input data collection efforts.

Data-driven models represent relationships between energy performance indicators and environmental parameters identified using historical data. These relationships are then applied to predict the building response when all or some environmental variables would change. Thus, data-driven models are based on developing correlations between desired input and output parameters using various statistical and machine learning approaches. In particular, the development as well as the accuracy level of data-driven models rely heavily on historical data for both input and output variables. Types and applications of data-driven modeling are discussed in detail in Section 2.3. Typically, the accuracy and interpretability levels of data-drive models are lower than those achieved by white-box models as data is usually noisy and the occupancy behavior is not consistent.

Hybrid, also referred to as grey box, models utilize data-driven analysis approach to tune and improve physics-based (also referred to as deterministic or white-box) models through value estimations of input parameters values using historical data. A common deterministic model using in hybrid analysis approach is based on Resistance and Capacitance (RC) modeling to account for building thermal mass. Piccinini et al. [25]. developed a framework for building a hybrid modeling approach using historical monthly electricity and

natural bills of a primary school building to calibrate a building energy model using the Dymola Environment. The study achieved a Normalized Mean Bias Error (NMBE) of 1.8% while using far less parameters compared to a white-box model developed using detailed simulation tools such as EnergyPlus or TRNSYS. Similarly, Giretti et al. [26] compared the performance of reduced-order modeling using Modelica with Buildings Library against calibrated detailed models belonging to three cases: a hospital, library, and an educational building. The calibrated reduced-order models obtained a Coefficient of Variation of the Root Mean Squared Error (CV(RMSE)) between 5% and 8% compared to the detailed models while using only 25 parameters that are categorized into building envelope, heating/cooling system, occupancy, and weather components.

Chen et al. [27] compared three energy modeling approaches including black, white, and grey box models. The comparative analysis considered several performance metrics including development efforts, computational times, and analysis limitations. Their study found a trade-off between each metric category with black-box models requiring the least efforts and times while white-box models having far more input parameters. Grey-box models are in the middle in terms of development effort and required input parameters as it still requires significant data correlating energy consumption and weather variables. In terms of interpretability, white-box models allow for better understanding of the impact contributed by each input on building energy performance followed by grey-box then black-box modeling approaches. This capability is due to the fact that relationships between energy consumption and input parameters are well-established for deterministic models based basic physical principles rather than inferred from historical data as required by the data-driven (black-box) models.

2.2 Data-Driven Trend in Building Energy Modeling

Data-driven approaches have gained an increasing popularity especially in the last 10 years as more methods have been proposed for baseline modeling and energy consumption forecasting due to the higher availability of reliable building energy consumption data and advances in machine learning techniques. Indeed, significant valuable and granular data can be collected from smart metering building technologies. For instance, measured building data currently have short frequencies of 15-minute or 1-hour intervals instead of monthly frequency data like utility bills. Additionally, the energy end-use granularity allows better evalua-

tion of various building energy systems (i.e., lighting, HVAC, and appliances). Moreover, higher data storage capabilities through databases and clouds permit access to a significant amount of building performance data for various analyses and applications. Due to the aforementioned factors, advances and interests in using data-driven approaches for building energy analyses have been significantly increased in the last decade as described in the following subsections.

2.2.1 Interest in Data-Driven Approaches

To gauge the level of interest in using data-driven approaches for building energy assessments, a bibliometric analysis is performed by using Web of Science database [28]. The analysis is specific to the literature published between 2010 and 2021 focusing on technical papers that develop and apply data-driven approaches to model or predict building energy consumption. Web of Science is used to identify the number of published papers that are relevant to data-driven approaches during a specific time period. Specifically, the following keywords are considered for the Web of Science search:

- Data-driven Building Energy Modeling.
- Building Energy Prediction.
- Building Electricity Prediction.
- Machine Learning Building Energy Modeling.

Figure 2.2 shows the results of the bibliometric analysis depicting clearly the rapid and consistent increase over the last decade in published papers with a focus on data-driven approaches applied to building energy analysis.

2.2.2 Data-Driven Approaches

The bibliometric analysis is also performed to understand the most common methods used in the literature for data-driven building energy modeling. The search query of peer-reviewed papers published from 2010 to 2021 to identify the used modeling approaches was categorized into machine learning methods

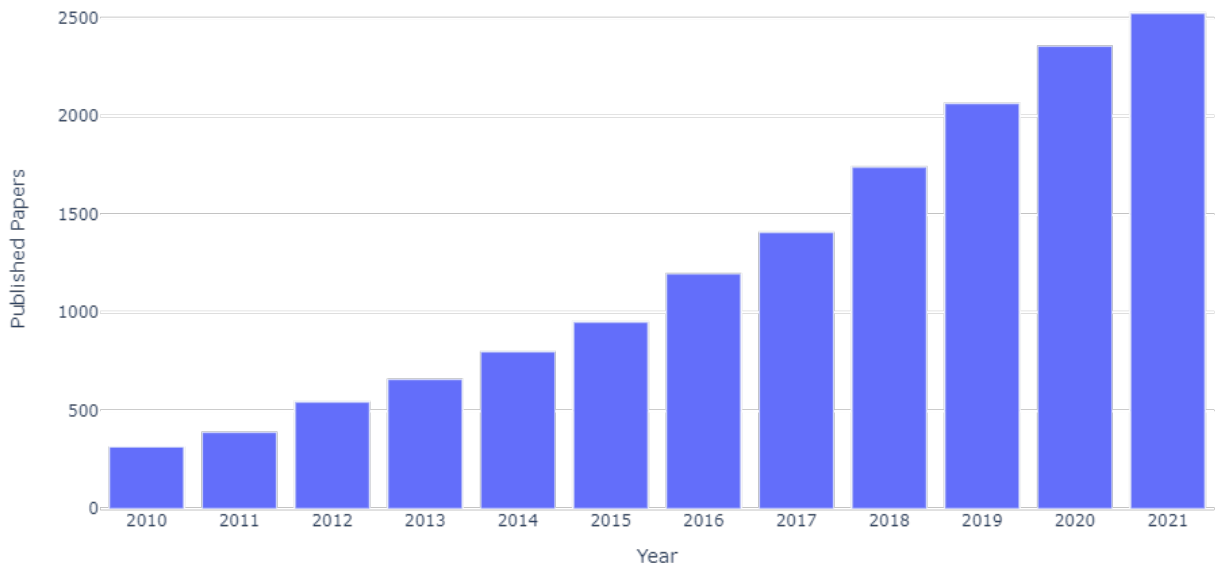


Figure 2.2: Number of research papers using data-driven approaches for building energy modeling

categories by Sklearn library [29] in Python. The models are further organized to represent relevant models since Sklearn library includes many models while the energy modeling research papers were more focused on a set of these models. However, this approach was not successful since the search query cannot determine the objective of the research paper but rather filter results based on matching keywords and similarity. This issue is amplified when the research paper address certain methods but does not actually use that method. Hence, an alternative approach is followed by using the resulted papers in the queries mentioned in Section 2.2.1 and randomly selecting papers until reaching 75 research paper. The randomness is introduced by exporting the list of papers and using Python to shuffle titles. If a certain paper does not utilize data-driven methods to model the energy consumption, the next paper on the shuffled list will be analyzed until reaching 75 papers. Figure 2.3 shows the result of this reading. The main level categories are:

- *Linear Regression (LR):* it is a category that involves linearly regressed models as described in Subsection 2.3.1.

- *Ensemble methods and Decision Tree (DT):* The two methods are grouped in one category based on

their similarity as explained in Subsection 2.3.2.

- *Support Vector Machine (SVM):* it is a modeling method of using supporting vectors to fit a hyper-plane for regression and classification as demonstrated in Subsection 2.3.3.
- *Artificial Neural Network (ANN):* it is a category that utilizes deep learning and human brain-inspired function of neurons and layers as discussed in Subsection 2.3.4.
- *Kernel regression:* it is a family of non-parametric techniques to fit changing coefficients on data points as outlined in Subsection 2.3.5.

From the main categories, the most commonly used models branch based on the frequency of occurrence in peer-reviewed papers. The plot does not branch into the specific models used in each study but rather into a general yet specifically sufficient level that can encompass the modeling approach.

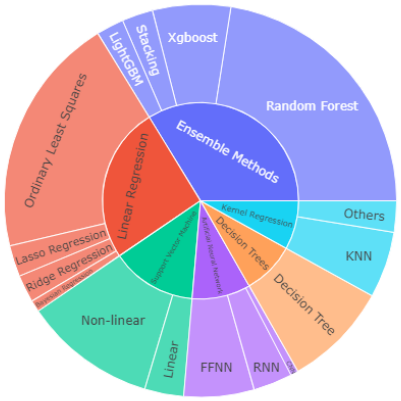


Figure 2.3: Sunburst chart of bibliometric analysis methods used in data-drive building energy modeling

2.2.3 Building Typologies

Data-driven models have been applied for several building types including commercial and residential buildings as well as individual and group of buildings. The typology can significantly impact the building energy performance due to occupant behavior and the operation of various systems. According to a study

performed by Liang et al [30] in Phoenix, Arizona using data for 636 commercial and 201 residential buildings collected from Energize Phoenix program, retrofits can save about 8% and 12% in annual energy consumption for residential and commercial buildings, respectively. Wang et al. [31] performed a comparison between different building typologies including offices, shopping malls, and educational buildings. The study concluded that shopping malls have the highest potential for energy savings, followed by multifunctional buildings and hotels. Figure 2.4 shows a sunburst chart of building typology reported in various energy modeling studies categorized using three levels including data source, building the main function, and sub-category of building type as the following:

- **Actual or metered data:** consist of energy consumption recorded using standalone measurement devices or Building Management Systems (BMS).
    - *Non-residential*: encompassing mostly commercial buildings and office spaces. Educational buildings represent cases where the building's purpose is mostly for classrooms and teaching such as schools and universities. Other buildings with commercial nature such as restaurants and retail buildings are grouped into one category.
    - *Residential:* including buildings that are used mostly for housing and living spaces. Residential buildings are divided into detached houses, apartment buildings, and other types of residential buildings.
- **Simulated or synthesized data:** are typically generated using simulation analysis tools such as EnergyPlus and DOE-2.
- **Public datasets:** are obtained from public databases such as Open Energy Data Initiative (OEDI) [32].

2.3 Data-Driven Approaches

This section outlines a brief description of each method used for data-driven modeling and the main reported applications for these methods. Each subsection discusses one of the main categories that are

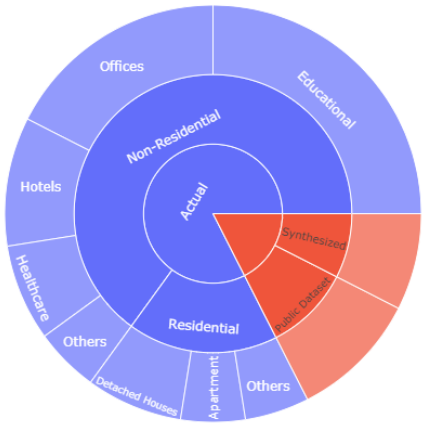


Figure 2.4: Sunburst chart of bibliometric analysis methods used in data-drive building energy modeling

mentioned in 2.2.2 with an explanation of the general algorithm, sub-models within the category, and a list of publications that utilized one or more of the category's models. Table 2.2 through Table 2.6 show summaries of such publications with a description of the applied case, data type, features, utilized category's models, and data granularity or frequency. The papers listed in Table 2.2 through Table 2.6 include data-driven modeling suitable not only for M&V analysis but also for baseline building energy development. Among the reported literature there are very limited papers that perform full M&V analysis using data-driven models as most of the reviewed applications evaluate the prediction performance of data-driven approaches. Features, predictors, and dependent variables are terms that are used interchangeably to list input parameters that are used to train the model to perform predictions about the response, target, or independent variable which represents the model's output. In each of the following sections' tables, the general category of the feature will be mentioned instead of the specific features for conciseness. In Section 2.4, the features will be explained further in terms of filtering and processing. Based on the conclusion reached by each study, Table 2.2 through Table 2.6 will show, if the results are clearly indicating one best model, the best model within the categories mentioned in this section in bold font. Data granularity represents the interval of prediction which can be 15-minute, hourly, daily, weekly, or monthly.

2.3.1 Linear Regression

2.3.1.1 Definition

LR is a term that encompasses a family of different techniques that aims to establish a linear relationship between the target y (i.e. output) and a set of predictors x_i (i.e. input parameters). Equation 2.1 shows the general form of a LR [33].

$$\hat{y}_i(\beta, X) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{2.1}$$

where

- β_i: LR coefficients.
- X_i: LR features or predictors.
- $\hat{y}_i(\beta, X)$: LR prediction of the output variable.

The LR modeling includes several methods with the most basic approach being the Ordinary Least Square (OLS). Other methods can be more complex involving other equation forms and algorithms for estimating the regression coefficients.

2.3.1.2 Applications

LR approach with its various forms is used extensively in building energy modeling including establishing baselines and benchmarks. Mathieu et al. [34] used an WLS method that is called Time of Week and Temperature (TOWT) to develop a building energy baseline model. The model considers two input parameters: time of the week and temperature. The time of week segments the week into 15-minute intervals while the temperature is featured into ranges that are a function of the maximum and minimum temperature from historical data. The ranges are fitted using piecewise LR analysis. Existing frameworks modified the method of TOWT by using a Weighted Least Squares (WLS) regression instead of OLS and allowed for recent data to be weighted more than old one. Granderson et al. [35] compared the prediction accuracy of 10 data-driven models including those based on LR methods using data from 537 buildings to gauge the accuracy of M&V modeling approaches. The study included two metrics where LR with appropriate feature engineering

showed similar accuracy to complex models. Kim et al. [36] modeled the energy use of an educational facility based on a set of metered data using LR methods along with more complex techniques over both working and non-working periods. In the study, Kim et al. [36] found that the LR method predicted building energy use less accurately than the complex model during non-working days when occupancy stochastic behavior is difficult to capture. Further applications are shown in Table 2.2.

Table 2.2: LR cases that are applicable to M&V analysis and utilize real historical data

Building Type and Number	Features	Data Granularity	Model Type	References
Bakery, office, and furniture store	Date and temperature	15-minute	OLS	[34]
537 commercial buildings	Varies from model to model with temperature and date as main features	15-minute	**OLS** and **MARS**	[35]
Educational building	Date, Weather, Occupancy	15-minute	OLS	[36]
Health Center	Temperature	Monthly	OLS	[37]
Office Building	Temperature and Occupancy	Monthly	OLS	[38]
Genome Project 2 open dataset of 1578 non-residential buildings	Date, and Meteorological data	3-hour	Baysian LR	[39]
Two office buildings, two shopping malls, one hotel, and one multi-function building	Date, Meteorological data, and Occupancy	15-minute	GPR	[40]
2 Educational buildings	Date and Meteorological data	Hourly	OLS	[41]

2.3.1.3 Discussion

Reported studies showed that LR approaches can vary in complexity and accuracy. The LR approach is often used as a benchmark for more complex models or even as a method with similar accuracy compared to more complex approaches for modeling the building energy consumption. Although LR cannot fit complex non-linear relations, the accurate selection of features, analysis before modeling, and checking LR assumptions can improve its accuracy greatly. Raw features with LR usually do not fit relationships easily while processing features with other models or simple methods allows LR to capture relationships better. This highlights the importance of LR approach to act at least as a benchmarking model.

2.3.2 Decision Tree and Ensemble Methods

2.3.2.1 Definition

DT is a basic non-parametric supervised learning method used for classification and regression analyses. The DT method can predict the value of a target variable using simple decision rules inferred from the data features. The training process for DT follows a piecewise constant approximation approach with different prediction models for various data groups [42]. In the context of M&V applications, decision trees act as regressors rather than classifiers using different metrics to measure their splitting homogeneity or commonly known as impurity. In regression, the case of M&V, the impurity of a leaf is measured by the residual sum of squares. The tree splits data points based on features until fitting the data or reaching specified stopping criteria. The splitting relies on different metrics to decide the goodness of the criterion set at a node which acts as a decision point that splits the data to minimize a specified cost function (i.e. residual sum of squares for M&V applications). Typically, DTs use the splitting criterion as described in Equation 2.2 [42].

$$R_1 j, s = X | X_j \leq s \text{ and } R_2 j, s = X | X_j > s \tag{2.2}$$

where

- s: A decision dividing a node into two leaves.
- R_i: Resulted leaf.
- X_i: Feature from the dataset.
- X: Realizations from the dataset.

Figure 2.5 shows a simple DT for regression where X represent that data points and X_i to X_r represent features from the dataset. At each decision node, the tree divides the data based on criteria, s_i to s_r, where the resulted leaves can have additional decision nodes. The tree keeps branching until minimizing the considered cost function (i.e. Residual Sum of Squares (RSS)) as shown in Equation 2.3 or reaching the

set stopping criteria. The end leaves represent the predicted value for the data points that fall into the leaf based on a series of decision nodes, y_j to y_r.

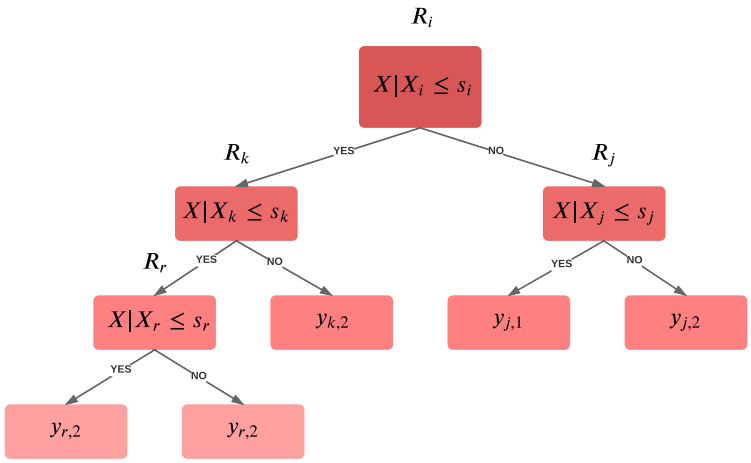


Figure 2.5: Simple structure of decision tree for regression

$$RSS = \sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1}) + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2}) \tag{2.3}$$

However, decision trees can form the basis for more complex models using ensemble methods. Random Forest (RF) is an ensemble method that fits several regressions or classification decision trees for various sub-samples of the dataset and aggregates them by averaging to improve prediction accuracy levels and control the overfitting problem. This ensemble approach is called "bagging" with sampling features and aggregating via averaging. Moreover, other ensemble approaches can be utilized instead of simple weighted averaging methods such as RF. "Stacking" is another ensemble process of generating several base models using training data such that meta-models use predictions from base models as features for out-of-sample predictions. "Blending" is a variation of stacking using testing data set to gauge the prediction accuracy of base models while a final test is applied for the meta-model [43]. The state-of-art ensemble methods include AdaBoost [44], Gradient Boosting Machine (GBM) [45], Extreme Gradient Boosting Machine (XGB) [46], and Light Gradient Boosting Machine (LGBM) [47]. All these methods use the principle of multiple learners except the boosting algorithm which introduces weighting penalization before each successive learner rather

than aggregating the final prediction from multiple learners directly. However, DT and DT-based models like RF and XGB are not effective at extrapolating beyond the range of the predictor's values [48]. Therefore, when a building's energy consumption data include values beyond the trained data for the predictors, other algorithms must be incorporated to overcome this issue.

2.3.2.2 Applications

DT is a machine learning method that is used in both classification and regression applications. Touzani et al. [49] used XGB to determine the improvements of boosting against TOWT by using date and temperature of over buildings where the accuracy metric boxplots showed an improvement over TOWT method. Afroz et al. [50] compared 6 data-driven models by predicting the energy consumption of 11 office buildings located in Ottawa, Canada. RF method is found to provide superior prediction accuracy levels than those of the DT method and even better than those achieved by some models except Nonlinear Autoregressive with Exogenous inputs (NARX). Agenis-Nevers et al. [51] applied 10 methods to model the energy performance of 11 UAE buildings including 10 commercial complexes and one housing unit. RF approach has achieved a global score that is above the average for the 11 buildings. Liu et al. [52] used simulated data generated using DesignBuilder model for an educational building in the Northern China region to compare the energy use predictions from three models. The study found that RF provides the highest prediction accuracy. Publications that utilized DT and ensemble methods are shown in Table 2.3.

2.3.2.3 Discussion

Ensemble methods can be used to develop a set of new models different from base models. With several sampling and aggregating techniques, the choice of the best category approach can pose some challenges. Indeed, the best suitable model depends on several factors and is often not possible to generalize for different building types and retrofit measures. However, reported comparative studies have indicated the appropriateness of certain ensemble methods over others. For example, in several analyses, RF approach outperforms DT in regression modeling as the former prevents overfitting by introducing randomness while the latter tends to branch out until overfitting the training data. On the other hand, approaches such as

stacking rely heavily on their base learners with different applications providing completely different results. While stacking can be effective, it is still a computationally expensive approach with vague transparency and unclear interpretability.

On the other hand, XGB and RF can indicate the contribution of each variable and increase the model interpretability. Given the results of the bibliometric analysis and reported applications of this modeling category, RF and XGB are the two most commonly suitable techniques in ensemble approach with limited drawbacks.

Table 2.3: Decision tree and ensemble cases that are applicable to M&V analysis and utilize real historical data

Building Type and Number	Features	Data Granularity	Model Type	References
410 Commercial building	Date and Temperature	15-minute	XGB	[49]
12 Office building	Date and Meteorological data	15-minute	**RF** and DT	[50]
10 Commercial and 1 residential buildings	Meteorological data	Monthly and Daily	**RF** and DT	[51]
2 Educational Buildings	Date, Meteorological data, and Occupancy	15-Minute	**RF** and DT	[53]
Residential Quarter	Date and Meteorological data	15-Minute	DT, GBM, **XGB**	[54]
507 Non-residential Buildings from Genome Database	Date and Meteorological data	15-Minute	Stacking	[55]
Educational Building	Date, Meteorological data, and Occupancy	15-Minute	Bagging Trees	[56]
Healthcare	Date, Meteorological data, and Occupancy	Hourly	XGB and **RF**	[57]
Hotel	Date, Meteorological data, and Occupancy	30-minute	RF	[58]
1325 air conditioners	Date, Meteorological data, and indoor environmental parameters	Hourly	**XGB**, RF, GBDT, AdaBoost	[59]
Heat pump in a residential building	Date, Meteorological data, and HVAC system operating parameters	30-minute	XGB, **Stacking**	[60]
House	Meteorological data, and indoor environmental parameters	10-minute	XGB	[61]

2.3.3 Support Vector Machine

SVM or in case of regression Support Vector Regression (SVR) is a common machine learning tool used for classification and regression analyses. A SVM model is developed by fitting a hyperplane that aims to determine the underlying relationship between predictors (i.e., input parameters) and target (i.e., output). The hyperplane is supported by two vectors as shown in Figure 2.6 such that the error measured with respect to these two vectors and the hyperplane is minimized by including the maximum number of points within the boundary lines and close to the hyperplane. The two parallel lines represent the supporting vectors while the middle line is the hyperplane. Equation 2.4 shows the hyperplane equation where the data is mapped to a higher dimension by a dot product between points and weights. Then, SVM aims to minimize the cost function which is shown in Equation 2.5 where ϵ represents the distance of the supporting vectors from the hyperplane and ζ represents the distance from the supporting vectors to the points outside the supporting vectors. The more the points that lie within the boundary, the less is the cost function [62].

$$F(x_i) = (w, \phi(x_i)) + b \tag{2.4}$$

where

- b: Model bias.
- $\phi(x_i)$: Kernel function that maps data to higher dimension.
- w: Model Weights.

$$Min(L(w, C)) = Min\left(\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*)\right) \tag{2.5}$$

- $L(w, C)$: Loss or cost function.
- C: Direction regularization coefficient.
- ζ_i: The distance from data observation to any of the supporting vectors which is minimized by the cost function.

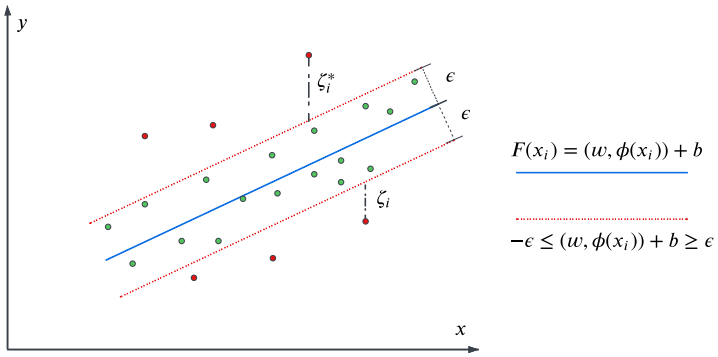


Figure 2.6: One-dimensional SVM for regression

2.3.3.1 Applications

Edwards et al. [63] compared using two variations of SVM against other modeling techniques including LR and ANN. The SVM is demonstrated to have a better performance compared to complex models when applied to residential buildings and to provide similar prediction accuracy levels compared to complex models for commercial buildings. Amber et al. [64] utilized parameters denoting working and non-working days to predict energy demand for an office building which resulted in SVM models that are trained on a subset of data denoting a specific type of day (i.e. work or non-work day) outperforming SVM models that were trained on all the data in prediction accuracy. This result highlights the importance of consistency in occupancy and how the model prediction accuracy can be degraded with more stochasticity in occupant behavior. Although SVM can be computationally expensive, several fitting algorithms can be utilized to minimize the computational time such as parallelizing the training work [65]. Zhao and Magoulès [66] utilized a parallel implementation approach for predicting a building's energy consumption that reduces the training time by parallelizing kernel evaluations and gradients compared to a sequential approach and provides similar prediction accuracy. Table 2.4 provides some reported studies applying SVM for building energy predictions.

2.3.3.2 Discussion

The SVM is a powerful yet computationally expensive algorithm. The mapping of observation to a higher dimension makes SVM superior in fitting complex relationships and minimizing the model prediction

Table 2.4: SVM cases that are applicable to M&V analysis

Building Type and Number	Features	Data Granularity	Model Type	References
3 Residential Buildings	Date, and Meteorological data	15-minute	SVM, **LS-SVM**	[63]
Simulated Office building	Date, and Meteorological data	15-minute	PI-SVM	[66]
4 Commercial Buildings	Temperature	Monthly	SVM	[67]
Hotel	Date, Meteorological data, and Occupancy	15-minute	SVM with RBF	[68]
Commercial Building	Date, Meteorological data	15-minute	SVM with RBF	[69]
60 Commercial buildings	Date, Meteorological data	15-minute	SVM	[70]
Hotel	Date, Meteorological data, and HVAC operation parameters	Hourly	SVM with RBF	[71]

errors. Parallelization can mitigate the slow-fitting performance of the SVM approach especially when dealing with large datasets and when accuracy distribution over the entire dataset is required. The proper choice of kernel when using SVM is not straightforward as the resulted mapped data points can change the prediction accuracy of the model and the process of fitting a hyperplane with no direct relation to model accuracy. Additionally, the choice of a kernel can be determined usually through k-fold cross validation. However, the number of studies using non-linear models as found by the bibliometric analysis suggests that using kernels such as Gaussian or RBF are more common. Furthermore, linear kernels can fit linear hyperplanes for non-complex applications as well as non-linear models that are more computationally expensive.

2.3.4 Artificial Neural Network

Deep learning or artificial neural network (ANN) is a subfield of machine learning where algorithms mimic the human brain functioning process. The ANN involves a set of neurons forming layers that are interconnected starting from an input layer to an output layer. The connections between neurons are determined using weight coefficients that are determined based on a training process using input-output data sets. As discussed in Subsection 2.2, the majority of ANNs used in data-driven building energy modeling are Feed Forward Neural Network (FFNN) [72] as detailed in the following sections.

2.3.4.1 Feed Forward Neural Network

FFNN is the most commonly used ANN-based approach in building energy modeling. Each layer's neurons, comprised of various features' signals, are multiplied by weights $w_{i,j,k}$ that connect them to the other layer's neurons. Followingly, a bias term $b_{j,k}$ is added to the summation of each weight and signal multiplication. The result is then inserted into an activation function which can be either a Rectified Linear Unit (ReLU) or a linear activation function. Without activation functions, the FFNN would be just a LR model. Equation 2.6 shows the process of multiplying weights with signals and adding bias [72]. Figure 2.7 illustrates the basic FFNN architecture. Figure 2.7 shows the same variables in Equation 2.6 with different indices where i denotes the layer number, j the neuron in certain layer, and k the connection. For example, $w_{i,j}$ represent the weight $w_{i,j}$ of the connection between node i and j.

$$\hat{y}_{X,W,b,g(h)} = g\left(W^T X + b)\right) \tag{2.6}$$

where

- W: Weights associated with the connection between neurons.
- X: Inputs form input layer or the output of an activation layer.
- b: Bias term for each neuron.
- $g(h)$: Activation function.

FFNN can have multi hidden layers (i.e. Multi-Layer Perceptrons MLP) or a single hidden layer (i.e. Single Layer Perceptrons SLP). Other forms can have different processes with the same network architecture such as Radial Basis Function Neural Network (RBFNN) [73] or Extreme Learning Machine (ELM) [52]. Both forms have instead of multi-hidden layers, a single hidden layer. RBFNN has radial basis functions that map data to a higher dimension instead of simply activating h. ELM is also a single hidden layer network where initial weights bias terms are initialized using a different method than MLP or SLP and fixed during the tuning phase.

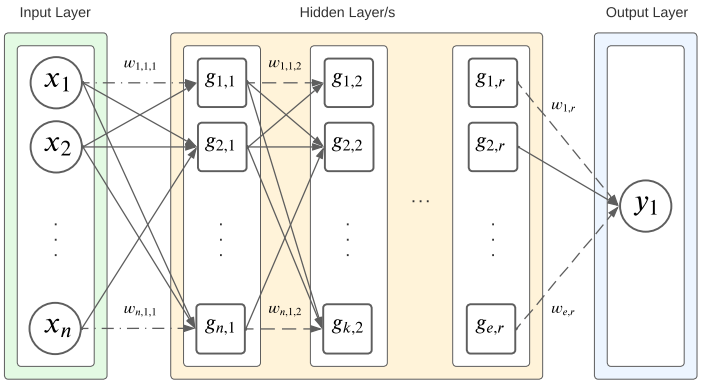


Figure 2.7: Feed forward neural network architecture

Table 2.5: Neural Network cases that are applicable to M&V analysis

Building Type and Number	Features	Data Granularity	Model Type	References
Biomedical manufacturing's chilled water system	HVAC system operating variables	15-minute to weekly	SLP	[74]
Office building HVAC hot water system	Outside dry-bulb temperature	Hourly and Daily	MLP	[75]
5 Office buildings	Date, Meteorological data, and HVAC loads	Hourly	SLP	[76]
1 Educational building, 1 real and 2 simulated office buildings	Date and Temperature	Hourly	SLP	[77]
47 buildings in an educational campus	Date and Meteorological data	Hourly	SLP	[78]
7 Dormitory buildings	Date, Meteorological data, and Occupancy	Hourly	SLP	[79]
Library and ASHRAE Energy Prediction Competition I dataset	Date and Meteorological data	Hourly	SLP	[80]
Educational building	Date, Meteorological data, and Occupancy	15-minute	SLP	[81]

2.3.4.2 Discussion

ANNs are gaining more popularity in building energy modeling due to the availability of better computing machines to perform cumbersome and time-consuming approaches. Furthermore, the development in ANN architecture and algorithms that enable the capture and identification of complex relationships. Nevertheless, the superiority of such methods remains the subject of debate since only slight improvements

in prediction accuracy can be achieved at the expense of significant computational efforts. FFNN-based models can take several forms with the choice between them being difficult to generalize to all building energy modeling applications. Typically, the development of FFNN-based models relies on a trial and error process using cross-validation to obtain the best model's parameters with no clear choice in the reported literature on the best general approach that leads to an accurate model's prediction. Some papers recommended certain methods to find the first iteration's parameters' values such as the number of hidden layers and neurons. Ahmad et al. [58] chose only a single hidden layer and performed a stepwise searching method to select the optimum number of neurons. On the other hand, Amber et al. [82] and Ye and Kim [83] relied on a formula that is a function of both the output and input layer sizes to determine the number of neurons.

2.3.5 Kernel Regression

Another category of data-driven approaches used for building energy modeling is kernel regression. This category of regression analysis approaches is also called time-varying coefficients where response values are predicted using different coefficients for different intervals. Kernel, in this context, is a function that assigns weights to data points based on a specific metric [84]. An example of kernel regression is K-Nearest Neighbor (KNN) [85] regression where Euclidean distance is used as a metric for weighing nearby points where a subset of all data points is selected and each is given an equal weight. Equation 2.7 defines the K-nearest neighbor regression. However, this method can have boundary issues as the regression becomes inaccurate at the endpoints. Additionally, the method generates a curve with several discontinuities as each point has an equal weight. Another approach is Nadaraya-Watson kernel-weighted average [86] which minimizes the weight points based on distance. Equation 2.8 shows the calculation of the model predictions.

$$\hat{y} = \frac{1}{h} \sum_{x_i \in N_{h(x)}} y_i \tag{2.7}$$

where

- N: Neighborhood of points similar based on Euclidean distance.
- h: Count of points in neighborhood N.

$$y = \frac{\sum_{i=1}^{N} K_\lambda(x_0, x_i) y_i}{K_\lambda(x_0, x_i)} \tag{2.8}$$

where

- K_λ: Kernel equation that weighs points in neighborhood N.

The kernel equation K_λ can be Epanechnikov quadratic, Tri-cubic, or Gaussian [84]. In each kernel function, a hyperparameter λ, named smoothing parameter, determines the local neighborhood's widths where lower and higher values can change the variance and bias of the model.

2.3.5.1 Applications

Ho and Yu [87] applied a kernel regression using KNN using measured data for an educational building with a special focus on the energy performance of a chilled water plant. The model included typical features for a building and chiller operating variables such as water flow rate, water supply, and return temperatures, as well as outdoor air dry-bulb temperature, and relative humidity. The model achieved reasonable prediction accuracy levels by selecting the optimal number of clusters based on the lowest mean square error. These results highlight the ability of kernel regression to consider several factors and weight them based on Euclidean distance. Gallagher et al. [88] modeled energy use of a biomedical facility using over 18 features (i.e., input parameters) including dry-bulb temperature data and equipment manufacturing variables such as production machinery electricity consumption, facility operation schedule, chilled water system electricity consumption. The study showed that KNN achieved the best accuracy metrics when using weekly data compared to SVM, ANN, LR, and DT. Wang et al. [89] compared energy use predictions for several data-driven models, stacking, RF, GBM, SVM, XGB, and KNN. The reported results indicate that the KNN-based model has mixed performance as it achieved better accuracy levels than RF and XGB in one case but provided the worst prediction accuracy in another case. Table 2.6 shows a summary of the reported studies using kernel regression for building and retrofit baseline energy modeling.

Table 2.6: Kernel Regression cases that are applicable to M&V analysis and utilize real historical data

Building Type and Number	Features	Data Granularity	Model Type	References
2 Educational buildings	Date and Meteorological data	30-minute and Hourly	KNN	[89]
Biomedical manufacturing facility	Date, Temperature, and manufacturing factors	Intervals from 15-minute to Monthly	KNN	[88]
Educational building	Date, Meteorological data, and chiller operating variables	15-minute	KNN	[87]
Chiller in a public building	Meteorological data and Chiller operating parameters	15-minute	KNN	[90]

2.3.5.2 Discussion

Kernel regression approach provides a powerful tool when modeling relations that are observed frequently over the dataset. By developing neighborhoods of similar points, the Kernel-based models can make predictions that are based on weighted values. The similarity provides a mean for the Kernel-based model to link the mapping between inputs and outputs and easily fit non-linear relations. However, several hyperparameters are encountered when selecting a Kernel-based modeling approach. From the reviewed applications, there appears to be no specific selection guidelines for these parameters other than experimentation and trial and error mechanism. Although complex kernels can produce smooth curves to fit the building energy consumption, there is no clear procedure to develop the set of complex kernels. The common recommendation from reported analyses is that kernel-based models need to be tested over a set of different data and compared against each other to determine the best modeling approach.

2.4 Feature Engineering

Feature engineering is an important step in developing data-driven models as it can significantly affect the models' accuracy. This process involves manipulating the available dataset to transform it into a set of features that data-driven models use to make predictions specific to a response variable. Feature engineering is similar to data processing where raw data is cleaned and any missing values are replaced or dropped.

However, feature engineering utilizes processed data instead of raw data to identify a set of features that enables the model to capture the relationship between predictors and the desired response variable. Machine learning models can be applied to cleaned data without performing any feature engineering, but multiple issues can arise reducing their predictive and explanatory capabilities. Typically, the prediction accuracy levels using training data can increase when a higher number of predictors are included. However, the use of several predictors can lead to over-fitting with data-driven models fitting the noise variations instead of the actual relationship and thus limiting the model's predictive and explanatory capabilities. The aforementioned detrimental effects are important for M&V applications as only historical data sets are available to tune and test the models. Thus, feature engineering can have a significant role in developing sound data-driven models that quantify energy savings accurately when conducting M&V analyses.

For M&V applications, feature engineering can be structured into two main goals: identifying the features to be used in the models and selecting the methods to perform feature engineering. The former emphasizes the mostly used features to predict the energy consumption regardless of the used models while the latter focuses on the approaches to adopt for selecting and/or creating relevant features.

2.4.1 Features

As discussed in Section 2.3, selected features for reported data-driven models vary significantly. For many studies, the model's features are considered based on their availability as limitations and challenges often arise in obtaining and collecting relevant data that are effective in predicting building energy performance. Table 2.7 lists categories and features used in predicting building energy consumption in the published literature. The most used categories of features include outdoor dry-bulb temperature and time-related parameters. Indeed, time-related features are often considered to predict the time dependency of the energy used by buildings. For instance, Mathieu et al. [34] used time of week feature where the week is segmented into 15-minute intervals to capture patterns and correlations that occur on a weekly basis. The process of converting a numerical time value into a grouping factor is usually called one hot-encoding [91]. Time-related features can include parameters such as month, day type, and holidays. The larger the interval of a time-related feature, the more historical data are needed to assess the contribution of that feature

in the development of a relationship between the predictors and the response variable. However, certain relationships to predict a building's energy consumption may target large time periods as demonstrated by Wang et al. [53] who evaluated the effect of occupancy levels during three academic semesters to determine the energy use of an institutional building at the University of Florida.

Another set of important features is meteorological data that include measurements of site weather parameters during the period when the building's energy consumption data are collected. The structure of such data can vary in terms of time granularity and the number of variables depending on the weather station's capabilities. In most weather stations, at least 6 variables denoting outdoor temperature, humidity, pressure, wind, and precipitation are recorded on a sub-hourly basis [92]. Table 2.7 lists common meteorological features used to develop the building's energy models and some reported studies. Some of the analyses have indicated the diminishing return of including all meteorological features as the model's prediction accuracy tends to increase slightly but its overfitting issues increase [34], [51]. When including all meteorological features, it is important to consider the multicollinearity between predictors. Indeed, highly correlated predictors can lead to inaccurate estimates of predictors' contributions and ultimately reduce the model's prediction accuracy levels.

The occupancy level is an important feature that can significantly reduce the unexplained variance in predicting a building's energy consumption. Anand et al. [93] used recorded occupancy presence based on Wi-Fi traffic monitoring devices to predict energy use for an institutional building. Although the study found that a large portion of the building's energy is due to office equipment and plug loads, occupancy level is determined to be a contributing feature. Time-related features can be used as occupancy indicators by using schedules. Zeng et al. [69] incorporated six buildings' schedules in modeling the energy consumption and found that the use of occupancy schedules has a significant impact on the model's prediction accuracy especially for buildings with stable occupants' rate.

Based on Table 2.2 through Table 2.6, the minimum required features for developing a baseline building energy model for M&V analysis include date and outdoor dry-bulb temperature. The date feature must be in a timestamp format to indicate the frequency of recording observations, starting, and ending date of historical values. Outdoor dry-bulb temperature is the main feature commonly reported for establishing the

relationship between weather and building energy consumption. Meteorological features, other than outdoor temperature, have been utilized by certain reported studies with no or little improvement in prediction accuracy [34, 51]. In a number of applications, occupancy rate can significantly improve the prediction accuracy of data-driven models if this data is readily and accurately available. As alternatives to occupancy rates, time-related features using operation schedules and building use patterns have been successfully considered in some applications. It is important to note, however, that as the modeling time step becomes shorter (e.g., hourly to daily), accurate estimations of occupancy patterns become more challenging.

Table 2.7: Features used in building energy consumption prediction

Feature Categories	Feature	References
Date-Related	15-minute of an hour	[34], [49], [50], [61], [68], [66], [73], [81]
	Hour	[57], [59], [71], [79], [94], [36], [81], [95]
	Day	[53], [36], [51], [81], [81], [71], [41], [95]
	Week	[53], [41]
	Holiday	[49], [51], [96]
	Month or/and Biannually	[81], [95], [53]
Meteorological	Outside Dry-Bulb and/or Wet-bulb Temperature	[34], [36], [49], [50], [61], [68], [66], [73], [53], [51], [81], [71], [41]
	Relative Humidity and/or Humidity Ratio	[53], [36], [81], [71], [41]
	Solar Irradiance	[53], [36], [51], [81]
	Enthalpy	[51], [41]
	Wind Direction and/or Speed	[53], [36], [51], [81]
Occupancy-Related	Infrared Sensors and/or recorded Equipment use	[36], [81], [71]
	Schedules and Records	[69], [53]
Operation-Related	Indoor Dry-Bulb Temperature	[59], [71]
	Building Systems' Operating Variables	[61], [71], [41], [95]

2.4.2 Feature Processing and Extraction

In processing analysis for developing baseline building energy models for M&V analysis, time and outdoor air temperature (i.e., dry-bulb and/or wet-bulb air temperatures) are the most commonly used features. Often, outdoor air temperature especially when it fluctuates significantly can reduce the ability of data-driven models in providing accurate predictions of building energy consumption. Therefore, this feature is usually processed to have a predictor with less variation frequency such that the building's thermal dynamics can be easily explained [97]. Table 2.8 lists some predictors that are calculated based on outdoor air temperatures including Cooling Degree Days (CDD), change-point temperatures, and piece-wise fitting models. Such predictors provide surrogate variables to determine the impact of the outdoor air temperature variations on the building's thermal performance. In terms of time-related features, hot-encoding and factoring provide a set of features with categorical instead of numerical variables. Examples of these categorical features are listed in Table 2.7 where timestamps of 15-minute intervals can be converted into a set of categorical variables such as hour, day type, and holidays. The last two processing methods indicated in Table 2.8 do not rely on domain knowledge but are based on algorithms such as PCA and Deep Learning Extraction to develop a set of new features to enhance the predictive capabilities of data-driven models.

2.4.3 Feature Selection

The selection analysis in feature engineering involves the conversion of the original dataset into a smaller dataset with fewer features. As noted earlier, increasing the number of features has not only a limited effect on improving the model's prediction accuracy but may result in an overfitted model with reduced explanatory capabilities. At the initial development phase of a data-driven model, EDA-based methods can be used to extract insights into the relationship between the predictors and the response variable. The EDA-based methods include pairwise correlation and plotting against the response variable [52]. However, data double-dipping must be avoided as identified relationships between the response variable and the predictors using the EDA approach based on the same training dataset can lead to enforcing relationships that do not necessarily exist outside the analyzed dataset. Gallagher et al. [74] used developed a M&V modeling methodology where a feature selection pipeline of two metrics reduced a dataset of 504 to 15 feature. The

pipeline sorts several features by iteratively using Spearman correlation coefficient between features to select the optimum set of features that maximize the coefficient of determination R^2. Followingly, Variance Inflation Factor (VIF) is utilized to remove any possible multicollinearity where features with a VIF greater than 5 being dropped.

Another method to filter the features is by forward and backward elimination approach with the model recursively trained on a different set of features and its performance is assessed using various evaluation metrics. These evaluation metrics such as Akaike Information Criterion (AIC) [98] can be used to balance the number of features and the accuracy level. Feature Importance methods [99], often used in DT-based models, such as RF and XGB, can be considered for the selection analysis with features splitting the leaves having a high contribution in improving the model's prediction accuracy are identified to have higher importance. In M&V applications, the process of selecting features rely on the availability of more features by having multiple meters, weather data that includes several parameters, or multiple occupancy sensors that are laid on several locations inside the building. Moreover, generated time-related features (e.g. day of a month, month of a year) might be dropped if the received information gain does not lead to an improvement in prediction accuracy. A list of selection processes in building energy modeling are summarized in Table 2.8.

Table 2.8: Feature engineering methods in building energy consumption prediction

Feature Engineering Category	**Feature**	**Method**	**Ref**
Processing	Outdoor Temperature	CDD and HDD	[41], [100]
		Change-Point	[50], [100]
		Piece-wise Fitting	[101], [34], [49]
	Time	Hot-encoding	[34], [101], [49], [51], [96]
	All or Multiple Features	Clustering	[101], [94]
	All or Multiple Features	PCA	[102], [80], [69]
	All or Multiple Features	Deep Feature Extraction	[103]
Selection	All or Multiple Features	Forward and Backward Selection	[63], [50]
	All or Multiple Features	Feature Importance	[61], [59], [51], [60]
	All or Multiple Features	EDA	[52], [79], [59], [93], [69]

2.5 Data Requirements

Data-driven models for M&V rely heavily on historical data to establish a relationship between input variables and building energy performance. The quality and quantity of historical data can significantly affect the accuracy of a data-driven model. In particular, the following three characteristics are often used to assess the quality and quantity of the data: time range, reporting frequency, and missing values. Data time range affects the re-occurrence of certain performance levels that can help models identify repeating patterns or ignore unusual activities. Grillone et al. [101] simulated 54 cases of three buildings with different parameters and trained two data-driven models using data specific to a period ranging from 9 to 12 months. The result showed that a significant decrease in the prediction accuracy distribution median and an increase in the distribution variance when using TOWT approach. OpenEEmeter [104] is an open-source framework used to calculate the energy use that could be avoided by retrofitting a building. The framework sets certain requirements on the data used for developing a building energy model including the data time range. For data with hourly and daily frequency, an OpenEEmeter compliant baseline building energy model requires at least data for 365 days.

The time frequency provides the level and type of information that can be gained from data through data-driven modeling. Using hourly or sub hourly data for energy consumption, any patterns and correlations can be learned better but more significant noise levels could be introduced as the building energy consumption becomes less consistent. On the other hand, aggregated consumption using daily or monthly frequencies exhibit less fluctuations at the expenses of extracting more information. Gallagher et al. [74] analyzed the effect of sub-hourly, hourly, daily, and weekly frequency on four data-driven models by using a recorded measurement of a chilled water system. They found that the frequency effect varied between models with daily frequency producing the lowest CV(RMSE) except for KNN where the hourly-based model resulted in a lower CV(RMSE).

Missing values represent another important requirement for the quality of data needed for training. Missing values are identified by periods of disconnected metering, irregular values, or missing some features' values during a given timestamp. Although each case of missing data is usually unique and requires a certain

imputation technique, several thresholds were established to prevent training models from using invalid data sets. CalTRACK [105] dictates missing data requirements for daily and hourly frequencies specific to data-driven models. Models based on daily data must not have more than 37 days (i.e. 10% for a full-year data) while hourly frequency data must have less than 10% missing hours of the total hours in every calendar month.

2.6 Existing Open-Source M&V Frameworks

Several papers developed specific data-driven models for building energy modeling and M&V analysis as shown in Table2.2 through Table 2.6. The process of developing such models involves specific knowledge of the underlying statistics and building energy consumption patterns and is limited to the considered building type and location. To automate this process, only limited analysis frameworks have been proposed with varying degree of capabilities and applications. Although such frameworks still require some knowledge of the underlying process, they still facilitated the development of baseline energy modeling needed for performing an M&V analysis. Table 2.9 lists the limited open-source frameworks and their characteristics suitable for M&V analysis.

Table 2.9: Sample of existing M&V frameworks

Framework	Models	Inputs	Frequency	Development Language	Ref
ECAM	LR	Date, Dry-bulb temperature, and Occupancy	Hourly, Daily, and Monthly	Excel add-in	[106]
EEMeter	LR	Date, Dry-bulb temperature, and Occupancy	Hourly, Daily, and Monthly	Python	[104]
NMECR	LR	Date, Dry-bulb temperature, Occupancy, and independent variables	Hourly, Daily, and Monthly	R	[107]
RMV2.0	LR, GBM	Date, Dry-bulb temperature, and Occupancy	Hourly	R	[108]

2.7 Models Evaluation

To accurately estimate energy savings associated with retrofitting buildings, baseline models need to be established to provide predictions with a satisfactory accuracy level using specific qualities. Several

of such qualities have already been discussed throughout this review including generality, predictive, and explanatory capabilities. The process of obtaining a model with the best aforementioned qualities require a selection of both evaluation metrics and approaches.

2.7.1 Evaluation Approaches

To have good predictive performance, data-driven models must balance bias and variance effects to provide accurate predictions even using unforeseen conditions. To enhance their prediction accuracy, models are often trained using a fraction of the available historical dataset until acceptable accuracy levels are reached without compromising other model performance metrics such as overfitting. Then, the trained models are tested on a new data set not used in the training analysis to evaluate their performance using metrics such as prediction accuracy level. A basic approach of evaluation is to split the data into training and testing dataset. With cross-sectional data, the split is random but with time series, the split must be carried to have two sets of data that are similar such that the model testing process is valid. An approach to split time series data is to set a date at the end of the data where observations after such date are left for testing. However, this approach is limited if the some of features' values does not occur in the training set, a case that can happen when splitting with a training data of less than a year time range.

Another approach is using k-fold cross validation where the whole data set is segmented into k consecutive blocks and the model is trained on all blocks except one which is left for testing. Followingly, the process is iterated by changing training and testing blocks k times and averaging the resulted evaluation metric. There is no specific number for the cross validation folds (i.e. blocks) but in practice, a value of 5 or 10 are usually chosen [33].

2.7.2 Evaluation Metrics

Regardless of the followed evaluation approach, different evaluation metrics can be used to indicate a desired quality of the trained model. One metric commonly used for model performance evaluation is the coefficient of determination, R^2, the explained variance as expressed by Equation 2.9 where $\hat{y}$ represents predicted values.

$$R^2 = \sum_{i=1}^{n} \frac{(\hat{y} - \bar{y})^2}{(y - \bar{y})^2} \tag{2.9}$$

However, R^2 does provide an accurate performance metric when comparing multiple models having a different number of predictors. Therefore, the adjusted coefficient of determination R^2_{adj} is used such that a complex model with a high number of predictors is penalized as shown by Equation 2.10:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - P - 1} \tag{2.10}$$

where p is the of predictors and n is the number of training data points. Similar metrics to R^2_{adj} are AIC and Bayesian Information Criterion (BIC) where the model is penalized as its complexity increases.

For M&V applications, Normalized Mean Bias Error (NMBE) and Coefficient of Variation of the Root Mean Squared Error (CV(RMSE)) are the most commonly used performance metric to assess the model's prediction accuracy levels. The NMBE measures the overall bias in the model's predictions as defined by Equation 2.11 with a positive value indicating that the model is on average over-predicting and a negative value being an indicator that the model is under-predicting.

$$NMBE = \frac{100}{n} \times \frac{\sum_{i=1}^{n} y - \hat{y}}{\bar{y}} \tag{2.11}$$

The NMBE is invariant of the time granularity as models with 15-minute or hourly interval predictions result in the same NMBE value. The CV(RMSE) measures the difference between actual and predicted values as expressed by Equation 2.12 where the differences are squared instead of being summed as the case for estimating the NMBE value.

$$CV(RMSE) = 100 \times \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2}}{\bar{y}} \tag{2.12}$$

The CV(RMSE) is sensitive to the model's prediction time granularity so its value for a model with 15-minute prediction intervals is different from the value obtained from another model with hourly prediction intervals. Both metrics are important in evaluating and comparing the performance of several models.

CHAPTER 3

MVBEP FRAMEWORK DEVELOPMENT

The literature review demonstrated several commonly used approaches when developing a data-driven model for predicting a M&V baseline for building retrofits. In particular, reported studies showed that several methods and tasks need to be incorporated into the process of developing a data-driven model. For example, several papers showed that processing timestamps into features such as hour of the day or day of the week improves generally the prediction accuracy. On the other hand, features such as relative humidity and solar irradiation do not always increase the prediction accuracy for data-driven models of building energy consumption. These variations in prediction accuracy suggest that the process of developing a data-driven model relies heavily on performing several steps within each task category to evaluate different combinations of such variations. However, exhausting all alternatives is cumbersome and computationally expensive. Therefore, a careful selection of fixed and/or changing methods must be performed to streamline the process of developing effective M&V data-driven models. Fixed methods are specific approaches that are always executed when developing a model while changing methods are approaches that can vary depending on the applied case study and the desired prediction accuracy levels.

This chapter described the methodologies used to develop the MVBEP framework suitable for data-driven models specific for M&V analysis of building energy retrofit projects. Section 3.1 discusses the framework conceptualization by describing the requirements and issues that the framework should consider when developing baseline models. Followingly, Section 3.2 discusses in detail the framework structure and process flow with performed the required tasks using 5 distinct modules. A clean dataset obtained for one office building is considered in Section 3.3 as the first application of using the MVBEP tool to perform M&V

analysis of a retrofit project using data-driven modeling.

3.1 Structure Conceptualization

The structure concept for the proposed MVBEP framework is derived from the guidelines set by the Cross Industry Standard Process for Data Mining (CRISP-DM) [109] which is a data science process for creating a Machine Learning (ML) model in six sequential steps starting from business understanding to model deployment. Each step serves a general purpose for any given data science project and is incorporated in the developed framework with some slight changes as detailed in the following sections.

3.1.1 Business Understanding

As described in Section 2.1, the main objective of the proposed MVBEP framework is to identify and select a sufficiently accurate predictive data-driven model that can estimate the building baseline energy consumption before retrofitting to predict the post-retrofit period.

3.1.2 Data Understanding

The data used for developing a data-driven model is expressed in timeseries for both dependent and independent variables. For the MVBEP framework, the data must include at least timestamps, outdoor dry-bulb temperature, and energy consumption. Based on Section 2.4, these features are the minimum required features to build a M&V data-driven model. However, the MVBEP framework should accept additional features that can include other meteorological variables or static variables indicating weekly schedules or annual holidays. Section 2.5 stated other data specifications in terms of date ranges, missing values, and frequencies. The MVBEP should be capable of considering similar specifications including handling different frequencies such that the framework can be applied to a wide range of M&V cases.

3.1.3 Data Preparation

The MVBEP framework needs to be able to process data before developing a data-driven model. The processing should begin by validating the input format and transforming raw features into useful variables for

training and testing the developed data-driven models. Data of features with irregular and or invalid values should be detected and solved by suitable data imputation techniques. In summary, the MVBEP framework needs to be able to identify, process, and select appropriate features to develop accurate predictions based on the provided data.

3.1.4 Modeling

Section 2.3 highlighted several commonly used modeling approaches from five distinct. These modeling approaches can be included in the developed MVBEP framework to aid in the data-driven model development and selection process. Additionally, The MVBEP framework needs to perform hyperparameter tuning for some or all considered models to determine the best combination of possible hyperparameters. This tuning process must include a selection of mostly tested hyperparameters to minimize the grid search process time. Moreover, if performing hyperparameter processing is not invoked and/or not desirable, The MVBEP framework should have default hyperparameters that generally lead to developing accurate models.

3.1.5 Evaluation

Based on Section 2.7, the evaluation process follows an approach of splitting the data for two tasks: training and testing the considered data-driven models. The training task involves the estimation of the best model's parameters including hyperparameters while the testing task is used to evaluate the prediction accuracies of the data-driven models considered by the MVBEP framework. The evaluation metrics include CV(RMSE) and NMBE depending on the user-defined criteria. In particular, the trained models could be ranked according to either CV(RMSE) or NMBE values obtained for the testing task. Moreover, the MVBEP framework should convey a certain level of interpretability for the users to better understand the contribution made by each feature to the resulting model's predictions. Although interpretation is not fully part of the evaluation process, it can assist in a better assessment of the model's capabilities.

3.1.6 Deployment

Unlike typical projects that follow the CRISP-DM methodology, the deployment for the MVBEP framework involves preparing a data-driven model to perform predictions by streamlining the processing of data suitable for generating predictions. Specifically, the MVBEP framework can be deployed as an effective tool that delivers data-driven models that are suitable for M&V analysis for a wide range of applications of building retrofit and measured data. The MVEBP framework should minimize the required knowledge of the underlying process without compromising the ability of the user of changing certain parameters.

3.2 MVEBP Structure

To meet the CRISP-DM's guidelines, the MVBEP framework is developed to have a modularized structure as depicted in Figure 3.1. As indicated by the flowchart of Figure 3.1, two steps are considered in the MVBEP framework structure including the MVBEP tool development and its application for baseline energy predictions. The first step is performed using pre-retrofit data obtained when the building's energy consumption is metered before the start of any retrofit. This step focuses on verifying the input features and the quality of the data used to develop accurate predictive data-driven models. The second step aims to generate predictions using the developed model applied to the post-retrofit data with the objective to quantify savings expressed by the Avoided Energy Use (AEU) compared to the actual retrofitted building's energy consumption. The following sections describe briefly the various modules specific to the MVBEP tool developed to generate, test, and apply data-driven models suitable to predict building baseline's energy consumption for retrofit projects.

3.2.1 Initialization

The MVBEP tool has an initialization module that can be used to assess the quality of the unprocessed input data. The data must conform to specific requirements to successfully initialize any data-driven model. These requirements are derived from previously reported studies as detailed in Chapter 2. These requirements can be categorized into three groups: input, general, and feature-specific considerations. Based on the initialization module results, the MVBEP tool generates a summary of the quality assessment of the used

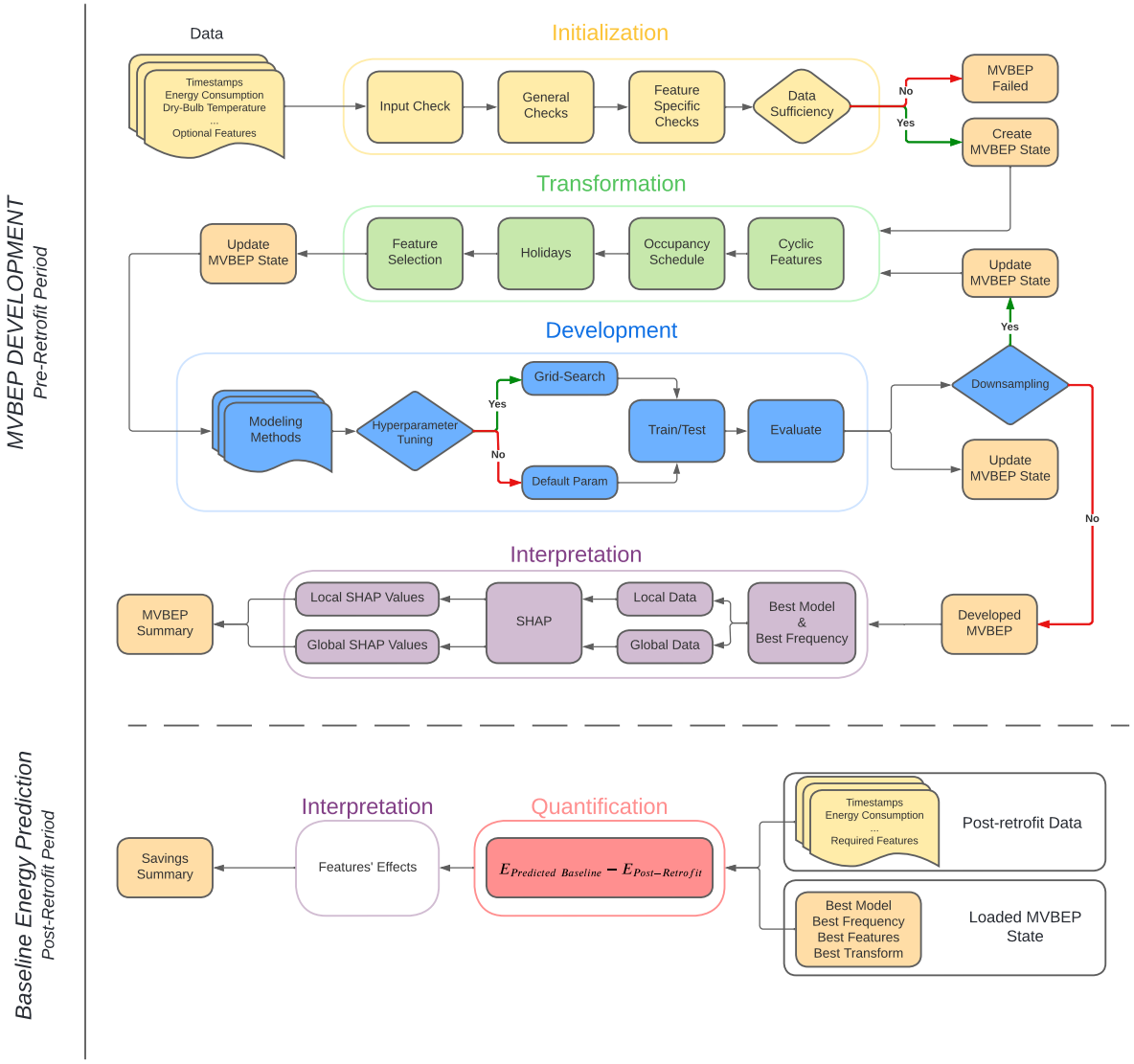


Figure 3.1: Flowchart for the development and application of the MVBEP framework

data. As part of the input quality requirements, the data is checked to include at least timestamps, outdoor dry-bulb temperature, and energy consumption. For the data to be sufficient to develop effective data-driven models, the following requirements extracted from reviewing previous M&V data-driven modeling literature are considered for checking the quality of input data as part of the MVBEP tool:

- Timestamp range: 365 days.
- Max Consecutive of invalid values or timestamp jumps: one day for 15-min and hourly frequency

and a week for daily frequency.

- Max total of invalid values or timestamps jumps: 10% of the data (i.e. 876 hours of hourly frequency and 37 days for daily frequency).

The first requirement states that the timestamps for daily aggregations must cover at minimum a full year. However, this default threshold can be edited by the user of the MVBEP tool. The other two requirements, which can also be changed by the user, are applied to all the remaining features. If any feature failed in meeting the last two requirements, it is automatically dropped from the data set except for energy consumption and outdoor dry-bulb temperature for which any failure to meet the requirements prevents the MVBEP tool to move forward with the development of data-driven models. General checks include checking the timestamp format, total data range, and other timestamp-related criteria. Feature-specific checks involve evaluating the remaining features' values which include irregular energy consumption and missing values. The initialization module takes actions to remedy minor data issues such as replacing values with locally weighted rolling averages and dropping optional features with many missing values. After the initialization step is completed, the MVBEP tool reports the results of the initialization process.

3.2.2 Transformation

The transformation module converts the initialization module's processed data into data that is ready for training and testing data-driven models. The transformation process starts by generating cyclic features depending on the timestamps' frequency which can be 15-min, hourly, or daily for the MVBEP tool. Cyclic features are developed by converting timestamps into time-related features as detailed below:

- **For 15-min datasets**:

  (1) 15-min of the hour (i.e. $\{1, 2, 3, 4\}$).

  (2) Hour of the day (i.e. $\{1, 2, \cdots, 23, 24\}$).

  (3) Hour of the week (i.e. $\{1, 2, , \cdots, 167, 168\}$).

  (4) Day of the week (i.e. $\{1, 2, \cdots, 6, 7\}$).

(5) Day of the month (e.g. $\{1, 2, \cdots, 30, 31\}$).

(6) Day of the year (e.g. $\{1, 2, \cdots, 364, 365\}$).

- **For hourly datasets**: All the features listed for the 15-min datasets except 15-min of the hour.
- **For daily datasets**: All the features listed for the hourly datasets except the hour of the day and the hour of the week.

The time-related features listed above are converted to cyclical features to better represent time patterns to train a data-driven model. For example, the possible values for hour of the day are integers between 1 and 24. Since 1 is close to 24, the transformation module can reduce the difference between these timestamps using periodic functions such as sine and cosine as shown in Equation 3.1 and Equation 3.1.

$$t_{\sin} = \sin\left(\frac{2\pi t}{n}\right) \tag{3.1}$$

$$t_{\cos} = \cos\left(\frac{2\pi t}{n}\right) \tag{3.2}$$

where

- : t: Encoded values of a time-related feature (e.g. hour of the day).
- : $t_{\sin}$: The sine cyclical value.
- : $t_{\cos}$: The cosine cyclical value.
- : n: The number of unique in a specific encoded time-related feature (e.g. 7 for hour of the day).

The use of both sine and cosine values for a single time-related feature helps converting the time-related feature into a cyclical variable. For example, when using only the sine function, hour of day values of 12 and 24 correspond to zero when transformed to sine values indicating that there is no difference. The cosine function complements the sine function and ensures that the transformed feature is cyclical as the cosine values for 6 and 18 are -1 and 1, respectively. However, such a transformation is not ideal as it assigns more weight to a single feature when used to train a model as one piece of information is segmented into

two features. For example, instead of having one feature representing hour of the day, two features are used to represent such a feature (i.e. cosine and sine values for hour of the day). Additionally, some models including those based on tree-based methods fail to process both the sine and cosine values simultaneously due to the need for one-feature process splitting at every decision node. Nevertheless, such transformation is often better than relying on the encoded values of time-related features.

If MVBEP was passed with arguments indicating a general occupancy schedule, the transformation process will add a feature called 'schedule' to indicate the building's general occupancy density. The passed arguments for creating an occupancy schedule can include recurring annual days of low occupancy density or general weekly hours indicating hours with low occupancy density. Additionally, the building's country code can be passed where MVBEP will utilize the holidays package [110] on Python Package Index (PyPI) repository to generate multiple features indicating whether the timestamp occurs on a public holiday or not.

The feature selection is the final step in the transformation step. Based on the input data's frequency, the model selects either one or two sets of features and datasets. The default resulting features and dataset whenever performing the transformation includes all the aforementioned transformations (i.e. from creating time-related features to public holidays). In the hourly frequency data, another dataset and set of features are created according to TOWT method along with the default dataset and features. The reason for creating a dataset according to TOWT method is that LR when trained with such data demonstrated in multiple reported studies to be accurate in predicting hourly energy consumption [35, 101]. Therefore, in the hourly frequency case, an additional dataset and features will be created according to what Mathieu et al. demonstrated [34].

3.2.3 Development

After completing the transformation processing, the MVBEP tool initiates the development including training and testing of data-driven models. As detailed in Section 2.3, five categories of data-driven modeling approaches can be considered for M&V analysis of retrofitting building energy systems. The simplest data-driven modeling option consists of LR with TOWT approach which is widely used for developing baselines of existing buildings. Ensemble modeling approach includes two prominently applied methods for assessing

building energy performance: RF and XGB. Several reported data-driven models have been developed using the SVM approach combined with a range of hyperparameters. However, there are no clear guidelines from the reported literature on determining the best combination of hyperparameters specific to SVM models to be suitable for M&V analysis in estimating building energy savings from retrofit projects. In addition, a wide range of FFNN-based models has been considered to predict building energy performance with different architectures and features. Among the reported FFNN's architectures, SLP is the mostly used in predicting building energy consumption. Lastly, kernel regression methodology has been applied for building energy prediction with KNN being the commonly used for M&V applications. Therefore, the implemented data-driven modeling approaches in the MVBEP tool include LR, RF, XGB, SVM, SLP, and KNN.

The review in Section 2.7 discussed several evaluation performance metrics and approaches to assess the prediction accuracy of baseline building energy models. In particular, evaluation metrics for both general building energy prediction and M&V analysis have been discussed. It is found that CV(RMSE) and NMBE are the mostly used metrics to evaluate the building energy models. These two metrics complement each other and convey better insights about the model's performance. Therefore, the performed evaluation in the development module for the MVBEP tool is based on both CV(RMSE) and NMBE metrics. The development process segments the data into two sets: one set for training and the remaining set for testing the prediction accuracy of the data-driven models. For the MVBEP tool, the default fractions for the training and testing sets are set to be 0.8 and t 0.2 with the option for the user to change these fractions to provide the desired testing period size. . The aforementioned modeling approaches generate models by using the training data to develop the relationship between the features and the target variable. On the other hand, the testing data is used for final evaluation with the main objective to select the best data-driven model for the considered retrofit project. Ultimately, the best model is directly selected by the user with the help of the MVBEP tool's generated development report, including the training and testing CV(RMSE) (i.e., default metric), and/or NMBE values. As stated in the transformation step description.

3.2.3.1 Hyperparameter tuning

Hyperparameter tuning is an optional step in the development phase of the data-driven models. The tuning process aims at identifying the best set of a model's hyperparameters that minimizes both variance and bias and ultimately improves the prediction performance of the model. Firstly, a set of hyperparameters are chosen using a specific range of discrete values. The Sklearn grid-search module considers all the combinations of hyperparameter values to minimize cross-validation's prediction accuracy metric (i.e. RMSE) [29]. To avoid the possibility of data leakage, the cross-validation training folds consist of observations that happened before the testing fold. Figure 3.2 illustrates the difference between K-fold and rolling cross-validation. In K-fold cross-validation, the training folds are selected regardless of the order for the training and testing folds. On the other hand, rolling cross-validation training folds are always before the testing folds [62].

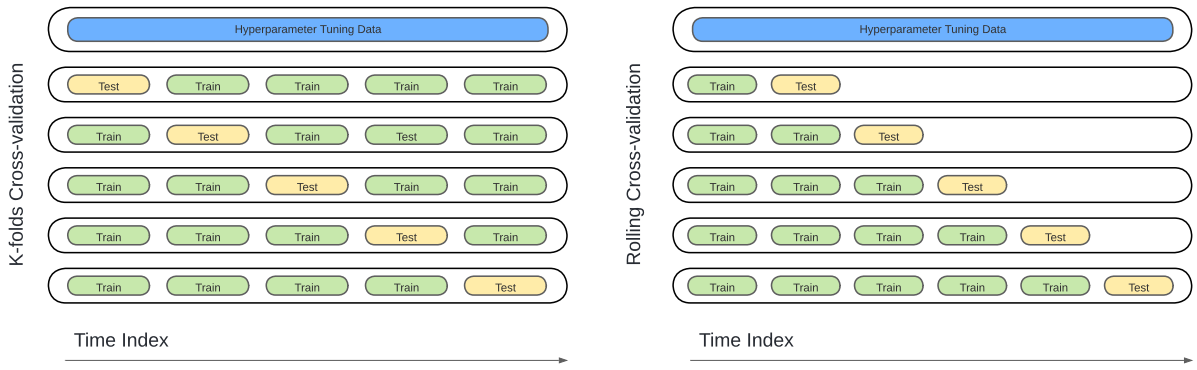


Figure 3.2: K-fold and rolling cross-validation

The hyperparameters that are considered by the MVBEP tool for various modeling approaches are listed in Table 3.1. The list of these hyperparameters is based on their importance in improving the model's prediction accuracies as well as on their computational time requirements. It is difficult to develop accurate data-driven models suitable for various case studies while restricting the hyperparameters to a specific set of values. Therefore, the MVBEP tool utilizes default values specific to the considered modeling approaches based on reported studies to be efficient along with the flexibility to change these hyperparameters by the user. The hyperparameter tuning is achieved using a grid-search strategy also known as the brute force hyperparameter search [111].

3.2.3.2 Random Forest Regressor

The selected hyperparameters for RF-based models include bootstrap, min samples leaf, and the number of estimators as shown in Table 3.1. These hyperparameters' names might vary depending on the programming language and used package. Probst et al. [112] summarized several reported studies on the hyperparameters' influence on RF's prediction accuracy and presented several strategies to tune RF-based models. Bootstrapping is included to examine the influence of introducing more randomness in the forest as trees could have different samples of data which may lead to reducing variance. The hyperparameters can have a negative effect on the model's prediction accuracy. While the number of estimators often reduces the model's bias, it may lead the RF-based model to overfit the data as more and more trees learn from the same repeated data. The last considered hyperparameter for RF-based models Consists of the minimum required samples to split at a leaf which tends to affect the model's variance. The remaining RF's hyperparameters are listed in SKlearn documentation for RF regressor [29].

3.2.3.3 Extreme Gradient Boosting Regressor

According to XGB's documentation [46], there is over 20 hyperparameter for XGB-based regressors with each having specific impacts. The selected hyperparameters for the MVBEP tool are limited to the important parameters in order to minimize the grid-search process time. Similar to the RF-based models, the number of estimators or boosting rounds is an important hyperparameter that affects the variance of XGB-based models. The boosting learning rate controls the learning rate in each boosting round. There is no specific value for such a hyperparameter but a small value is selected as default for the development of XBG-based modeling using the MVBEP tool as shown in Table 3.1. The model's complexity can be reduced by using a regularization hyperparameter that is called Gamma where information across trees is used to reduce complexity. This hyperparameter regularizes the complexity of the model instead of the specific complexity of each tree. Again, the selection of the best value for the regularization parameter is dependent on the training data and the other selected hyperparameters. However, an additional value that is closer to the default value set by the documentation [46] is selected as shown in Table 3.1.

3.2.3.4 Support Vector Regressor

The SVR approach involves computationally expensive algorithms as reported by several studies discussed in Section 2.3.3. Only two hyperparameters are selected for the MVBEP tool for SVR-based models since training a single combination of hyperparameters requires significant computational time compared to other modeling approaches. Yang and Shami [113] listed the regularization parameter C and SVR kernel as the most important hyperparameters to tune when modeling with SVR methodology. The mostly used kernels are usually Radial Basis Function (RBF) and polynomial kernel. As C increases, there is less regularization and supporting vectors start to include more points which increases bias. As C decreases, supporting vectors become closer to the hyperplane which increases variance [29]. For the MVBEP tool, the SVR-based model development is carried out using the SVR algorithm obtained from Sklearn [29].

3.2.3.5 Single-Layer Perceptron

For the MVBEP tool, the FFNN-based models are implemented using the Sklearn MLP module [29]. This module trains regression neural networks using different architectures and hyperparameters. One of the most important hyperparameters for FFNN-based models is the FFNN architecture which is mainly represented by three components: hidden layers, hidden neurons in every hidden layer, and activation functions. Based on the reported studies specific to FFNN-based modeling for predicting building energy consumption, SLP is the mostly used architecture. The number of hidden neurons varies from one case study to another. Studies have suggested some formulas based on input and layer sizes to obtain the number of hidden neurons [80, 114]. For the MVBEP tool, the number of neurons can vary from 5 to 21 with a step of 4 neurons which covers ranges used in reported SLP-based studies. The used activation function in the MVBEP tool is Rectified Linear Unit (ReLU) which is the most commonly used in deep learning regression as it has several advantages over other activation functions such as the sigmoid or tanh functions in issues such as exploding or vanishing gradients [115]. The remaining hyperparameters for FFNN-based models include the initial learning rate and the optimization method (i.e., the solver). Table 3.1 shows that the initial learning rate is set around the default value in Sklearn MLP module documentation [29] as there is no general rule for choosing a learning rate. Two solvers are selected in the MVBEP tool: Stochastic

Gradient Descent (SGD) and Adaptive Moment Estimation (Adam) which are two popular optimization algorithms in deep learning [72]. For the development of the MVBEP tool, the Sklearn library [29] was chosen over Tensorflow [116] because it is readily available when installing Python. For simple regression neural networks, the performance between the two frameworks is similar.

3.2.3.6 K-Nearest Neighbor

The most important hyperparameter in the KNN approach is the number of neighbors which can significantly vary from one case to another. Therefore, the selected range considered for the KNN-based models in the MVBEP tool is between 5 and 75 to have better generalization over various M&V case studies. The MVEBP tool always performs hyperparameter tuning for the KNN-based models regardless of whether performing hyperparameter tuning is selected by the user or not as there is no single default value for such a model. The KNN-based model development in the MVBEP tool is implemented using the Sklearn KNN regressor algorithm [29].

3.2.4 Interpretation

The interpretation is a step that explains features' effects and the contribution of each feature in making predictions. The interpretation is provided only for complex models which are RF, XGB, SVR, SLP, and KNN as such models are difficult to interpret. The interpretation is accomplished by using SHapley Additive exPlanations (SHAP) framework [117] which estimates the local importance of each feature. By using background data, a sample from training data, and the trained model, SHAP produces an explainer that outputs the contribution of each feature to every prediction. Equation 3.3 shows the relation between features' contribution and the predicted value.

$$f'(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x'_i \tag{3.3}$$

where:

- $f'(x)$: The trained data-driven model (e.g. RF).

Table 3.1: Modeling approaches and their selected hyperparameters

Modeling Approach	Hyperparameter	Grid-search values	Default values	Hyperparameter Description
RF	bootstrap	True, False	True	Whether bootstrap samples are used to build each tree or the entire dataset
	min_samples_leaf	3, 5	1	The minimum number of samples at the resulted leaf
	n_estimators	100, 200, 500	100	The number of trees in the forest
XGB	n_estimators	100, 200, 500	100	The number of gradient boosted trees
	eta	0.05, 0.3	0.3	Boosting learning rate
	gamma	0, 0.4	0	The required minimum loss reduction to split a leaf node
SVR	C	0.1, 1, 10	1	Regularization parameter
	kernel	rbf, poly	rbf	The kernel that maps values to a higher dimensions
SLP	learning_rate_init	0.0001, 0.0005, 0.001	0.001	The initial learning rate to control the update's step size
	hidden_layer_sizes	5, 7, 14, 18	13	The number of hidden neurons
	solver	adam, sgd	adam	The used optimization algorithm
KNN	n_neighbors	5, 10, 15, ..., 75	N/A	The number of neighbors to use when making predictions

- $g(x')$: The explanation model produced by the SHAP framework.
- x: A vector of features' values referring to a single observation (e.g. energy consumption at a specific timestamp).
- x': A vector of simplified SHAP inputs that are mapped from x.
- ϕ_0: The expected value (i.e. mean) of the trained data-driven model's predictions for the background data (i.e. the data that is used to build the explanation model).
- ϕ_i: The feature contribution for a single observation.

Each value of $\phi_i x'$ represents the contribution of every feature in the prediction and when summed with ϕ_0 (i.e. the expected value of the model prediction), it gives the exact prediction of the data-driven model for a single observation. Such an interpretation allows for local predictions to be understood with some uncertainty in the feature effect as the explanation model (i.e. weighted LR) is trained on data that will always contain noise. Global interpretation can be obtained by producing multiple random local interpretations and observing the feature's effect by either a boxplot of the SHAP-generated values or by plotting SHAP values against the feature's actual values.

3.2.5 Quantification

The Quantification module is invoked as the final step for performing the M&V analysis by the MVBEP tool. In this module, the developed data-driven model is used to determine the baseline energy performance of the post-retrofitted building and ultimately the energy savings from the retrofit project. Specifically, energy use predictions are made by the developed data-driven model using post-retrofit input data. Then, the energy savings are estimated by subtracting from the best model's predictions the actual building energy consumption metered after the completion of the retrofit project.

3.3 Application to a Medium Office in Boulder, CO

To demonstrate the described process in Section 3.2, this section applies the MVBEP framework on a synthesized data. The synthesized data is obtained from the End-Use Load Profiles (EULP) for the U.S. [118]. The EULP database is developed by using residential and commercial buildings' physics-based models. The models were calibrated and validated against empirical datasets of actual energy use. For each building model, a one-year dataset of whole building energy consumption and end-uses is provided with a 15- min frequency. Currently, the EULP database has 347,144 building models with different typologies and shapes. With a combination of 3,089 AMY data from different U.S. counties, the database has approximately 550,000 commercial buildings that represent 1.8 million different commercial buildings. However, to apply the MVBEP tool one-year data is not sufficient as the analysis period is segmented into two periods: the pre-retrofit period used to train and test data-driven models as well as the post-retrofit period when the models

are used to estimate the retrofit savings. Additionally, the reported dataset does not include a building case that has been retrofitted. Hence, one building model is selected and used to simulate an additional year after 2018 (i.e., the AMY used for the EULP dataset). The building selected consists of a medium office space located in Boulder, CO, with the characteristics summarized in Table 3.2. Figure 3.3 shows a 3-D rendering of the office building from OpenStudio software [119].

The office building model is simulated by using weather data that spans two years from 2018 to 2019. The simulation results are obtained for 15-min whole-building energy use as well as for weather parameters. For the second year (i.e., 2019), the building model is adjusted to represent a retrofitted HVAC system. Specifically, the HVAC retrofit consists of upgrading the chiller, the boiler, and the water heater. For this retrofit, the building model is adjusted by changing the chiller's Coefficient of Performance (COP) as well as the thermal efficiencies of the boiler, and the water heater. These changes are compared to the baseline settings as shown in Table 3.3. The building simulation is performed for 2019 with the changed values listed in Table 3.3 to generate the retrofitted building energy consumption. For this application, the retrofitting date is set to be on the 1st of April 2019 which delineates the retrofit from the pre-retrofit periods. Figure 3.3 shows the segmentation of the data into three groups. The lines representing pre-retrofit and post-retrofit whole-building energy consumption are the required data to apply the MVBEP tool.

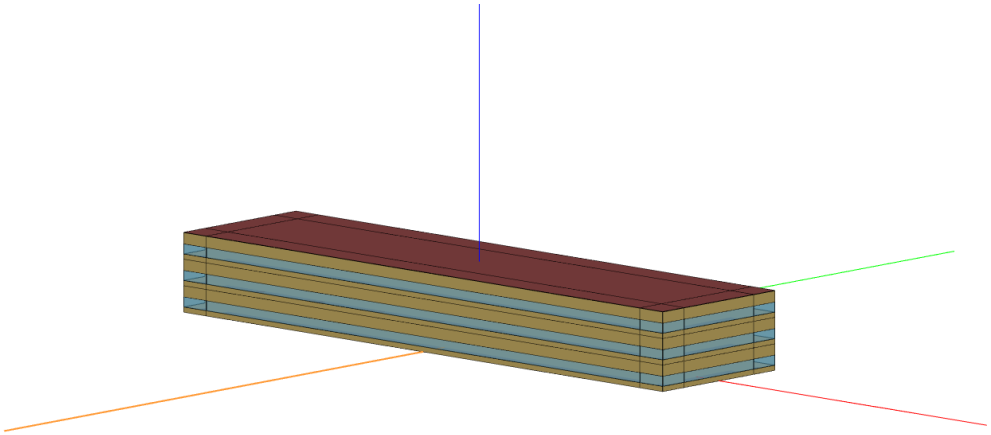


Figure 3.3: Simulated office rendered image from OpenStudio

Table 3.2: Main characteristics of the simulated office building in Boulder, CO

Variable Category	Variable	Value
Building shape	Total floor area (m^2)	7,000
	Number of floors	3
	Floor to ceiling height (m)	4
	Length to width ratio	4
Building typology	Activity	Office
	weekday operating hours	9
	weekday opening time	5:45 AM
	weekend operating hours	15.75
	weekend opening time	7:30 AM
	Occupant density (per/m^2)	0.054
Building systems	Chilled-water loop	FCU with Water-cooled Chiller and cooling tower
	Hot-water loop	FCU with Boiler
	Cooling source	Electricity
	Heating source	Natural gas
	Lighting density (W/m^2)	7.3
	Equipment Load (W/m^2)	10.8
Building weather	Location	Boulder, CO U.S.
	ASHRAE Zone 2004	5B
	Climate Zone	Cold
	Weather file 2018	CO Buckley
	Weather file 2019	Broomfield Jeffco Boulder
Building envelope	Reference	Office - ASHRAE 169-2013-5B

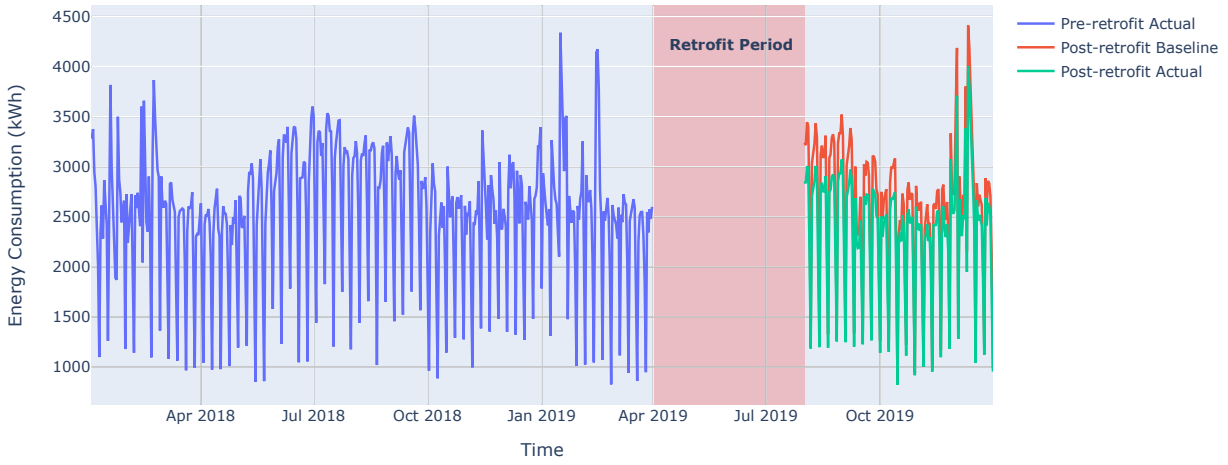


Figure 3.4: Building energy consumption segmentation before and after retrofitting

Table 3.3: Efficiencies of the office building HVAC system for both pre-and post-retrofit periods

System	Variable	Pre-retrofit value	Post-retrofit value
Water-cooled chiller	COP	3.8	6
Boiler	Efficiency	0.8	0.95
Water heater	Efficiency	0.82	0.95

3.3.1 Initialization Processing

The initialization starts by creating an object of MVBEP tool to fit data-driven model using the office building data. The initiation process provides a summary results and a set of plots for analyzing the data and pinpointing any failures in meeting the data quality requirements. Table 3.4 shows the descriptive summary for the office building data while Figure 3.5 shows a timeseries plots of both outdoor dry-bulb temperature and whole-building energy consumption with a daily frequency. The plot shows the segmentation of the training, testing, and quantification periods where the training period is one year of the pre-retrofit energy consumption and the testing period is three months. The main statistical features of the variables part of the dataset are summarized in Table 3.4. These variables include:

- t: Time (15-min).
- E: Energy (kWh).
- T_{dry}: Dry-bulb temperature (°C).
- RH: Relative humidity (%).
- W_v: Wind speed (m/s).
- W_D: Wind direction W_D.
- GHI: Global horizontal radiation (Wh/m^2).
- DNI: Direct normal radiation (Wh/m^2).
- DIF: Diffused horizontal radiation (Wh/m^2).

Table 3.4: Main characteristics of the variables of the Office building dataset

Statistics	t	E (kWh)	T_{dry} (°C)	RH (%)	W_v (m/s)	W_D	GHI (Wh/m^2)	DNI (Wh/m^2)	DIF (Wh/m^2)
Mean	-	25.42	10.53	50.76	3.84	168.49	196.65	248.35	56.40
Min	2018/1/1	5.52	-23.3	2	0	0	0	0	0
Q1	2018/7/4	17.59	2.2	30.02	2.6	90	0	0	0
Q2	2019/1/1	23.10	10.6	48.57	3.6	180	11	1	8
Q3	2019/7/1	33.45	18.3	69.79	5.1	220	348.5	507	83.5
Max	2019/12/31	96.05	37.5	100	21.6	360	1087	1079	522
Std	-	10.58	10.84	24.61	2.29	100.27	281.61	354.80	85.34

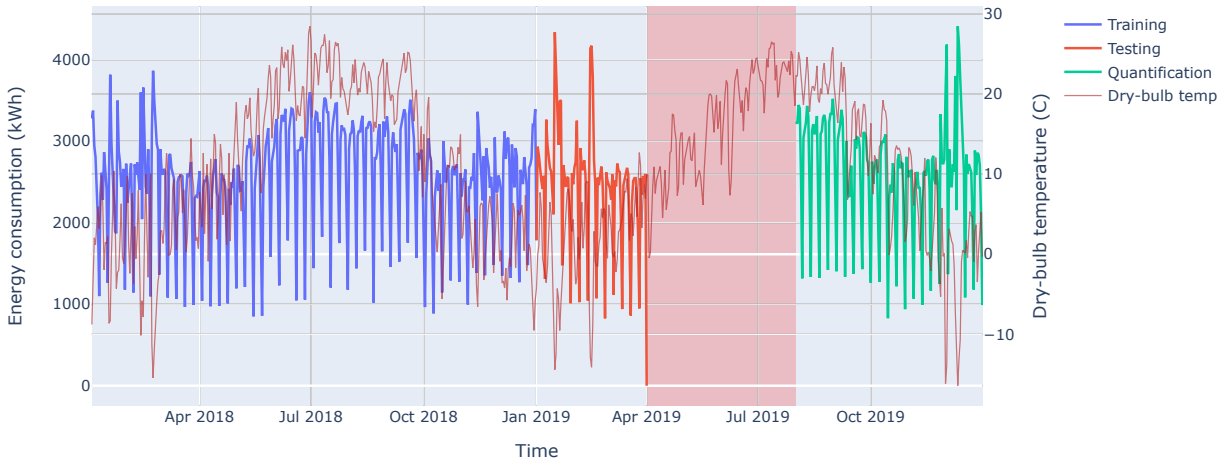


Figure 3.5: Office building energy consumption and outdoor dry-bulb temperature vs time

3.3.2 Transformation Processing

After the initial processing of the dataset including checking that all the data meet the quality requirements, the MVPEB tool invokes the transformation process to clean and prepare the data in order to be suitable for training data-driven models. The process of transformation depends on the frequency of the data and downsampling (i.e. reducing the frequency). For the case of office building case study, the

data has a 15-min frequency resulting in 4 transformation options considered by the MVBEP tool including 15-min data with default features, hourly-data with default and TOWT features, and daily data with default features as indicated in Table 3.5. For the 15-min dataset, only timestamps are transformed. As noted in Table 3.5, the datasets and associated transformed features are used to train two types of data-drive models including group 1 and group 2 models as discussed in the following sections. For the hourly datasets, two sets of features are considered for the development of the data-driven models. The features for the hourly default dataset are the same as those used for the 15-min dataset except that the 15-min of hour feature is not included and that all the values for the other features are aggregated on an hourly basis such that the whole-building energy consumption values are obtained by summing the 15-min values while the other features are averaged. The second hourly dataset is transformed according to the TOWT method which considers only one time-related feature (i.e., hour of the week). The temperature values are segmented into bins of equal length which, for this case study consist of 6 bins. Occupancy is a feature that is determined following the TOWT approach as described by Mathieu et al. [34]. For the daily dataset, the outdoor dry-bulb temperature is transformed into heating and cooling degree-days (i.e., HDD and CDD) while the remaining features are averaged on a daily basis.

3.3.3 Development Processing

After each transformation processing of the dataset, the MVBEP tool considers a set of data-driven models based on the user selection and the dataset type. When all the available modeling approaches are selected by the user of the MVBEP tool for the 15-min dataset, the development process generates 15 data-driven models. The difference between the hourly datasets and other datasets for 15-min and daily data frequencies is the dataset that is used to train the LR model. Indeed, for the TOWT hourly dataset, MVBEP tool considers only the LR approach for training and testing the data-driven models. Specifically, a weighted LR model is considered for every hour of the week. Therefore, the resulting data-driven model for the TOWT hourly dataset has 168 factor variables to cover all hours for one week. For the 15-min and daily datasets, LR models are trained and developed based on the default features.

After completing the data transformation processing, the MVBEP tool identifies the best modeling

Table 3.5: Transformed features and datasets for the resulting transformations based on model group and timestamps frequency

Frequency	Dataset type	Feature category	Feature	Models group
15-min	Default dataset	t	15-min of hour Hour of day Hour of week Day of week Day of month Day of year	Group 1
		T_{dry}, RH, W_v, W_D, GHI, DNI, DIF		
Hourly	Default dataset	t	Hour of day Hour of week Day of week Day of month Day of year	Group 1
		T_{dry}, RH, W_v, W_D, GHI, DNI, DIF		
	TOWT dataset	t	Hour of week	TOWT LR
		T_{dry}	Segment 1 $[T_{dry,1,s}, T_{dry,1,e}]$ Segment 2 $[T_{dry,2,s}, T_{dry,2,e}]$ ... Segment 6 $[T_{dry,6,s}, T_{dry,6,e}]$	
		Occupancy		
Daily	Default dataset	t	Day of week Day of month Day of year	Group 1
		T_{dry}	HDD CDD	
		RH, W_v, W_D, GHI, DNI, DIF		

approach and timestamps frequency based on either CV(RMSE) or NMBE values. Additionally, the MVBEP tool saves the results of the other models for all datasets for comparative analysis to give further insights into the performance of each data-driven model. The accuracy metrics for the data-driven models considered by the MVBEP tool for the office building case study are shown in Table 3.6. Based on the CV(RMSE) metric, the MVBEP tool indicates that the XGB-based model trained with a daily dataset is the most accurate model for the simulated office building energy consumption. Moreover, Table 3.6 indicates that the training times required for KNN and SVR models are the longest among all the models considered by the MVBEP tool. In general, the values of CV(RMSE) decrease as the frequency of the timestamp decreases which might be due to high-frequency datasets containing more noise. For daily frequency, all models achieve similar training and

testing performance metrics except for the XGB-based model which demonstrates a large difference between the training and testing prediction accuracies. This result might be because the XGB-modeling approach overfits the training data. In addition, the variations in the sign of the NMBE values indicate that some models tend to underestimate the building's post-retrofit baseline energy consumption while other models overestimate the building energy demand. The lowest NMBE value is obtained with the KNN-based model trained with the daily dataset. However, NMBE metrics should not be used solely to make the selection of the most data-driven model to predict building energy consumption as explained in Section 2.7.

Table 3.6: Training and Testing prediction accuracies for various data-driven models for the office building in Boulder, CO

Frequency	**Models**	**Training**		**Testing**		**Training Time (s)** [1]
		CV(RMSE)	**NMBE**	**CV(RMSE)**	**NMBE**	
15-min	LR WLS	31.39	0.00	39.04	5.86	0.077
	RF	5.22	-0.00	20.86	-0.90	1.074
	XGB	9.85	-0.12	18.46	-0.40	1.95
	SVR	16.94	1.53	25.99	2.19	512
	SLP	16.80	-0.09	25.73	0.61	34.75
	KNN	17.77	0.61	30.15	-2.08	531
Hourly	LR TOWT	14.87	0.00	20.23	0.87	0.298
	RF	5.25	0.00	18.29	-0.57	0.31
	XGB	8.50	-0.11	17.15	-0.22	0.318
	SVR	18.48	2.23	27.60	4.17	7.798
	SLP	19.91	0.12	27.74	-1.12	5.727
	KNN	17.26	0.66	29.30	-2.17	32.723
Daily	LR WLS	20.96	-0.00	23.98	6.06	0.015
	RF	5.34	-0.01	10.46	2.39	0.108
	XGB	4.33	-0.09	9.01	1.44	0.054
	SVR	22.34	-4.87	25.17	-1.92	0.035
	SLP	20.82	-0.04	23.82	4.20	0.41
	KNN	17.56	0.00	23.53	4.08	0.539

[1] The training process was performed on a Laptop with 8 GB RAM and Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz

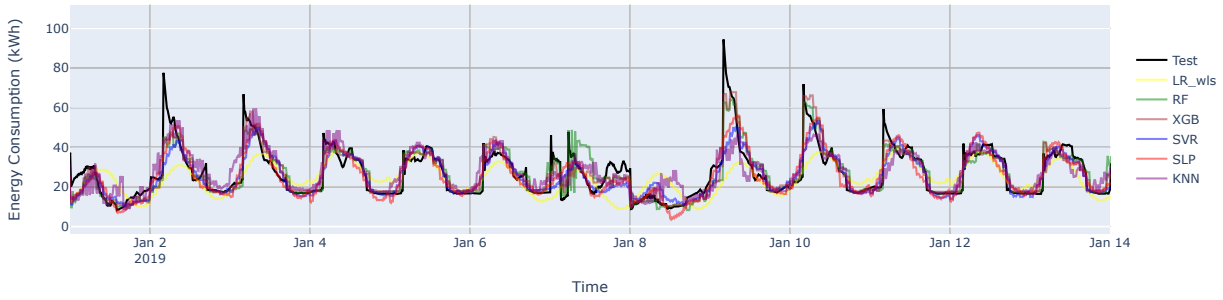
To better understand the prediction performance of various data-driven models considered by the MVBEP tool, timeseries plots are generated for all models and datasets for one week during the testing period as shown in Figure 3.6. For the 15-min dataset, the prediction accuracy metrics exhibit significant fluctuations among various models since these models are subject to larger and more noisy data than for other data frequencies (i.e., hourly and daily). The variations in performance metrics among models are

significantly lower for hourly and especially daily datasets. Indeed, for daily dataset, the predictions including the CV(RMSE) values for all models are closer between training and testing periods resulting in better prediction accuracies as depicted in Figure 3.6. However, moving from hourly to daily predictions reduces the level of information provided by the models about the energy performance of the building during a specific day. LR-based models that are trained on the default and TOWT datasets have less variance as they predict the general pattern of building energy consumption. On the other hand, complex models can predict better the building energy performance but with higher tendencies to either underestimating or overestimating during certain periods.

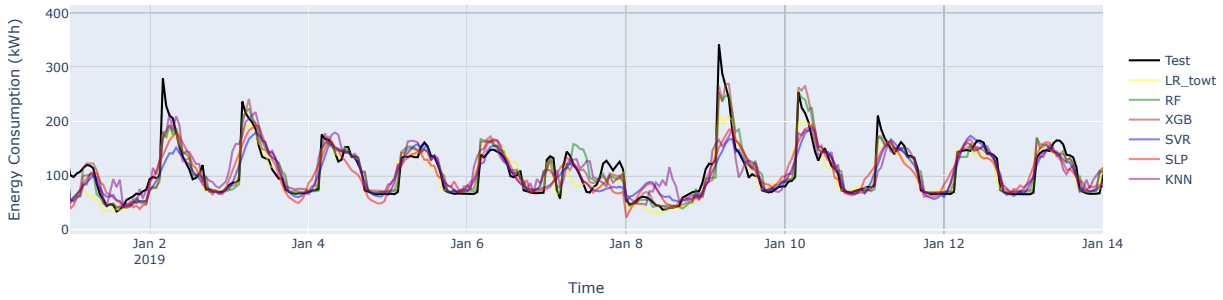
3.3.4 Interpretation Processing

As noted earlier, it is generally difficult to determine the contribution of various features in a data-driven model's predictions, especially for complex models. Given that a XGB-based model gave the best CV(RMSE) value for predicting the office building energy consumption, it is the chosen model to perform an interpretation processing by the MVBEP tool. Similar to the transformation processing, the interpretation analysis depends on the dataset frequency. In this section, the interpretation processing is carried out for the XGB-based model trained and tested using the daily dataset. Figure 3.7 shows the results of a global interpretation analysis of the XGB-based model developed to predict daily energy consumption for the office building located in Boulder, CO. The SHAP values shown in Figure 3.7 provide the contributions of individual features in the model's predictions. The global importance of a feature is generated by obtaining SHAP values for multiple random data points in the testing dataset resulting in boxplots that highlight the distributions of various features' SHAP values. The higher the SHAP absolute value, the stronger the importance of the feature in the model predictions. Figure 3.7 shows that the most important features for XGB-based model predictions include HDD, day of week sine, and cosine values. Given that the testing data include mostly building energy consumption during the winter, the HDD feature is expected to have more importance than the CDD feature. Other meteorological values showed distributions with small SHAP values indicating that their influence on the model's predictions is minimal.

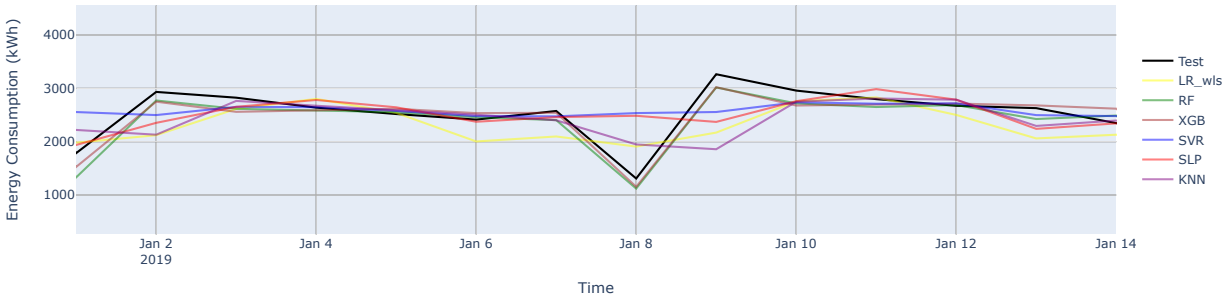
For local interpretation analysis, a series of consecutive predictions are evaluated as shown in Figure



(a) Timeseries for 15-min frequency



(b) Timeseries for hourly frequency



(c) Timeseries for daily frequency

Figure 3.6: Predictions of whole-building energy consumptions using (a) 15-min, (b) hourly, and (c) daily frequencies from various data-driven modeling approaches using one week of the testing data

3.8. The plots in Figure 3.8 compare the actual and predicted values for daily building energy consumption

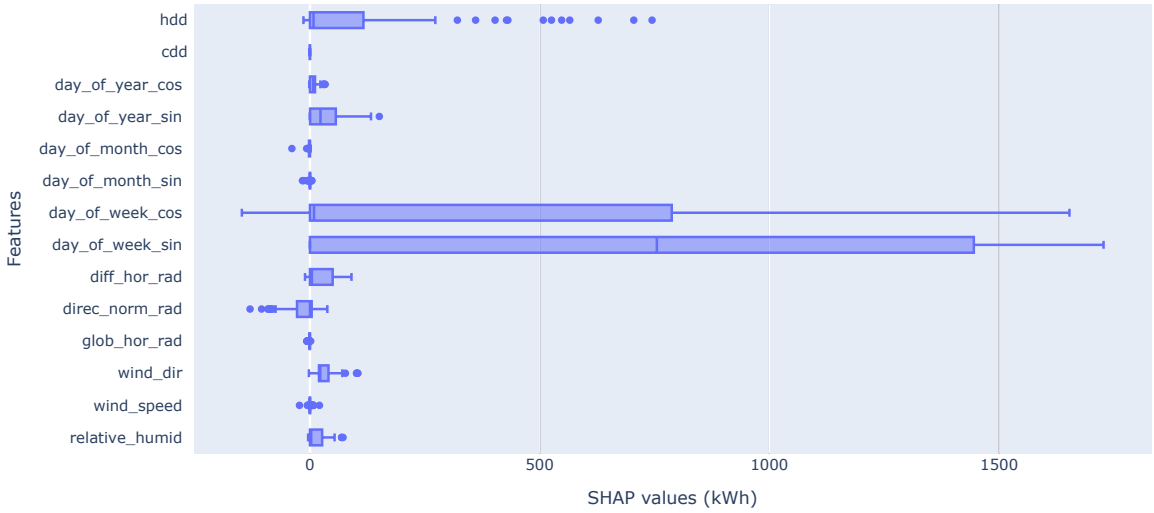


Figure 3.7: Results of a global interpretation for the XGB-based model for daily predictions of energy consumption for an office building located in Boulder, CO

as well as the SHAP contribution values for each feature of the daily XBG-based model. The summation of SHAP values at any timestep provides the model's prediction for the building energy consumption.

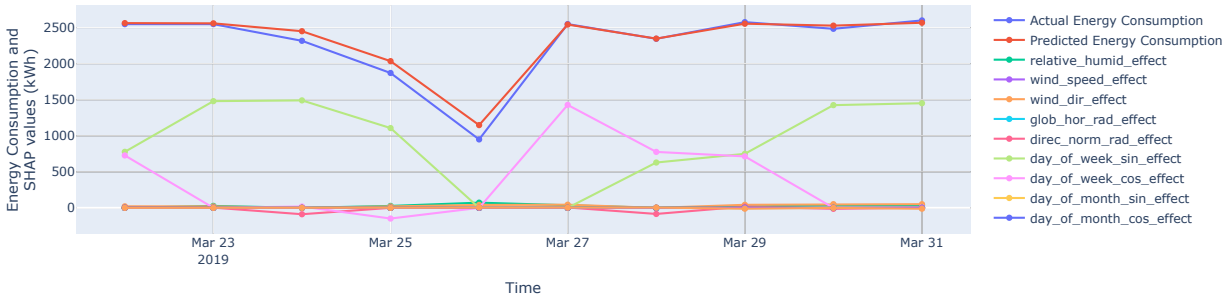


Figure 3.8: Results of a global interpretation for the XGB-based model for daily predictions of energy consumption for an office building located in Boulder, CO

For models using hourly and 15-min datasets, the outdoor dry-bulb temperature variations are plotted against the building energy consumption along with the XGB-based model feature SHAP values as shown in Figure 3.9. Specifically, Figure 3.9 shows a random sample of 1,000 observations of outdoor dry-bulb temperature and building energy consumption extracted from the testing dataset. The SHAP values indicate

that outdoor dry-bulb temperatures that are lower than 0 °C or higher than 17 °C have high importance in the prediction of the XGB-based model for the building energy consumption.

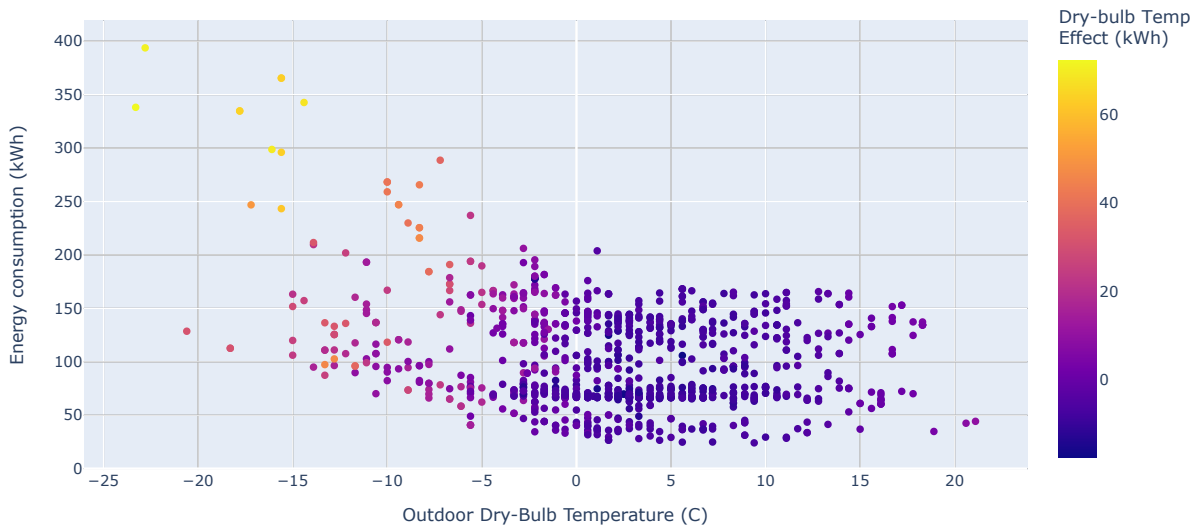


Figure 3.9: Variations of outdoor dry-bulb temperature against energy consumption with feature SHAP values using XGB-based model developed based on hourly dataset

3.3.5 Quantification of Energy Savings

When the MVBEP tool develops a satisfactory data-driven model to estimate the baseline building energy consumption during the post-retrofit period, the quantification of energy savings achieved by the retrofit project can be considered. The quantification process is initiated by the MVBEP tool by providing the developed and selected data-driven model using the pre-retrofit dataset with the post-retrofit dataset. This post-retrofit dataset needs to have the same features and frequency as the pre-retrofit dataset used to train and test the developed data-driven model. For the case of the office building in Boudler, CO, the energy savings or avoided energy use (i.e., AEU) for the post-retrofit period ranging between August 1 and December 31 in 2019 is estimated by the selected best model (i.e. XGB with daily dataset) 35,761 kWh using Equation 3.4. The actual energy savings from the retrofit project or AEU is determined to be 37,858 kWh obtained through the simulation results of the office building with and without the retrofit changes.

Specifically, the actual AEU is calculated by subtracting the energy consumption obtained for the retrofitted building model (i.e., with retrofit changes) from the energy demand achieved by the baseline model (i.e., using the baseline settings without the retrofit changes) during the period between 2019/08/01 and 2019/12/31.

$$AEU = \sum_{t}^{n} (E_{\mathrm{PB},t} - E_{\mathrm{A},t}) \tag{3.4}$$

where

- AEU: Avoided energy use (kWh).
- $E_{\mathrm{PB},t}$: Predicted building baseline post-retrofit energy consumption (kWh).
- $E_{\mathrm{A},t}$: Actual post-retrofit building energy consumption (kWh).

As discussed earlier, the selection of the best data-driven model developed by the MVBEP tool is based on two performance metrics (i.e. CV(RMSE) and NMBE). For the office building located in Boulder, CO, the data-driven model that provides the best AEU estimate is based on the CV(RMSE) since using only the NMBE values can be inaccurate as the NMBE metric does not prevent error cancellation and may provide erroneous AEU estimation for periods other than the testing period. Thus, the selection of the best data-driven model needs to be based on both NMBE and CV(RMSE) values. Reddy et al. [120] have introduced a Goodness-of-Fit (GOF) metric that combines both CV(RMSE) and NMBE metrics as expressed by Equation 3.5. The constants $w_{CV(RMSE)}$ and w_{NMBE} represent the weights for both CV(RMSE) and NMBE metrics. In the application of the GOF, Reddy et al. chose a 1:3 weighting ratio for ($w_{CV(RMSE)}$:w_{NMBE}) which favors having low bias over low variance. For this study, both metrics are weighted equally when comparing the achieved savings for the models.

$$GOF = \sqrt{\frac{w_{CV(RMSE)}^2 CV(RMSE)^2 + w_{NMBE}^2 NMBE^2}{w_{CV(RMSE)}^2 + w_{NMBE}^2}} \tag{3.5}$$

Figure 3.10 shows the GOF metrics against the percentage difference between the estimated and actual AEU values with symbols indicating the developed models and colors showing the frequency. For each data-driven model, the AEU values are obtained by using Equation 3.4 for the predicted baseline

energy consumption for the post-retrofit period data between 2019/08/01. Figure 3.10 shows that there is a correlation between achieving a low GOF and low absolute percentage difference between the estimated and true AEU. Table 3.7 summarizes the results shown in Figure 3.10 with the actual energy difference between the baseline and post-retrofit is known to be 37,858 kWh.

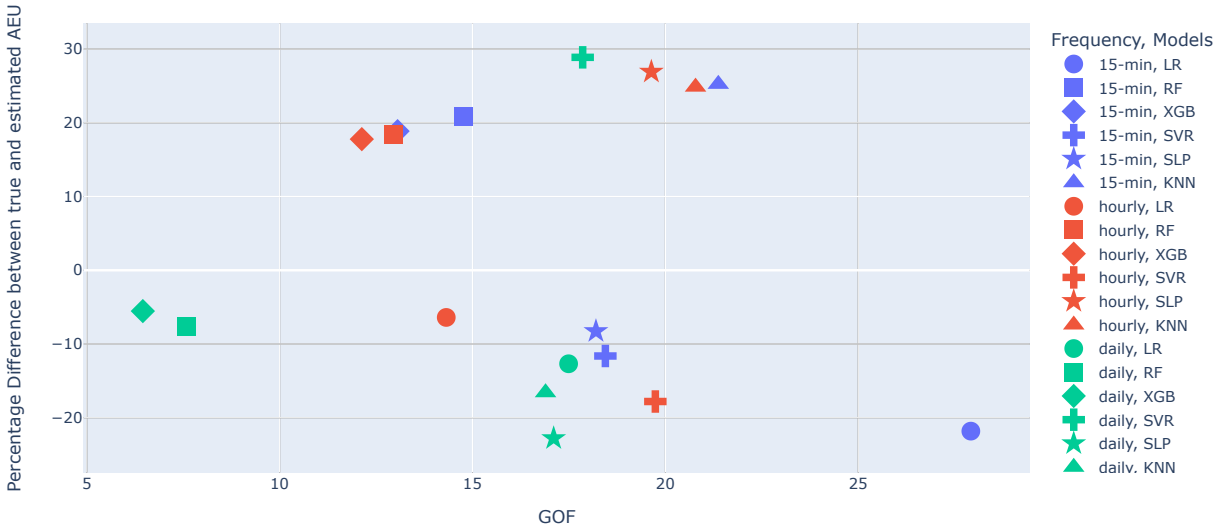


Figure 3.10: Variations of GOF with AEU estimation accuracy for various data-driven models

Table 3.7: Variations of both CV(RMSE) and NMBE with AEU estimation accuracy for various data-driven models

Frequency	Models	CV(RMSE) (%)	NMBE (%)	GOF (%)	Savings (kWh)	Diff % [1]
15-min	KNN	30.15	-2.09	21.37	47413	25.24
	LR	39.04	5.87	27.92	29597	21.82
	RF	20.86	-0.90	14.77	45754	20.86
	SLP	25.73	0.61	18.20	34726	8.27
	SVR	25.99	2.20	18.44	33455	11.63
	XGB	18.46	-0.40	13.06	45000	18.87
Hourly	KNN	29.31	-2.17	20.78	47269	24.86
	LR	20.23	0.88	14.32	35436	6.40
	RF	18.30	-0.58	12.94	44807	18.36
	SLP	27.75	-1.12	19.64	48045	26.91
	SVR	27.60	4.17	19.74	31118	17.80
	XGB	17.15	-0.22	12.13	44589	17.78
Daily	KNN	23.54	4.08	16.89	31577	16.59
	LR	23.98	6.07	17.49	33056	12.68
	RF	10.47	2.39	7.59	34964	7.64
	SLP	23.82	4.20	17.10	29229	22.79
	SVR	25.18	-1.93	17.86	48790	28.88
	XGB	9.01	1.44	6.45	35761	5.54

[1] The values are obtained by calculating the percentage differance between the true AEU (i.e the difference between the simulated baseline and retrofitted building) and the estimated AEU by the resulting models

CHAPTER 4

MVBEP APPLICATIONS

Data-driven models are developed in a process that trades between multiple aspects to achieve an accurate prediction accuracy. Even with a successful application of a data-driven model that produced a satisfactory prediction accuracy for a specific building, there is no guarantee that the same modeling approach provides a consistent level of prediction accuracy when applied to other datasets even specific to the same building and climate. Therefore, testing the MVBEP tool should be carried out over several case studies involving a wide range of buildings with different climates, building typologies, and varying quality levels in terms of coverage and data range. A suggested approach to generate a clean but synthetic dataset for building energy performance is by using physics-based simulation tools to model several building types with different features and locations. However, the more common and practical approach is to monitor actual energy consumption of existing buildings with records of Actual Meteorological Year (AMY) data. In this chapter, the performance of MVBEP tool in developing data-driven models is demonstrated using measured datasets. This chapter applies the developed MVBEP framework outlined in Chapter 3 to a dataset to test the generalization of the MVBEP framework in predicting energy consumption for different building types and climates. The used data for MVBEP applications is described in Section 4.1. Section 4.2 details the analysis carried out for hyperparameter tuning as part of improving the performance of data-driven models generated by the MVBEP tool. Finally, Section 4.3 demonstrates several features of the MVBEP tool over multiple case studies to better understand the ability of the tool to generate suitable data-driven models for various climates, building typologies, time ranges, and training periods.

4.1 The Building Data Genome 2 (BDG2) Data-Set

BDG2 is an open dataset of approximately 1,600 buildings with hourly measurements of energy consumption and meteorological data spanning at least one year. Some buildings have several meters to measure various energy end-uses. Miller et al. [121] collected the dataset from 11 sites with some data obtained from publicly available sources while the remaining data was collected via private channels. The weather data for the sites are obtained from nearby weather stations through the National Centers for Environmental Information (NCEI). Miller et al. [121] have processed and cleaned the data by removing outliers, removing duplicate values, and ensuring that reported units are consistent throughout the whole dataset. Table 4.1 summarizes the main features of the BDG2 dataset including the numbers of buildings and meters as well as the locations and associated climate zones. However, to make sure that the data is suitable for training data-driven models, buildings selected for use with the MVBEP tool have to meet the requirements listed and described in Section 3.2. Based on these requirements, out of 1578 buildings, only 601 buildings met the data quality considerations and are used for further analysis with the MVBEP tool as shown in Table 4.1.

Table 4.1: Main features of the BDG2 dataset (Source: [121])

Site	Actual Site Name	Location	Climate	Buildings
Bear	Univ. of California - Berkeley	Berkeley, CA	3C	34
Bobcat	Anonymous	N/A	5B	1
Bull	Univ. of Texas Austin	Austin, TX	2A	4
Eagle	Anonymous	N/A	4A	65
Fox	Arizona State Univ. (ASU)	Tempe, AZ	2B	118
Hog	Anonymous	Anonymous	6A	132
Lamb	Cardiff - City Buildings	Cardiff, UK	4A	83
Panther	Univ. of Central Florida (UCF)	Orlando, FL	2A	27
Peacock	Princeton University	Princeton, NJ	5A	31
Rat	Washington DC City Buildings	Washington DC	4A	7
Robin	Univ. College London (UCL)	London, UK	4A	50
Swan	Anonymous	N/A	3C	15
Wolf	Univ. College Dublin (UCD)	Dublin, Ireland	5A	34

4.2 Hyperparameter Tuning

Whenever developing a ML model, it is always encountered to make design choices that define a model's architecture. Many times, an optimal selection of such choices is not possible without iterating over several combinations of the model's hyperparameter which is computationally expensive. This section aims to test whether finding the optimal set of hyperparameters for the modeling approaches in Section 2.3 is possible or not. Therefore, this section shows an analysis of hyperparameter tuning effect on prediction accuracy and tries to obtain a set of hyperparameters that generally increase the prediction accuracy. The used dataset for testing the effect of hyperparameter tuning is BDGP 2 given that it covers different buildings from different climates. 10 random buildings in the dataset were selected to minimize the unnecessary training time while maintaining a sufficiently large sample of buildings data as shown in Table 4.2. The selection of hyperparameters is not intended for picking the perfect set of hyperparameters but rather a better set of default hyperparameters other than the ones set by Sklearn library [29]. As highlighted in Section 3.2, the mentioned hyperparameters in Table 3.1 are one of the important hyperparameters to test in most cases of ML. Therefore, this approach will focus on such hyperparameters.

As part of the development of any data-driven model, the best features and hyperparameters have to be defined. The optimal selection of the model's structure including its hyperparameters requires extensive computational efforts especially when the modeling approach is complex. This section aims to assess if the optimal selection of hyperparameters is feasible and practical for the various MVBEP tool's modeling approaches listed in Section 2.3. Therefore, this section determines the effects of hyperparameter tuning on the prediction accuracy of various modeling approaches considered by the MVBEP tool and tries to obtain a set of hyperparameters that generally increase the prediction accuracy. The used dataset for testing the effects of hyperparameter tuning is BDGP 2 which covers several buildings from different climates. For this analysis, data for 10 buildings are selected randomly from the BDGP 2 dataset to minimize unnecessary training time while maintaining a sufficiently large sample to represent all the buildings in the BDGP 2 dataset. Table 4.2 shows the information about the randomly selected data where each building has historical data of 2 years. The selection of hyperparameters is not intended for identifying the best set of hyperparameters for

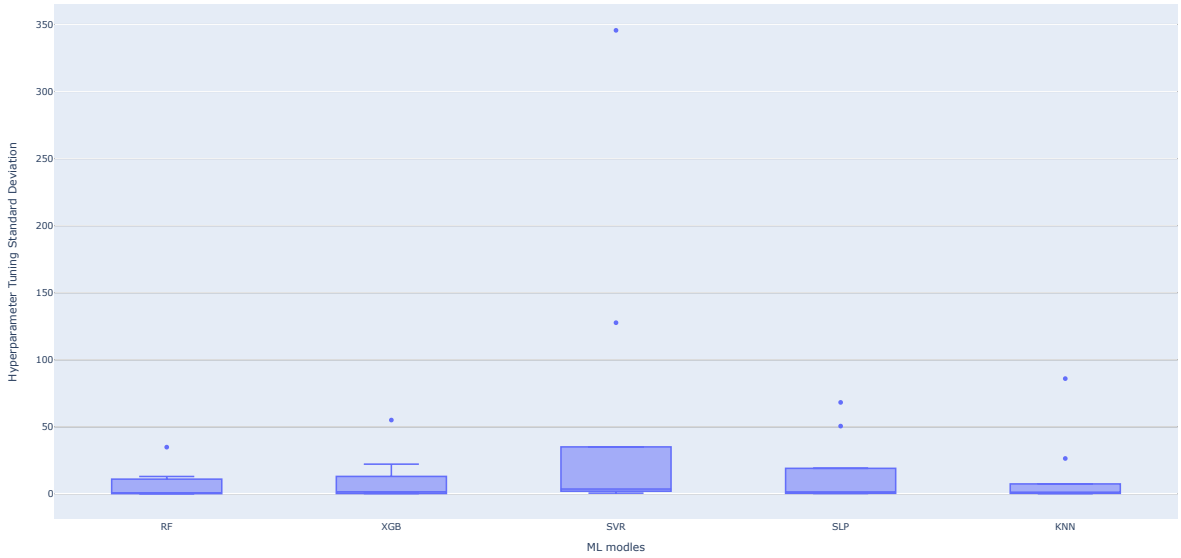
each model approach but rather to determine better default hyperparameters than those considered by the Sklearn library [29]. As highlighted in Section 3.2, the hyperparameters listed in Table 3.1 are previously identified hyperparameters for the modeling approaches considered by the MVBEP tool and will be further evaluated in this section.

Table 4.2: Main features of the BDG2 dataset for hyperparameter tuning

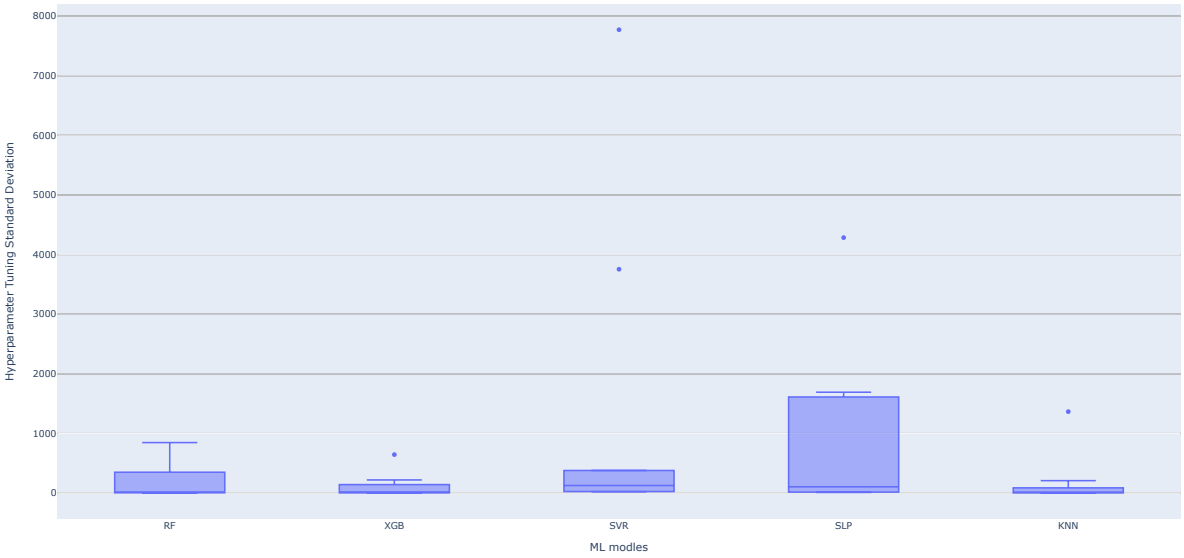
Site	Building	Actual Site Name	Location	Climate
Bear	Public Jocelyn	Univ. of California - Berkeley	Berkeley, CA	3C
Fox	Assembly Lakeisha	Arizona State Univ. (ASU)	Tempe, AZ	2B
Fox	Assembly Sheldon	Arizona State Univ. (ASU)	Tempe, AZ	2B
Fox	Food Francesco	Arizona State Univ. (ASU)	Tempe, AZ	2B
Fox	Office Demetrius	Arizona State Univ. (ASU)	Tempe, AZ	2B
Fox	Warehouse Pearl	Arizona State Univ. (ASU)	Tempe, AZ	2B
Hog	Office Elke	Anonymous	Anonymous	6A
Hog	Public Brad	Anonymous	Anonymous	6A
Robin	Education Julius	Univ. College London (UCL)	London, UK	4A
Wolf	Office Emanuel	Univ. College Dublin (UCD)	Dublin, Ireland	5A

To evaluate the effectiveness of the hyperparameter tuning process on the prediction accuracy metrics (i.e., RMSE) of the data-driven models included in the MVBEP tool, the analysis is carried out using datasets from 10 buildings selected from the DBGP 2 database. The RMSE standard deviation distribution is obtained for all combinations of data-driven models in every building and available dataset frequency. Figure 4.1 shows the analysis results using two boxplots of RMSE standard deviation based on each modeling approach and data frequency considering all 10 buildings. Models based on the SLP approach show a wider distribution compared to the modeling approaches when the dataset has a daily frequency. For hourly datasets, the modeling approaches provide similar RMSE standard deviation distributions. This result indicates that all the modeling approaches' combinations in the hyperparameters grid search can provide significant prediction accuracy differences. Moreover, the standard deviation boxplots illustrate that all models achieve different prediction accuracy levels when hyperparameter tuning is performed.

The result of the hyperparameter tuning process is a data table that shows several metrics about each combination of hyperparameters. To quantify the effect of a certain hyperparameter value, a simplistic approach is followed in the following analysis. For each combination, the model is tested by using rolling cross-validation as explained in Section 3.2. The combinations are ordered based on the RMSE column



(a) RMSE variance distribution for hourly frequency data



(b) RMSE variance distribution for daily frequency data

Figure 4.1: RMSE variance distribution in hyperparameter tuning grid search results

which shows the mean test score for every 5 folds. To quantify the effect of having a specific value for a certain hyperparameter, the difference percentage $RMSE\ diff\ \%$ between the hyperparameters combination $RMSE_i$ and the best score $RMSE_B$ is obtained as shown in Equation 4.1. Followingly, the average difference

percentage is obtained for all combinations containing that specific hyperparameter's value in a single building for every building.

$$RMSE\ diff\ \% = \frac{RMSE_i - RMSE_B}{RMSE_B} * 100 \tag{4.1}$$

4.2.1 Random Forest

Figure 4.2 shows the RMSE percentage difference boxplots for the RF selected hyperparameters (i.e., number of trees, bootstrapping, and minimum leaf samples). The number of trees in the forest does not seem to change the prediction accuracy levels for models using datasets with both hourly and daily frequencies. However, performing bootstrapping on the trained dataset significantly reduces the RMSE metric for both hourly and daily datasets. Moreover, the minimum number of samples in a leaf does not significantly reduce the performance when the frequency is hourly but the RMSE difference distribution became narrower for the daily datasets. The recommendation based on the hyperparameter tuning results is to set the default number of trees to 100 and perform bootstrapping when training RF-based models. The minimum number of samples that can be left at a leaf is set to the default value of 3 as shown in Table 3.1.

4.2.2 Extreme Gradient Boosting

Figure 4.2 indicates that the mean RMSE difference percentage distribution becomes wider when the boosting rounds are increased from 100 to 200 and 500. A 0.4 value for Gamma parameter results in shifting the RMSE percentage difference distribution down which suggests an increase in the required split minimum loss from 0 to 0.4 generally improves the prediction accuracy of the XGB-based models. Finally, the boosting learning rate of 0.05 provides a narrower and lower distribution with few outliers compared to a learning rate of 0.3. Based on the analysis results, it is recommended to use 100 for boosting rounds, 0.4 for the gamma value, and 0.05 for the learning rate as the default hyperparameters for XGB-based models.

4.2.3 Support Vector Machine

Figure 4.4 illustrates two boxplots for mean RMSE percentage difference distribution for both the regularization parameter and the SVR kernel. The use of 0.1 as the value for the regularization parameter produced a lower and narrower RMSE percentage difference distribution compared to using 1 and 10. This result suggests that based on the 10-building sample, a higher regularization produces better prediction accuracy for the SVR-based models. The RBF kernel produced a significantly better prediction performance than the polynomial kernel which indicates that polynomial mapping does not generate accurate SVR-based models based on the buildings sample. Therefore, the default values for developing a SVR-based model are a C value of 0.1 and a RBF kernel.

4.2.4 Single-Layer Perceptron

The boxplots in Figure 4.6 show the mean RMSE percentage difference distribution of the hidden layer's number of hidden neurons, the optimization algorithm, and the initial learning rate. The initial learning rate of 0.00055 and 0.001 produces similar distributions for hourly data while for daily data, the distribution gets lower and narrower. Similarly, using 13 and 17 hidden neurons generates a trade-off where a certain frequency's prediction accuracy is improved while the other is reduced. The use of a SGD produces a significantly lower and narrower distribution for daily data while with hourly data, the distributions are similar. The choice of optimal hyperparameters, in this case, is not straightforward given that there is not a single set that suggests an improvement over all the cases. The default initial learning rate is set to 0.00055 as 0.001 does not provide a significant improvement with daily frequency data. For the number of hidden neurons, a value of 13 is selected as the default hyperparameter's value for the same reason. The default optimization algorithm is SGD as it significantly narrowed the distribution while maintaining a similar distribution for hourly frequency data.

The boxplots shown in Figure 4.6 indicate the mean RMSE percentage difference distributions that are associated with the number of neurons in the hidden layer, the optimization algorithm, and the initial learning rate. The initial learning rates of 0.00055 and 0.001 provide similar distributions for the hourly datasets while the RMSE percentage difference distribution gets lower and narrower for the daily datasets.

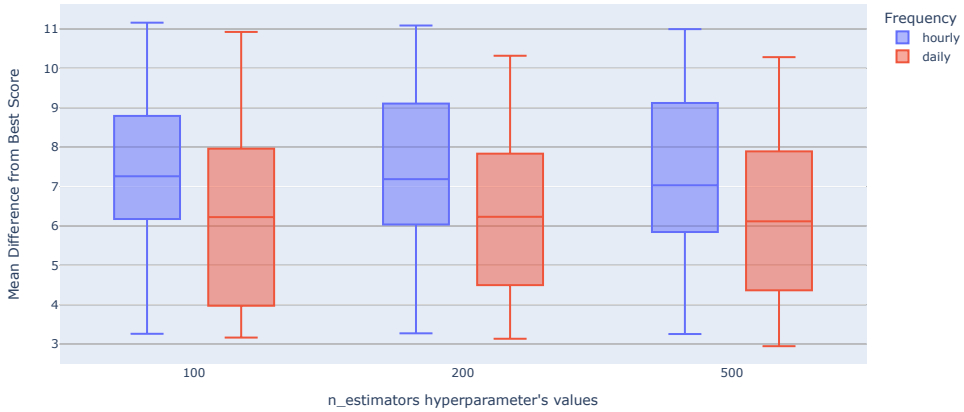
Similarly, using 13 and 17 hidden neurons generates a performance trade-off for the SLP-based models since the prediction accuracy is improved for one frequency but it is reduced for the other. The use of a SGD produces significantly lower and narrower distributions for daily datasets but similar distributions for hourly datasets. The choice of optimal hyperparameters for the SLP modeling approach is not straightforward given that there is not a single set of values that result in performance improvements for all the cases. The default initial learning rate is set to 0.00055 as 0.001 does not provide a significant improvement in the SLP-based models' performance when trained with daily datasets. For the number of hidden neurons, a value of 13 is selected as the default hyperparameter's value for the same reason. The default optimization algorithm is SGD as it significantly narrowed the distribution while maintaining a similar distribution for hourly datasets.

4.2.5 K-Nearest Neighbor

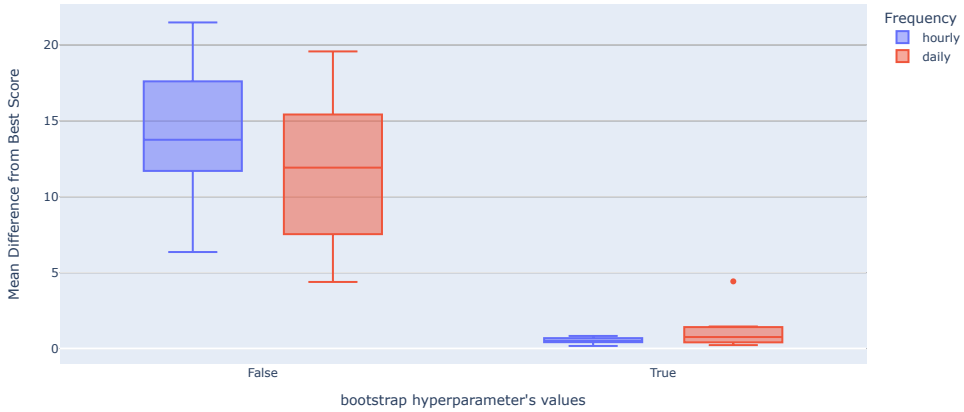
Although hyperparameter tuning is always performed for KNN, this analysis aims to check that the selected range is sufficient to achieve the optimum number of neighbors for the KNN-based models. Figure 4.5 shows a boxplot of mean RMSE percentage difference distributions for both the dataset frequency and the number of neighbors. The distributions become narrower as the number of neighbors is increased until 40 when the distributions tend to produce more outliers. To generalize the application of the KNN-based modeling approach in the MVBEP tool, the default selected range for neighbors is set between 5 and 70 with a step of 5.

4.2.6 Cross-Validation Method Impact

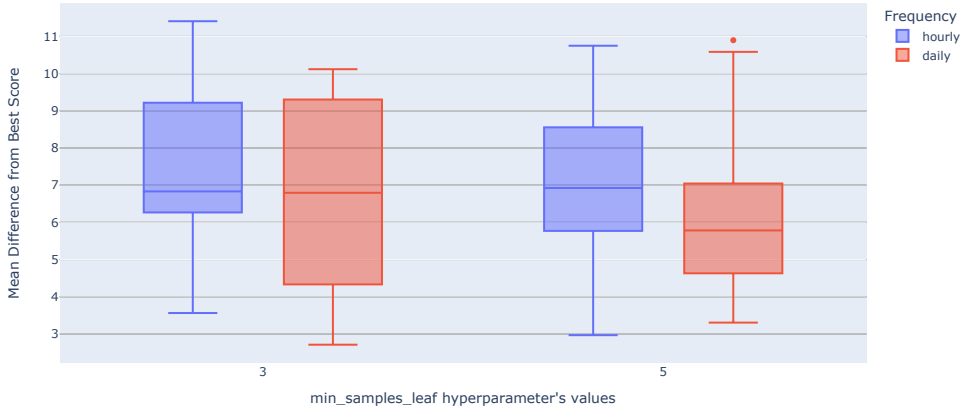
As highlighted in Section 3.2, the use of K-folds cross-validation can cause data leakage given that timeseries data usually has serial correlation between timestamps. The impact of choosing a specific cross-validation method is investigated by performing the same hyperparameter tuning analysis with K-folds cross-validation. Figures 4.2 through 4.6, show the hyperparameter tuning analysis results using rolling cross-validation while Figures 4.7 through 4.11 summarizes the results of hyperparameter tuning analysis using K-folds cross validation. For all the considered data-driven models, no significant changes are observed in terms of the best hyperparameter values that were obtained using rolling cross validation.



(a) RMSE percentage difference boxplot for number of trees

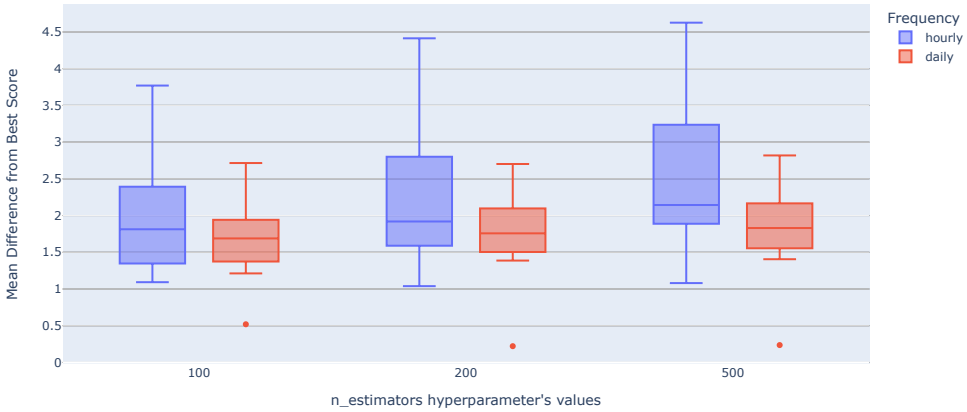


(b) RMSE percentage difference boxplot for including and excluding bootstrapping

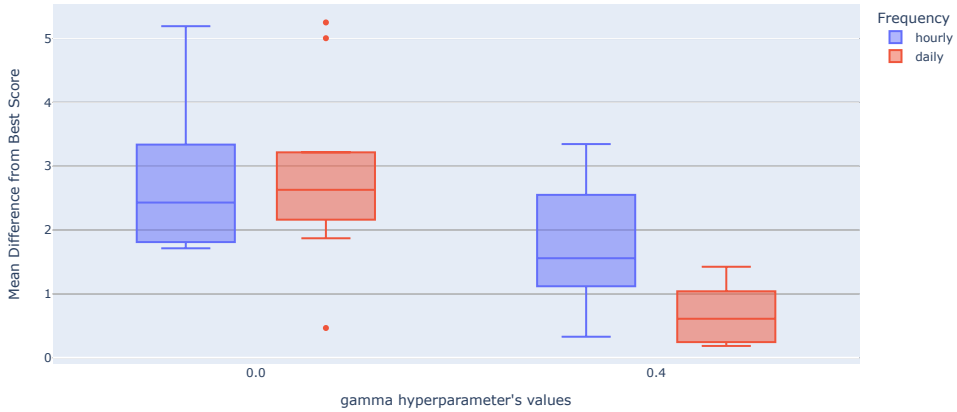


(c) RMSE percentage difference boxplot for minimum leaf samples

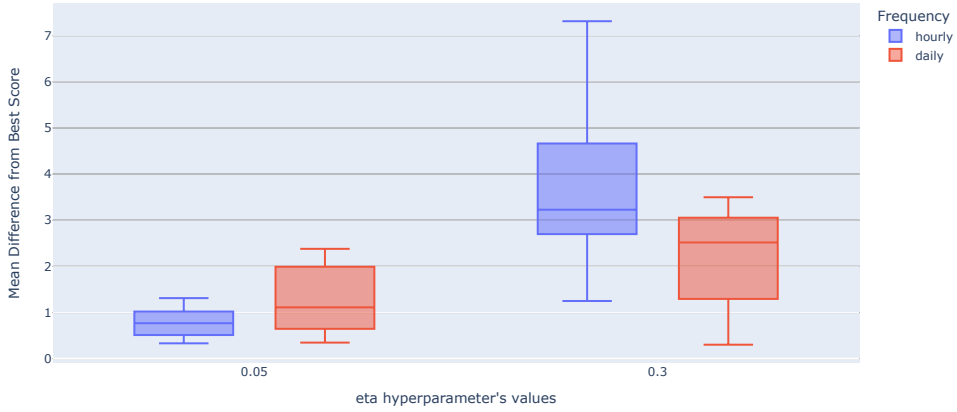
Figure 4.2: RMSE percentage difference boxplots for RF hyperparameter tuning analysis using rolling cross validation



(a) RMSE percentage difference boxplot for number of boosting rounds

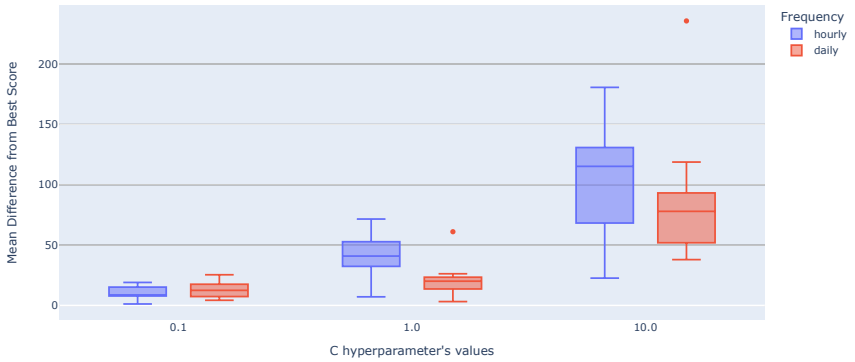


(b) RMSE percentage difference boxplot for the required split minimum loss

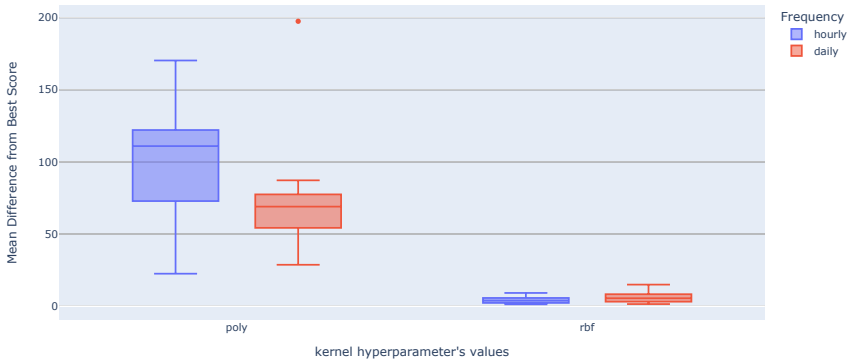


(c) RMSE percentage difference boxplot for boosting learning rate

Figure 4.3: RMSE percentage difference boxplots for XGB hyperparameter tuning analysis using rolling cross validation



(a) RMSE percentage difference boxplot for number of the regularization parameter



(b) RMSE percentage difference boxplot for the number of the SVR kernel

Figure 4.4: RMSE percentage difference boxplots for SVR hyperparameter tuning analysis using rolling cross validation

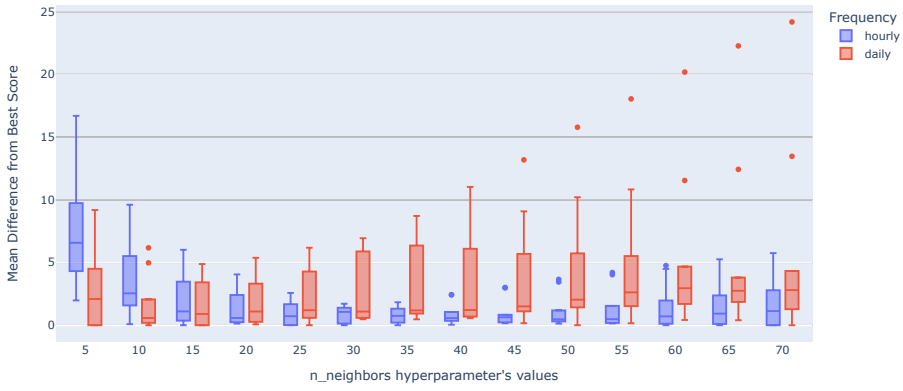
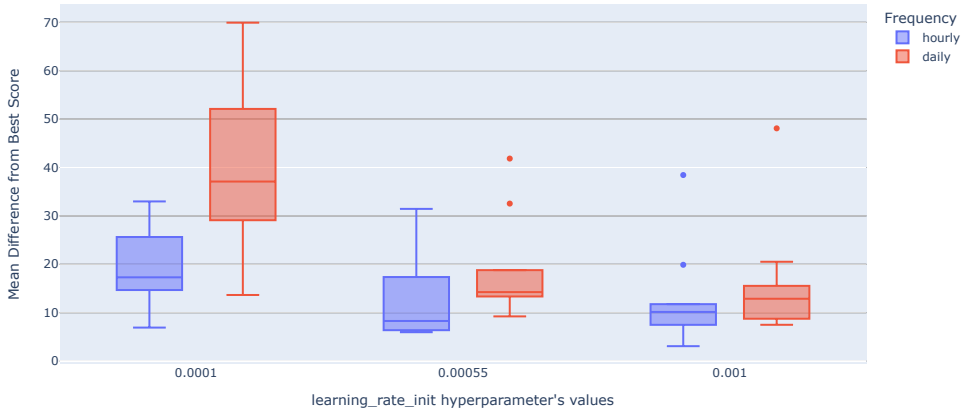
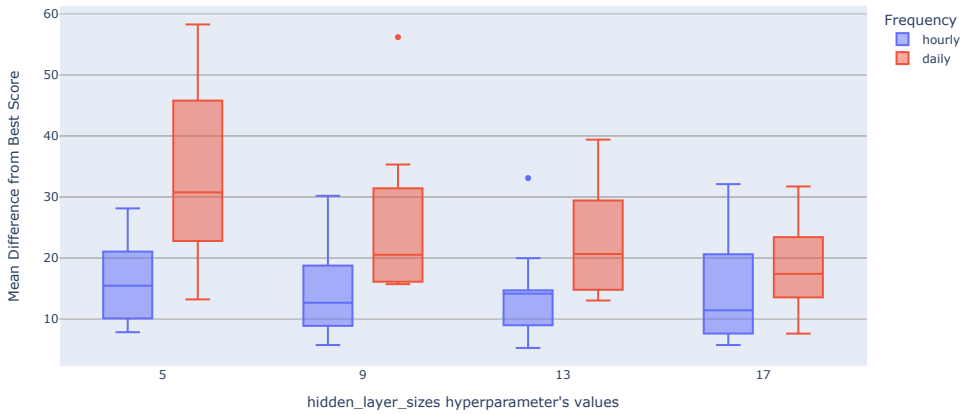


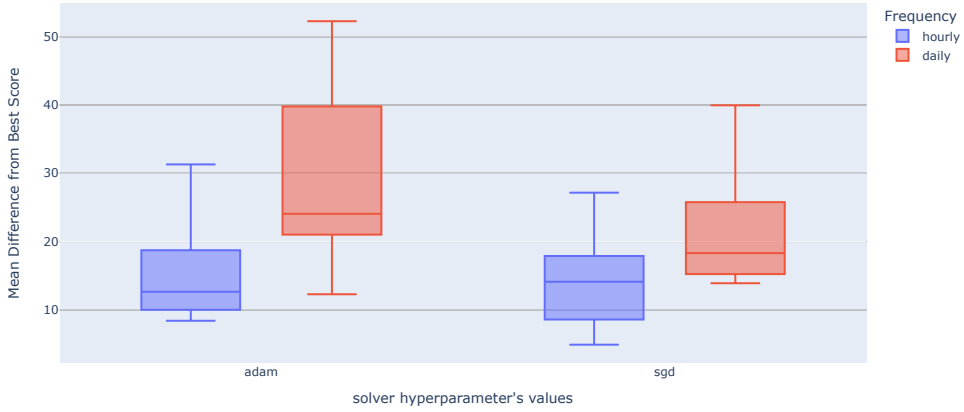
Figure 4.5: RMSE percentage difference boxplot for KNN number of neighbors tuning analysis using rolling cross validation



(a) RMSE percentage difference boxplot for number of initial learning rates

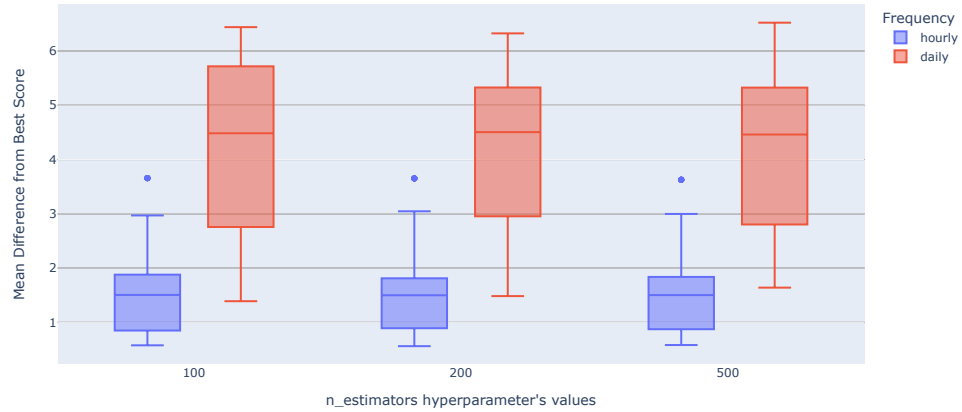


(b) RMSE percentage difference boxplot for the number of hidden neurons in the hidden layer

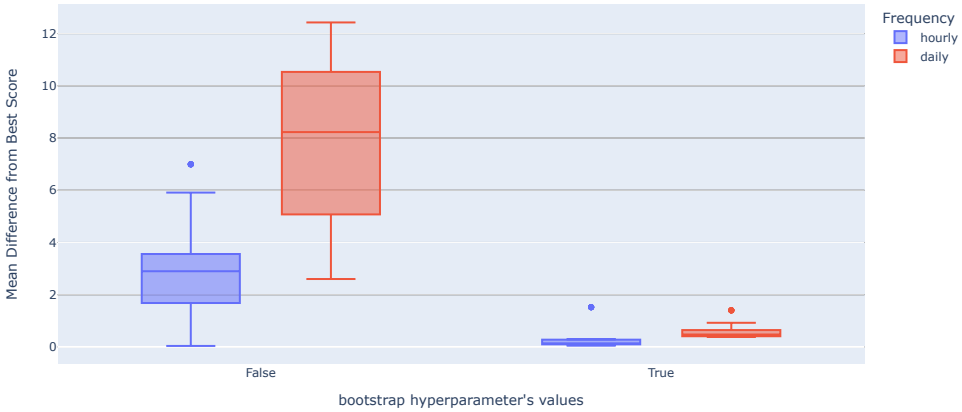


(c) RMSE percentage difference boxplot for optimization algorithm

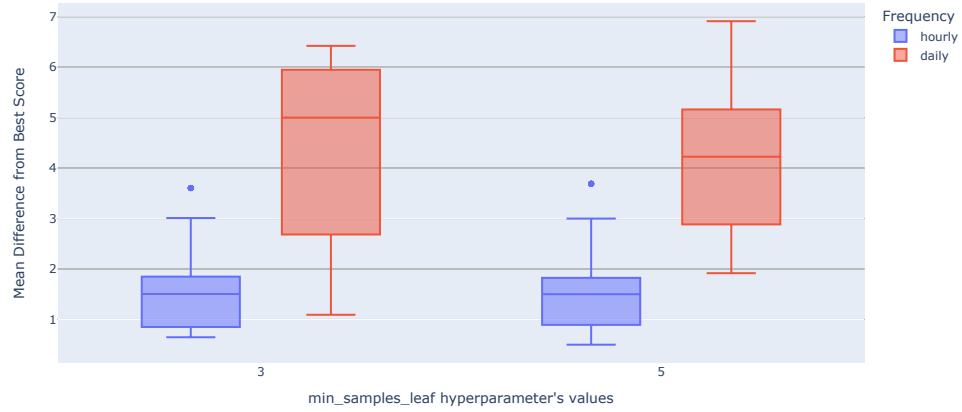
Figure 4.6: RMSE percentage difference boxplots for SLP hyperparameter tuning analysis using rolling cross validation



(a) RMSE percentage difference boxplot for number of trees

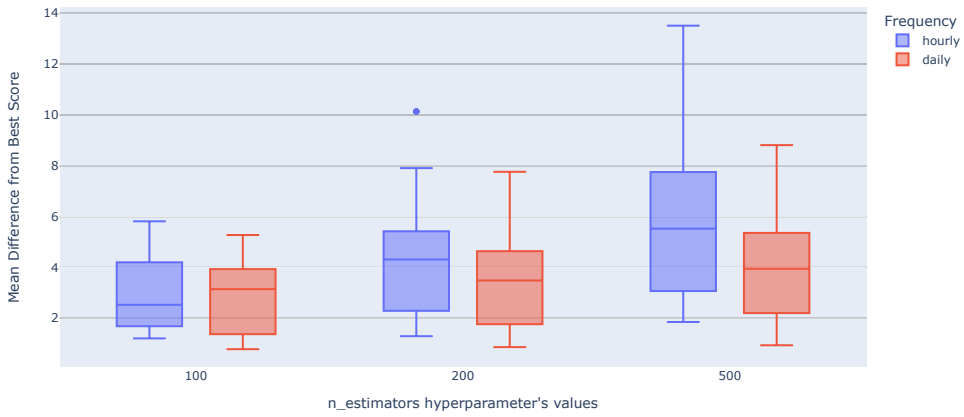


(b) RMSE percentage difference boxplot for including and excluding bootstrapping

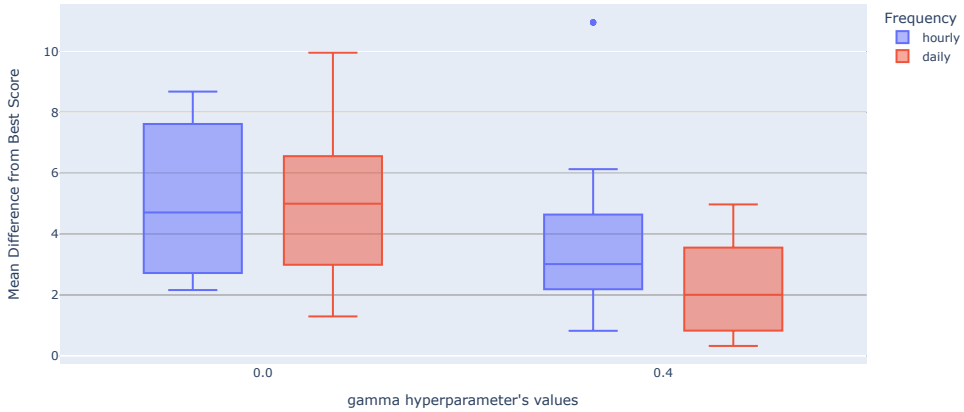


(c) RMSE percentage difference boxplot for minimum leaf samples

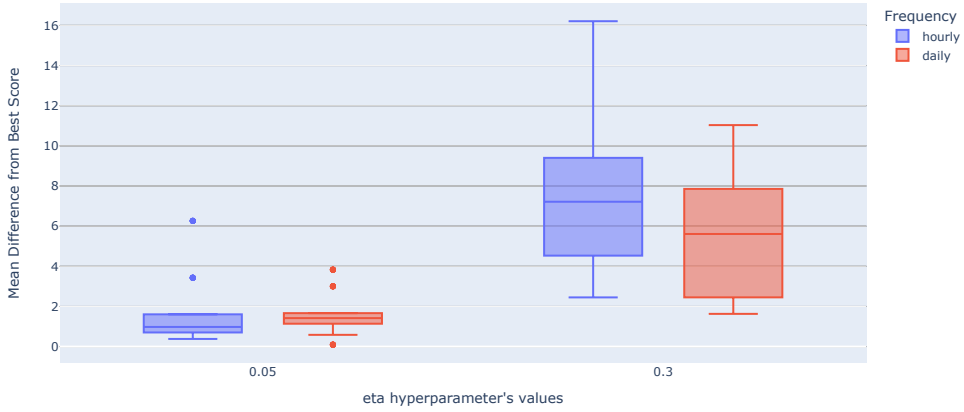
Figure 4.7: RMSE percentage difference boxplots for RF hyperparameter tuning analysis using K-fold cross validation



(a) RMSE percentage difference boxplot for number of boosting rounds

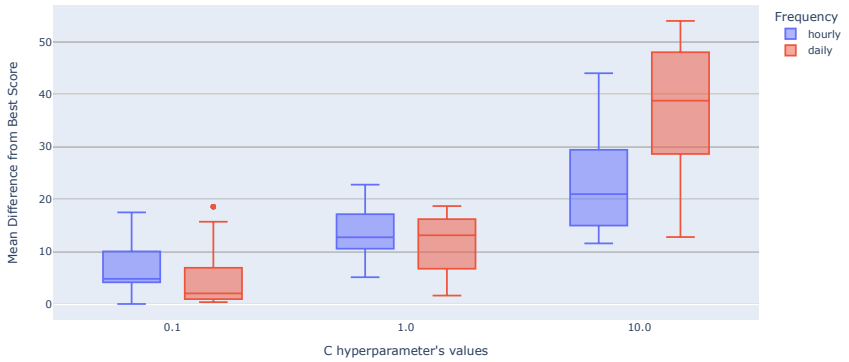


(b) RMSE percentage difference boxplot for the required split minimum loss

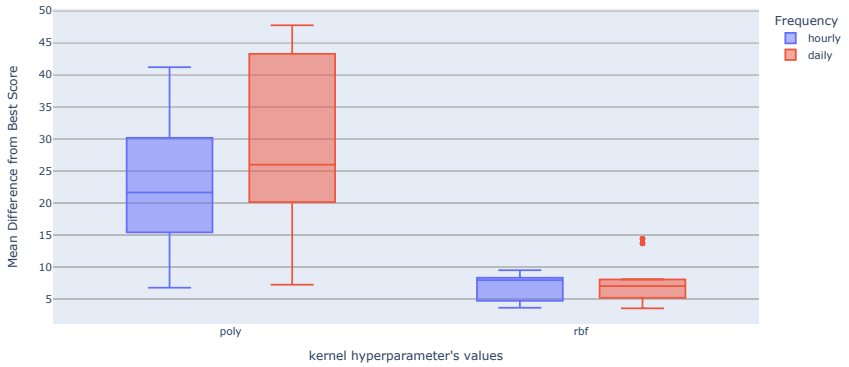


(c) RMSE percentage difference boxplot for boosting learning rate

Figure 4.8: RMSE percentage difference boxplots for XGB hyperparameter tuning analysis using K-fold cross validation



(a) RMSE percentage difference boxplot for number of the regularization parameter



(b) RMSE percentage difference boxplot for the number of the SVR kernel

Figure 4.9: RMSE percentage difference boxplots for SVR hyperparameter tuning analysis using K-fold cross validation

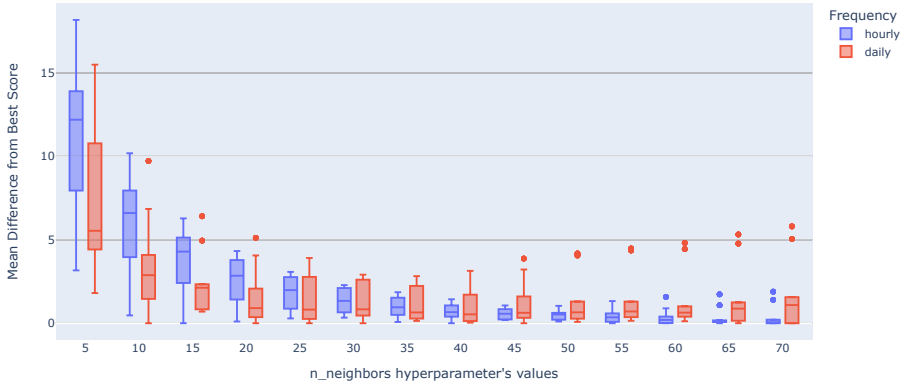
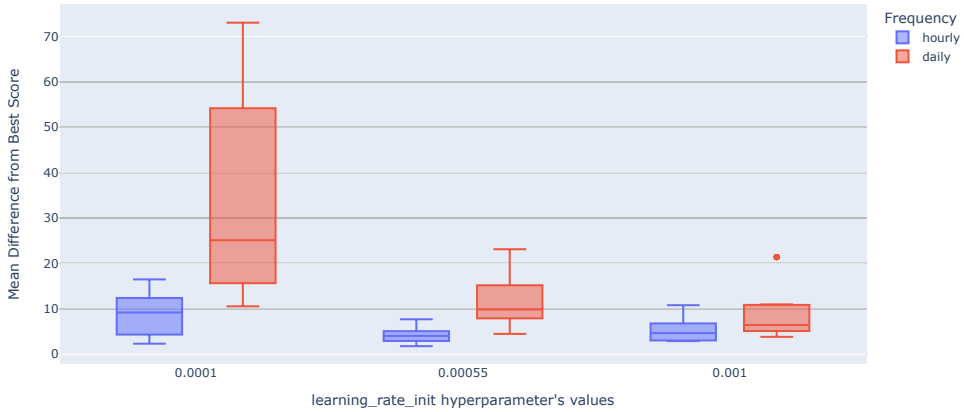
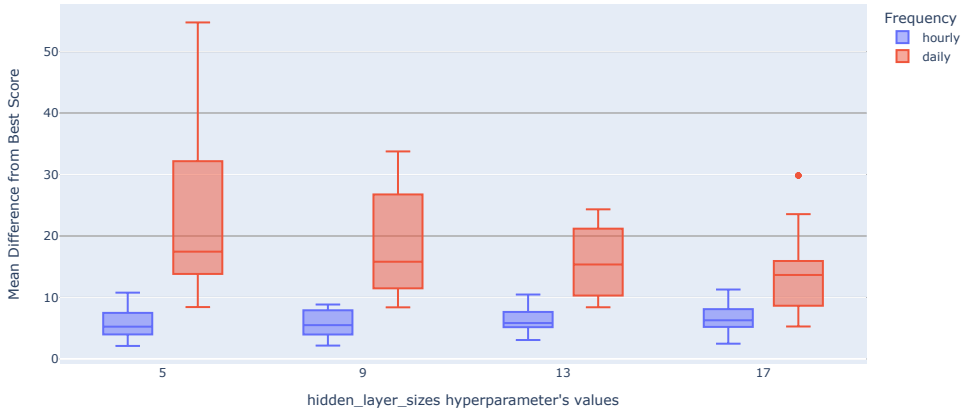


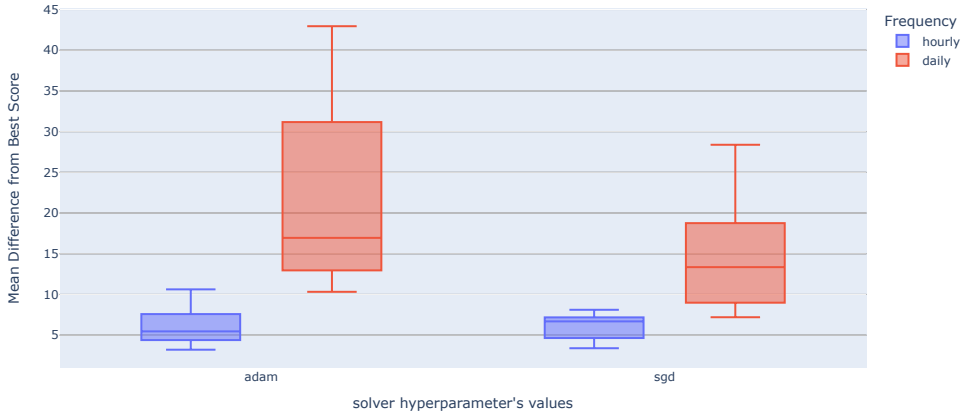
Figure 4.10: RMSE percentage difference boxplot for KNN number of neighbors tuning analysis using K-fold cross validation



(a) RMSE percentage difference boxplot for number of initial learning rates



(b) RMSE percentage difference boxplot for the number of hidden neurons in the hidden layer



(c) RMSE percentage difference boxplot for optimization algorithm

Figure 4.11: RMSE percentage difference boxplots for SLP hyperparameter tuning analysis using K-fold cross validation

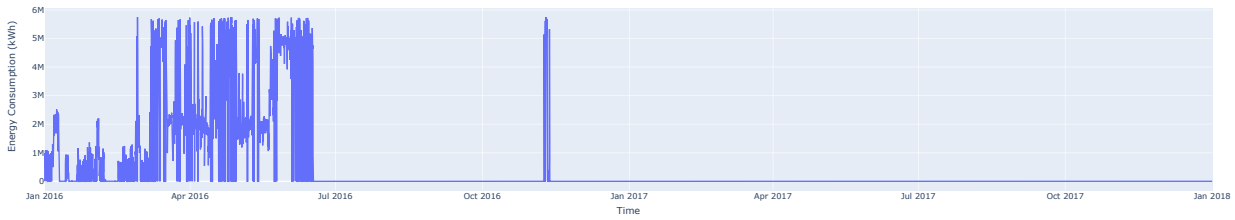
4.3 Genome Project 2: Multiple-cases Application

Section 3.3 has considered the application of the MVBEP tool using a synthetic dataset from a single building. However, energy demand from buildings can be unpredictable and highly affected by occupants' behavior and other factors. Testing the accuracy of any data-driven model using clean data set is not sufficient to assess the MVBEP tool's capabilities to generate accurate data-driven models. In this section, the MVBEP tool is applied to address datasets based on measurements and multiple buildings using the BDG2 database. The dataset includes 601 buildings that meet the initialization data sufficiency requirements discussed in Section 3.2.

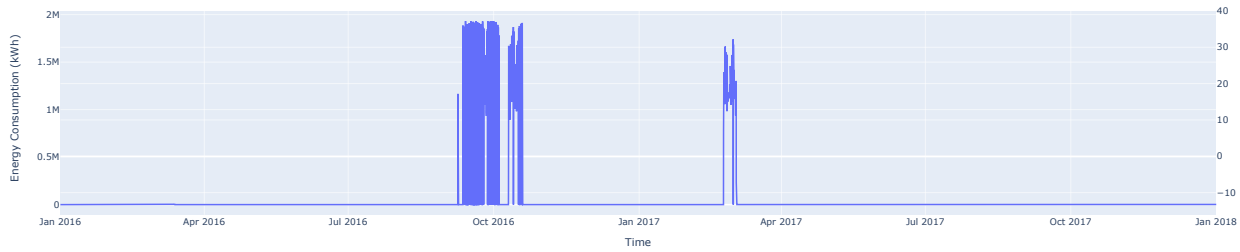
4.3.1 Model Development Failures

The MVBEP tool is applied to all BDG2 buildings that meet the data sufficiency requirements. No hyperparameter tuning of the data-driven models is performed for each building. Instead, the stated optimal sets of hyperparameters in Section 4.2 for various modeling approaches are utilized for all buildings. As explained earlier, the data sufficiency requirements limit potential failures but do not guarantee a successful development for data-driven models. Such failures have been observed for multiple buildings part of the BDG2 databases as datasets from some buildings provided models with CV(RMSE) score exceeding 46,500% for the testing phase as is the case for the datasets specific to the Hog education Robert and Peacock assembly Dena. The energy consumption plot shown in Figure 4.12 indicates that the data is not suitable for building a data-driven model.

Identifying potential failures for a large dataset is often cumbersome and difficult to achieve. For this study, datasets are filtered based on the CV(RMSE) scores obtained for the training periods. This approach mitigates the effects of buildings having inaccurate data while providing a general filtering approach that can be implemented for large datasets. According to the ASHRAE Guideline 14, a data-driven model during development must not exceed a CV(RMSE) score of 25% in the training set when 12 months of post-retrofit data is available. This requirement is only for daily and monthly datasets. With frequencies higher than daily, ASHRAE guideline 14 requires that they should be aggregated to at least a daily frequency. In this



(a) Hog education Rober building



(b) Peacock assembly Dena building

Figure 4.12: Examples of inaccurate datasets of energy consumption for two buildings

analysis, the 25% threshold is also considered for hourly datasets. Therefore, buildings with an average training set CV(RMSE) exceeding 25% for all models in an hourly or daily frequency are dropped from the analysis conducted in this section. The reason for considering averaged CV(RMSE) scores based on all models instead of a specific CV(RMSE) score for each model is to avoid overestimating the model's prediction accuracy. Datasets for certain buildings could deliver both successful and unsuccessful models based on the desired prediction accuracy threshold. If the average score exceeds the set threshold, the reason is most likely due to the fact that the data is not accurate and is not suitable for developing data-driven models that can be used for M&V analysis. However, if the average score does not exceed the threshold while some models may produce CV(RMSE) scores that exceed the set threshold, then the failure is mostly due to the model rather than the building's dataset. The resulting number of buildings after performing the filtration process is 208 buildings out of 601 as listed in Table 4.3.

4.3.2 Total Data Range Analysis

The MVBEP tool is used to develop data-driven models for all 208 buildings with both hourly and daily datasets. Figure 4.13 presents the testing accuracy metrics including CV(RMSE) and NMBE distributions

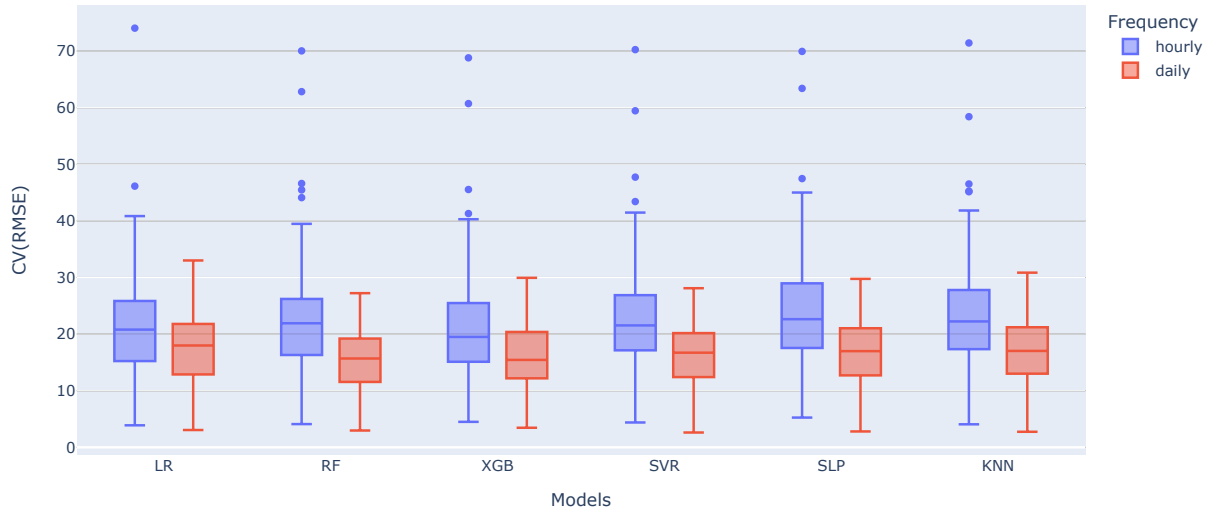
Table 4.3: Main features of used BDG2 data after filtering the datasets for all buildings

Site	Actual Site Name	Location	Climate	Buildings
Bear	Univ. of California - Berkeley	Berkeley, CA	3C	30
Eagle	Anonymous	N/A	4A	14
Fox	Arizona State Univ. (ASU)	Tempe, AZ	2B	43
Hog	Anonymous	Anonymous	6A	52
Panther	Univ. of Central Florida (UCF)	Orlando, FL	2A	7
Peacock	Princeton University	Princeton, NJ	5A	8
Rat	Washington DC City Buildings	Washington DC	4A	3
Robin	Univ. College London (UCL)	London, UK	4A	35
Swan	Anonymous	N/A	3C	7
Wolf	Univ. College Dublin (UCD)	Dublin, Ireland	5A	9

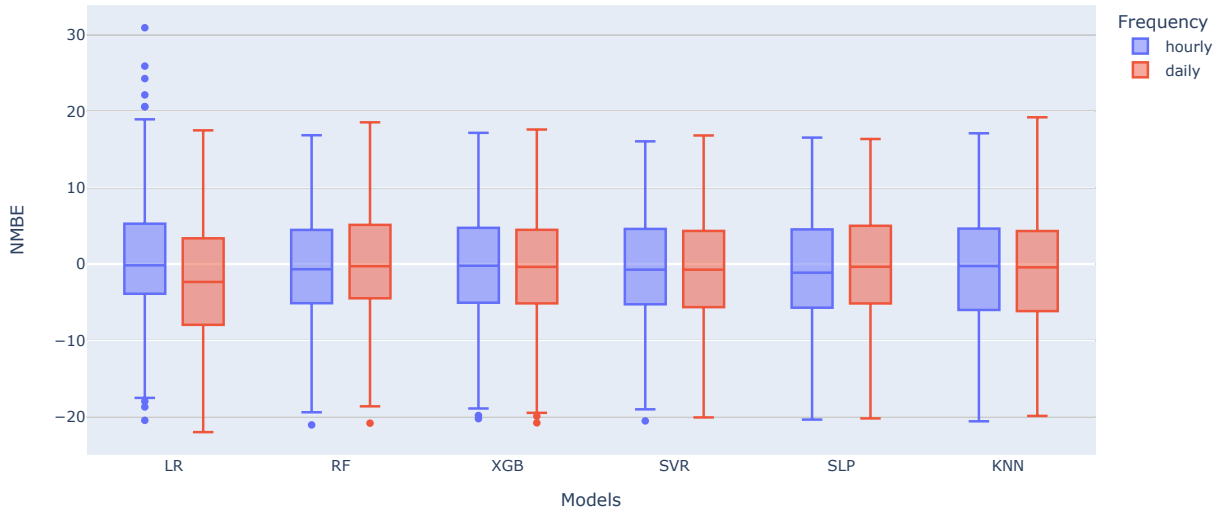
of all modeling approaches and buildings. For the 25th percentile (Q1), Interquartile Range (IQR), and 75th percentile (Q3) values for all models, the prediction accuracy metrics for both training or testing sets are shown in Figure 4.13 and listed in 4.4. The evaluation metrics in Table 4.4 are provided for both training or testing datasets to better assess if any model overfits the data by checking for a high prediction accuracy for the training dataset and low prediction accuracy for the testing dataset.

Table 4.4: Prediction accuracy metrics for both training and testing datasets for all hourly and daily modeling approaches

Evaluation			Frequency											
			Daily						Hourly					
Set	Metric	Q	KNN	LR	RF	SLP	SVR	XGB	KNN	LR	RF	SLP	SVR	XGB
Training	CV(RMSE)	1	11.20	11.73	7.86	7.53	7.97	1.20	13.09	13.89	13.75	7.09	9.41	6.41
		2	15.65	15.99	11.06	10.29	10.83	1.69	18.79	19.95	18.82	10.75	13.92	9.08
		3	19.09	21.01	14.03	13.26	14.46	2.55	24.07	25.41	23.63	14.80	18.73	12.20
	NMBE	1	-0.22	0.00	-0.04	-0.08	-0.78	0.00	-0.06	0.00	-0.01	-0.52	-0.59	0.00
		2	0.19	0.00	0.00	-0.01	-0.05	0.00	0.22	0.00	0.00	0.01	-0.15	0.00
		3	0.59	0.00	0.04	0.04	0.49	0.00	0.46	0.00	0.01	0.40	0.25	0.00
Testing	CV(RMSE)	1	13.06	12.90	11.82	12.98	12.65	12.25	17.47	15.78	16.67	17.94	17.41	15.23
		2	17.40	18.37	15.74	17.20	17.01	15.49	22.39	21.08	22.16	22.98	21.62	19.82
		3	21.49	21.81	19.21	21.05	20.16	20.16	27.82	25.91	26.36	29.05	26.88	25.64
	NMBE	1	-6.14	-7.96	-4.45	-5.15	-5.63	-5.13	-5.97	-3.88	-5.10	-5.71	-5.27	-5.00
		2	-0.43	-2.34	-0.29	-0.36	-0.74	-0.37	-0.26	-0.17	-0.68	-1.13	-0.74	-0.22
		3	4.28	3.35	5.10	4.98	4.34	4.42	4.58	5.25	4.35	4.50	4.54	4.64



(a) CV(RMSE) boxplots



(b) NMBE boxplots

Figure 4.13: Testing set accuracy metrics boxplots for all models using both hourly and daily frequencies

4.3.2.1 CV(RMSE) Based Performance

The results of Figure 4.13 indicate that all the modeling approaches have similar CV(RMSE) distributions for a given frequency setting. Few outliers are found for each model for the hourly frequency and outliers are detected for the daily frequency. Table 4.4 indicates that different models achieve differences between the training and testing set CV(RMSE) values. The CV(RMESE) values obtained for the testing datasets are higher as expected for data-driven models. For hourly frequency, SLP, SVR, and XGB-based models provide CV(RMSE) values for the training set that are half those obtained for the testing datasets which could indicate that these models are overfitting the training data. For daily frequency predictions, the differences in CV(RMSE) values between training and testing datasets are smaller for SLP and SVR-based models while are similar to those obtained for hourly frequency for RF, KNN, and LR-based models. However, XGB-based models produced the highest differences between training-testing CV(RMSE) values for daily frequency which is a strong indication that the models are overfitting the training dataset for a significant number of buildings. Although the CV(RMSE) value for testing sets are XGB-based models is comparable to those obtained by the other modeling approaches, overfitting is a potential indicator that the models may generate unreliable predictions. It is difficult to pinpoint the reasons for overfitting when considering performance metrics achieved for building datasets without examining the impacts of hyperparameter tuning. Nevertheless, XGB-based models provide the lowest testing set median CV(RMSE) values for both hourly and daily predictions. In terms of CV(RMSE) value distribution, RF-based models provide the narrowest distribution with values close to the median compared to other models.

4.3.2.2 NMBE Based Performance

Figure 4.13 shows that NMBE outliers are less compared to the CV(RMSE) outliers for most models except for LR-based models. Unlike the observations made using the CV(RMSE) metrics, daily frequency-based models achieve few outliers for RF and XGB-based models. Table 4.4 indicates the NMBE values for both the training and testing sets are similar. However, NMBE values are subject to error cancelations as no absolute or squared values are used to determine the differences between actual and predicted values as is the case for the CV(RMSE) values. All the NMBE median values are close to zero with distributions of most

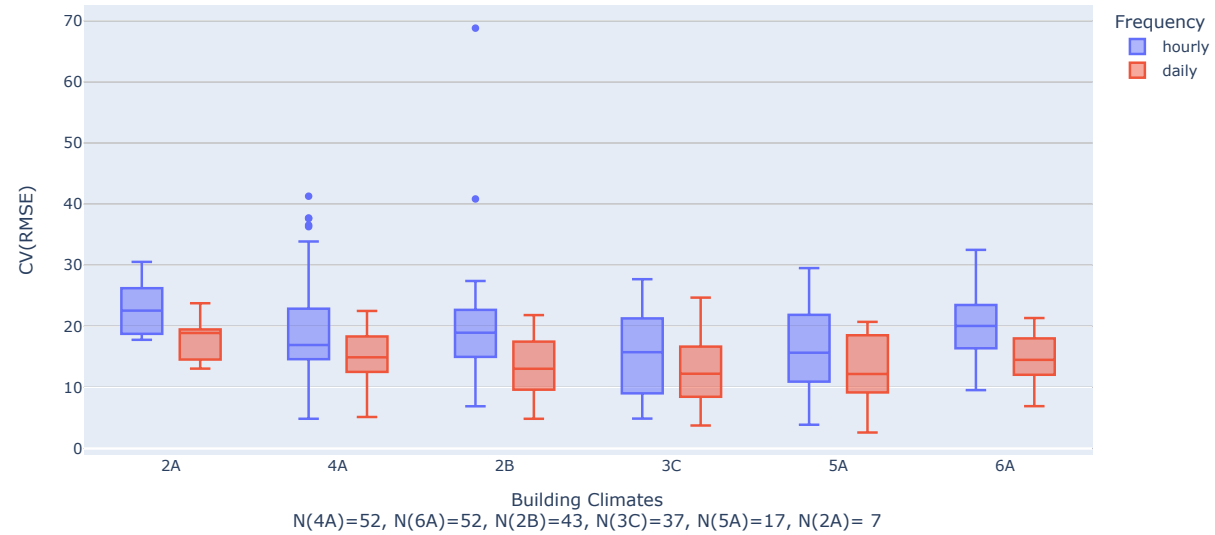
models almost symmetrical around zero. Such distribution patterns indicate that the modeling approaches, in general, do not have biases for under or overestimation of energy consumption of buildings. The NMBE distribution for the LR-based model trained using an hourly frequency varies slightly from that trained with a daily frequency most likely due to the differences in datasets associated with both frequencies as explained in Section 3.2. The NMBE distributions for the remaining modeling approaches do not exhibit significant variations in the frequencies.

4.3.3 Impact of Climate

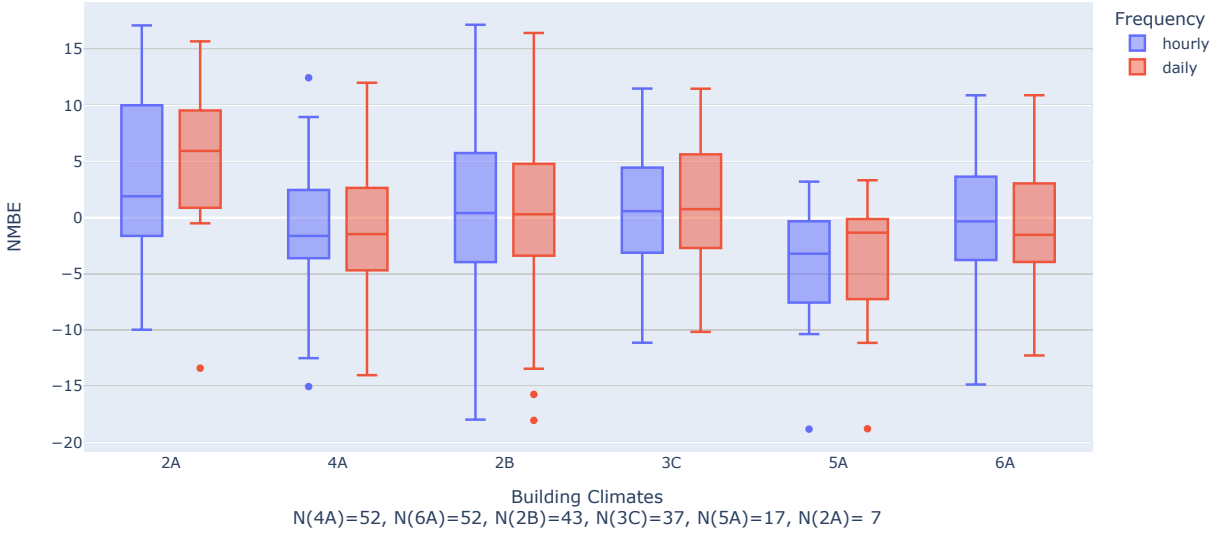
The effect of climate is analyzed by aggregating the performance of all resulting best hourly and daily models for each building and plotting distributions for both CV(RMSE) and NMBE values obtained for locations associated with specific ASHRAE climate zone and data frequency as shown in Figure 4.14. The classification of the climate zones is based on ASHRAE 90.1 (2019) [122]. Figure 4.14 does cover all AHRAE climate zones as the BDG2 dataset does not include buildings from all climate zones. The CV(RMSE) distributions are significantly different in terms of quantile values depending on the climate zone. The CV(RMSE) distributions have higher means and standard deviations for hourly frequency than those for daily frequency. In particular, the CV(RMSE) distributions specific to climate zones 4A and 2B for hourly frequency depict a significant number of outliers compared to the other climate zones. There is no clear indication that the performance of data-driven models is significantly affected by the climate zone.

4.3.4 Impact of Building Typology

Figure 4.15 shows the CV(RMSE) and NMBE distributions based on the dataset frequency and building type achieved by all the modeling approaches considered by the MVBEP tool. The 208 buildings dataset has 16 building types but only 5 building types have more than 10 buildings. For the analysis illustrated by Figure 4.15, building types with less than 10 buildings are not considered to assess the impact of topology on the performance of data-driven models. As indicated in Figure 4.15, the features of the CV(RMSE) distributions including median values and standard deviations do not change significantly with the building type. The CV(RMSE) distributions for daily frequency have lower medians and standard



(a) CV(RMSE) boxplots

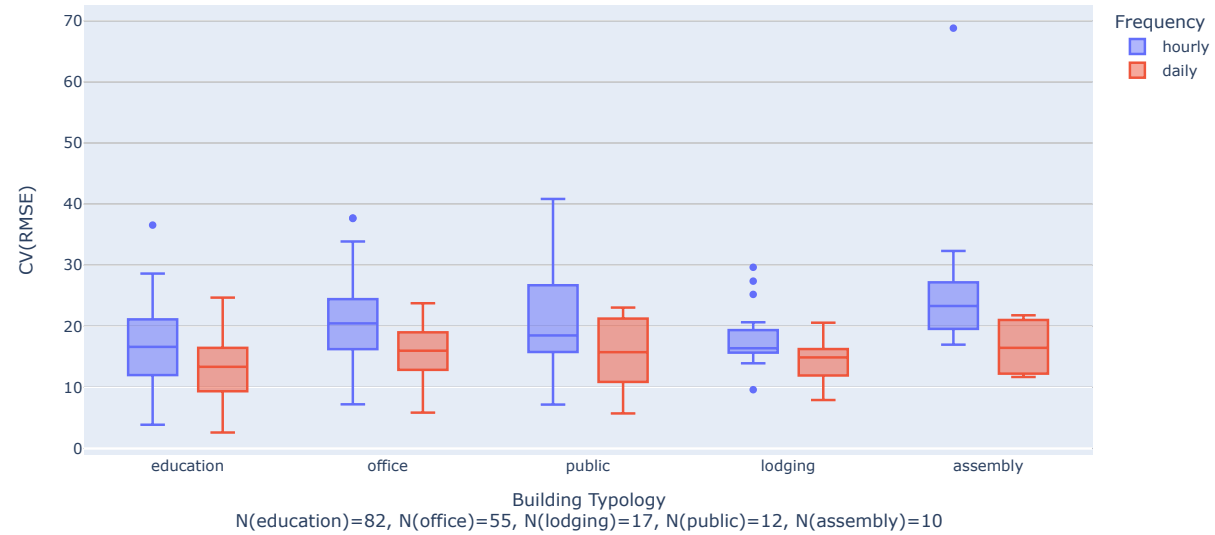


(b) CV(RMSE) boxplots

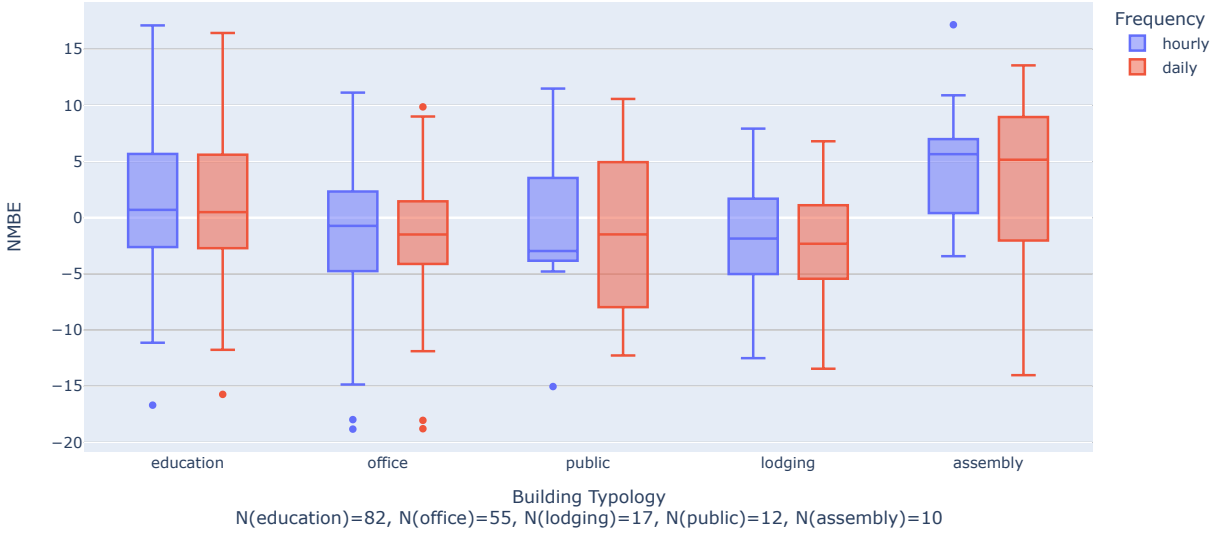
Figure 4.14: Testing set prediction accuracy metrics boxplots for various climates using hourly and daily frequency

deviations than those obtained for hourly frequency regardless of the building type. Public buildings have a relatively wide CV(RMSE) distribution (i.e., large standard deviation) possibly due to irregular schedules

which cannot be directly accounted for by the used features to develop data-driven models. Buildings with relatively high outliers are assembly buildings which suggest that data becomes more regular on a lower frequency. The NMBE distributions shown in Figure 4.15 exhibit similar median values and standard deviations regardless of the data frequency. Unlike the case for the CV(RMSE) distributions, the number of outliers for the NMBE distributions is not necessarily reduced when daily rather than hourly frequency is considered to develop data-driven models. All building types' NMBE distributions have median values close to zero indicating that the building typology does not affect the models' bias.



(a) CV(RMSE) boxplots



(b) NMBE boxplots

Figure 4.15: Testing set prediction accuracy metrics boxplots for various building topologies using hourly and daily frequency

4.3.5 Impact of Training Periods

In this section, the impacts of both time range and period selection for training the performance of data-driven models are investigated. Generally, the larger the dataset used for training models, the better the prediction accuracy. However, for M&V applications, complete and high-quality datasets useful for training models are often limited. Although M&V protocols and guidelines set some thresholds, the impact of the amount of data to train various MVBEP tool models is considered in this section. For the analysis, a subset of the 208 buildings dataset is considered to select only buildings with 2-year data for total energy consumption. For the initial analysis, the first year's data is used for training while the second year's data is used for testing the MVBEP tool models. Then, the training period is gradually reduced from 12 months to just 3 months with an increment of 3 months. For each case, the remaining first-year data is not used for testing the performance of the data-driven models. Thus, regardless of how many months a model was trained on for the first year, the testing is performed by using only the whole second year. Table 4.5 lists the main characteristics of the 194 buildings with at least 2 years of energy consumption data.

Table 4.5: Main characteristics of the BDG2 dataset with 2-year building energy consumption

Site	Actual Site Name	Location	Climate	Buildings
Bear	Univ. of California - Berkeley	Berkeley, CA	3C	30
Eagle	Anonymous	N/A	4A	14
Fox	Arizona State Univ. (ASU)	Tempe, AZ	2B	42
Hog	Anonymous	Anonymous	6A	51
Panther	Univ. of Central Florida (UCF)	Orlando, FL	2A	2
Peacock	Princeton University	Princeton, NJ	5A	8
Rat	Washington DC City Buildings	Washington DC	4A	3
Robin	Univ. College London (UCL)	London, UK	4A	35
Wolf	Univ. College Dublin (UCD)	Dublin, Ireland	5A	9

As noted earlier, training periods of 12, 9, 6, and 3 months are considered for training each of the MVBEP tool's data-driven models. For simplicity, the training periods are segmented using annual quarters so different 3-month periods can be considered to constitute a training period. For instance, in a 9-month training period, different continuous quarters could be used from the first-year dataset including from the first to the third quarter or from the second to the fourth. Similarly, a 6-month training period can include either the first or last two quarters. Lastly, a 3-month training period can have four variations with each

quarter of the first-year dataset. These variations aim to test the effects of selecting reduced training periods as well as the specific seasons that these periods cover on the performance of data-driven models. There are 9 training period options for each of 194 buildings and 2 frequencies with each having 6 modeling approaches (e.g. LR and XGB) resulting in 20,952 different data-driven models.

First, the CV(RMSE) and NMBE distributions are evaluated for various training periods, model approaches, and data frequencies as shown in Figure 4.16. The exact CV(RMSE) quartile values are listed in Table 4.6 for various training periods, model approaches, and data frequencies. The reductions in both the median values and standard deviations for CV(RMSE) when changing from hourly to daily frequencies are observed for almost all models regardless of the training period. The differences between the CV(RMSE) distributions for a given data frequency between 12-month and 9-month training periods are small for all the models as indicated by Table 4.6. The CV(RMSE) distributions are also relatively similar between 9-month and 6-month training periods for all models except SLP-based models (both frequencies) and LR-based models (for daily frequency) where the testing CV(RMSE) distributions have higher median values and standard deviations. Finally, when the training period is limited to 3 months, there is a significant deterioration of the performance of all the models with their CV(RMSE) distributions having more outliers and third quantiles exceeding 40%.

For the NMBE distributions depicted in Figure 4.17, the impacts of changing the training period are different from those noted for the CV(RMSE) distributions. Indeed, Figure 4.17 indicates that the NMBE distributions show small differences when training from 12-month to 9-month periods regardless of the data frequency. However, the differences between 9-month to 6-month training periods are relatively high for SLP-based models with hourly frequency. Some models show quantile values that are closer to zero when trained with shorter periods but the distributions have larger standard deviations. For 3-month training periods, the NMBE distributions exhibit similar quantile values to those obtained for longer training periods using the same frequency and the same modeling approaches. However, using 3-month training periods generates distributions with a high number of outliers.

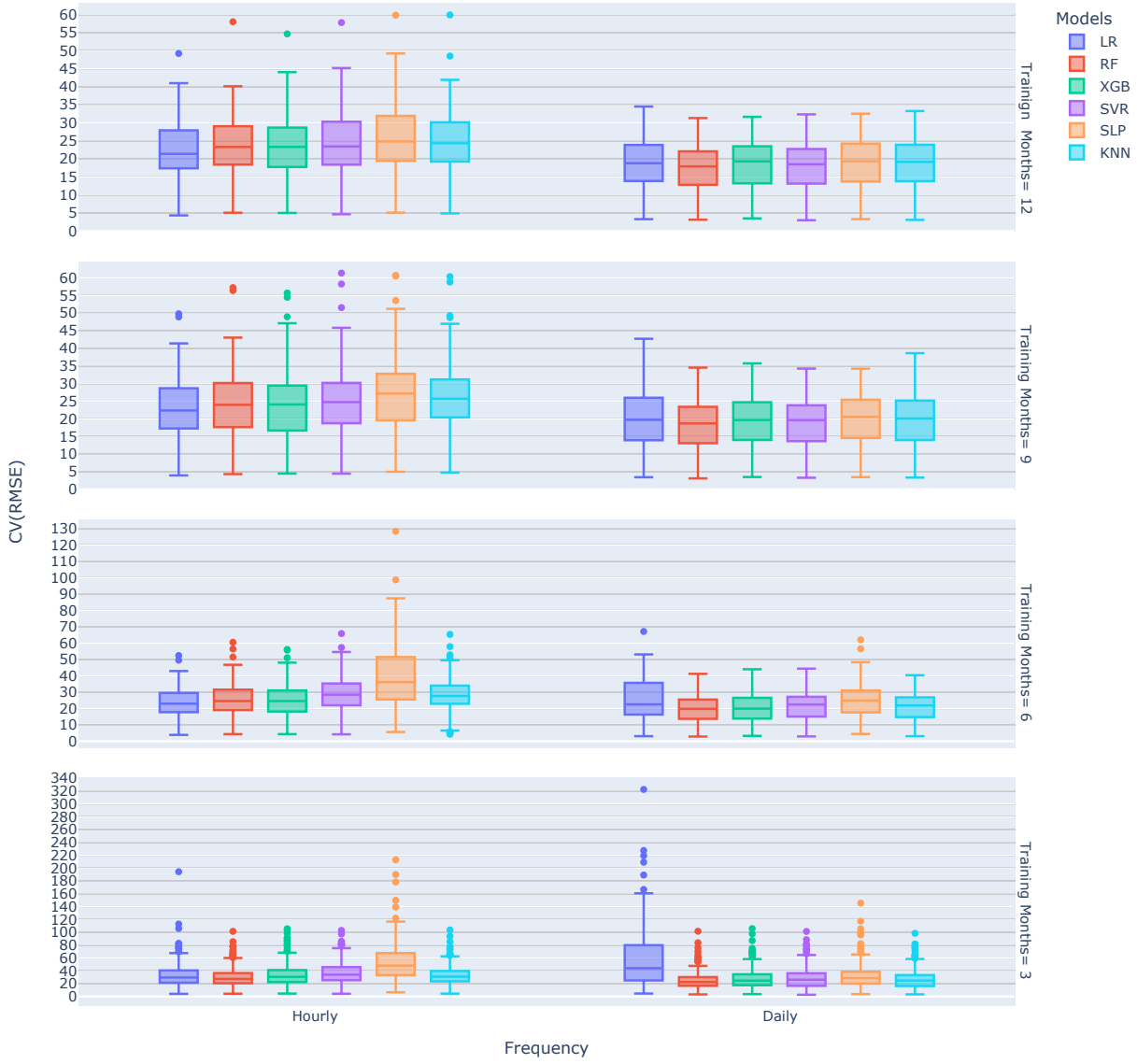


Figure 4.16: CV(RMSE) metric boxplots for various modeling approaches, training periods, and frequencies

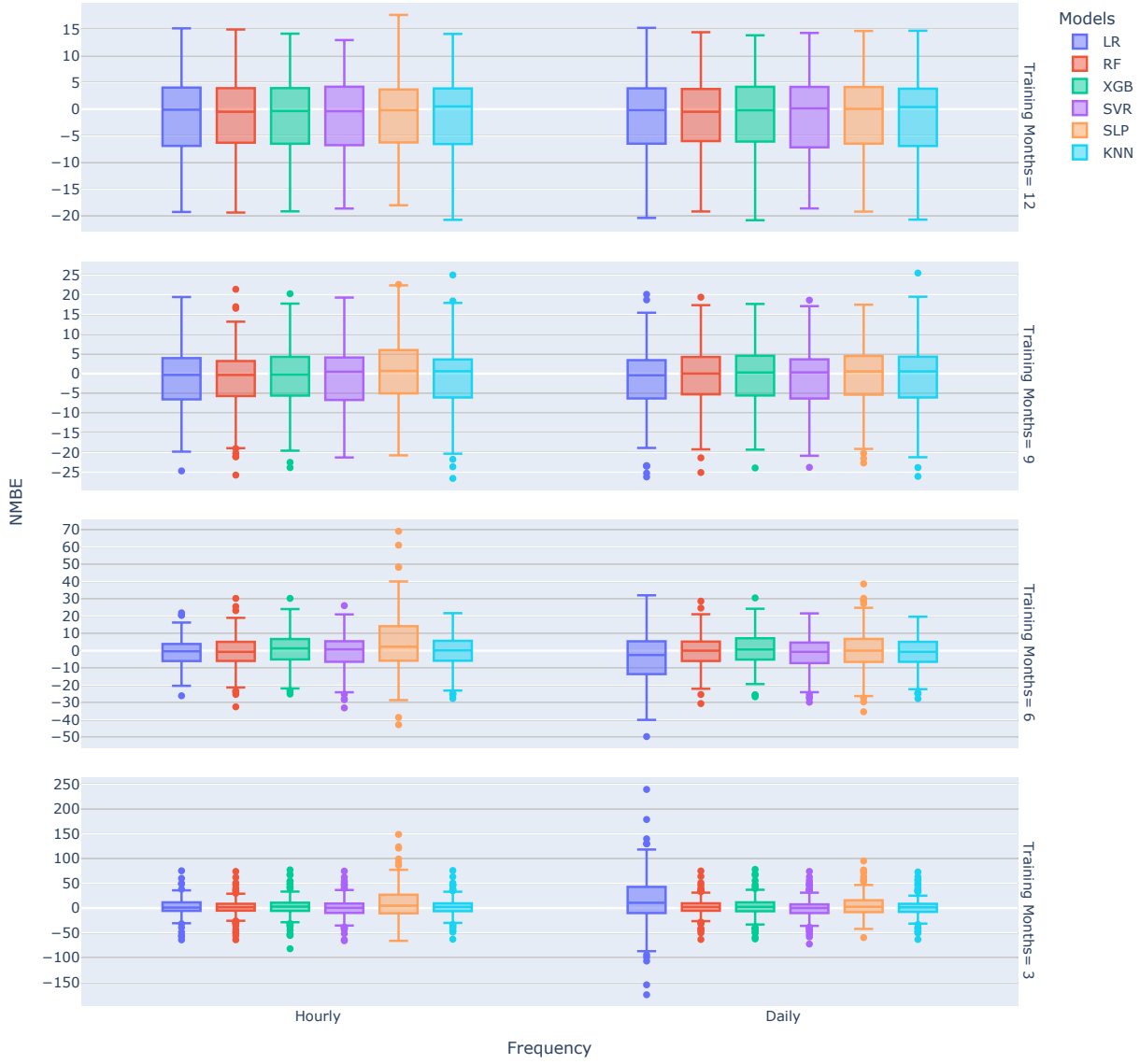


Figure 4.17: NMBE metric boxplots for various modeling approaches, training periods, and frequencies

Table 4.6: CV(RMSE) quartile values for various training periods, modeling approaches, and frequencies

Frequency	Models	Q1 Training Months				Q2 Training Months				Q3 Training Months			
		3	6	9	12	3	6	9	12	3	6	9	12
Daily	KNN	16.4	14.9	14.0	13.9	24.8	21.9	20.1	19.2	33.4	26.8	25.1	23.8
	LR	25.0	16.4	14.0	13.9	44.1	22.6	19.7	18.8	79.9	35.7	26.0	23.7
	RF	16.7	13.8	13.1	12.8	22.7	19.8	18.7	17.9	30.2	25.5	23.4	22.1
	SLP	20.0	17.7	14.6	13.8	28.6	25.0	20.6	19.4	38.4	30.9	25.4	24.0
	SVR	16.7	15.2	13.8	13.3	26.2	22.5	19.6	18.5	36.3	27.2	23.8	22.7
	XGB	17.7	13.9	14.0	13.3	24.7	20.0	19.7	19.3	34.6	26.6	24.7	23.4
Hourly	KNN	24.0	23.3	20.7	19.4	30.9	27.7	25.7	24.3	39.7	33.9	31.0	29.8
	LR	21.5	18.1	17.3	17.5	29.6	23.0	22.4	21.4	40.6	29.5	28.7	27.9
	RF	20.7	19.1	17.8	18.8	27.4	24.6	24.0	23.3	36.4	31.4	30.1	29.0
	SLP	33.2	25.6	19.6	19.7	48.1	36.2	27.2	24.9	67.6	51.2	32.8	31.7
	SVR	25.4	22.2	18.8	18.4	34.0	28.5	24.8	23.5	45.7	35.3	30.2	30.1
	XGB	22.5	18.2	16.8	17.9	30.6	24.6	24.1	23.3	41.1	30.8	29.4	28.5

Table 4.7: NMBE quartile values for various training periods, modeling approaches, and frequencies

Frequency	Models	Q1 Training Months				Q2 Training Months				Q3 Training Months			
		3	6	9	12	3	6	9	12	3	6	9	12
Daily	KNN	-7.6	-6.4	-6.0	-6.5	1.4	-0.6	0.6	0.4	8.5	5.1	4.2	3.7
	LR	-10.0	-13.2	-6.3	-6.1	10.6	-2.5	-0.4	-0.2	42.5	5.5	3.4	3.8
	RF	-5.4	-5.8	-5.2	-5.8	2.1	0.0	0.0	-0.5	9.6	5.3	4.2	3.6
	SLP	-8.1	-6.2	-5.3	-6.2	2.5	0.0	0.6	0.1	15.5	6.8	4.5	4.1
	SVR	-10.0	-7.0	-6.3	-7.1	0.1	-0.6	0.3	0.1	7.4	4.5	3.6	3.9
	XGB	-6.2	-5.1	-5.4	-5.9	2.4	0.7	0.3	-0.2	11.5	7.1	4.5	4.1
Hourly	KNN	-6.3	-5.8	-6.0	-6.2	2.1	0.2	0.6	0.5	9.5	5.7	3.6	3.7
	LR	-5.6	-5.9	-6.5	-6.7	1.0	-0.3	-0.3	-0.1	11.5	3.7	3.9	3.9
	RF	-5.2	-5.9	-5.7	-6.2	1.6	-0.6	-0.3	-0.5	8.6	5.1	3.1	3.9
	SLP	-10.6	-5.8	-5.0	-6.0	4.8	2.3	0.7	-0.2	26.8	14.2	5.9	3.6
	SVR	-9.8	-6.4	-6.6	-6.6	0.4	0.8	0.5	-0.4	9.0	5.4	4.1	4.0
	XGB	-5.6	-5.0	-5.6	-6.3	2.9	1.4	-0.2	-0.3	10.7	6.7	4.3	3.9

Figure 4.18 shows the aggregated CV(RMSE) distributions for all modeling approaches, ASHRAE climate zones, and training periods expressed in yearly quarters. The boxplot whiskers are extended to the maximum value instead of only 1.5 of the interquartile range (IQR) beyond the third quantile as the results showed a high number of outliers. Using these boxplots, more insights can be gleaned from the CV(RMSE) distributions. For 9-month training periods, the CV(RMSE) distributions do not show any significant variations when using the first three quarters instead of the last three. There are, however, differences in CV(RMSE) distributions when comparing different ASHRAE climate zones even though no

specific conclusions can be made about the impacts of climates on the performance of data-driven models as established in Section 4.3.3. For 6-month training periods, the selection of the first or last 2 quarters does not show a significant effect on the CV(RMSE) distributions. When reducing the training periods to only one quarter, changes in CV(RMSE) distributions become more evident for some climate zones and quarters. Climate 5A, characterized by cold and humid conditions, shows that selecting the second or third quarter generates the narrowest CV(RMSE) distributions with median values being close for all quarters. This result is counter-intuitive as the second and third quarters do not cover the coldest seasons. In contrast, in climate 6A, also featuring cold and humid conditions, the CV(RMSE) distributions for the first and the last quarters exhibit the smallest IQRs and the closest to zero median values. For climate 4A with mild and humid conditions, the second and third quarters' CV(RMSE) distributions are narrower suggesting data obtained for quarters covering spring or fall are more important to train data-driven models for mild climates. Climate 2B, featuring hot and dry conditions, CV(RMSE) distributions for the second and third quarters provide the best performance. However, the analysis results conclude that the selection of specific periods for given climates has no significant impact on the prediction accuracies of data-driven models.

Figure 4.19 shows the NMBE distributions for various training periods and ASHRAE zone climates. Figure 4.19 shows different results from those obtained from Figure 4.18 specific to the CV(MSE) distributions. Some of the distributions with 9-month training periods for a given climate zone vary with the selected quarters. Similar differences are observed using the same climates when reducing the training to 6-month periods. Based on the results of Figure 4.19, there is no clear indication that the climate has any significant effect on the NMBE distributions. Furthermore, climate 5A and 6A are very similar yet they produce different results for the NMBE distributions which suggests that such variations are attributed to either noisy datasets or other factors related to the building types available for each of the climate zones considered in the analysis.

The analysis results shown in Figure 4.18 and 4.19 do not demonstrate a clear impact that can be associated with choosing a specific training period for a certain climate. The observed significant impacts are related only to the training period and dataset frequency. The absence of a consistent climate effect might be attributed to either the data or the process of training. The former is an effect of having buildings

without any controlled characteristics such as type or activity between different climates which could be why the results were contradicting between similar climates. On the other hand, the process of training with insufficient data might have contributed to making the accuracy metric distributions noisy. Granderson et al. [35] performed a similar analysis with 209, 196, and 30 buildings from California, Washington D.C., and Seattle, respectively. The study used datasets with daily frequency and for the same mentioned training periods (i.e. 12, 9, 6, and 3 months) but without specifying the training data coverage over the year. The study showed that the CV(RMSE) distributions of testing set between buildings in California are more similar to Seattle than those in Washington D.C. To investigate such effects, more comprehensive datasets from different climates with randomly selected building typologies need to be considered.

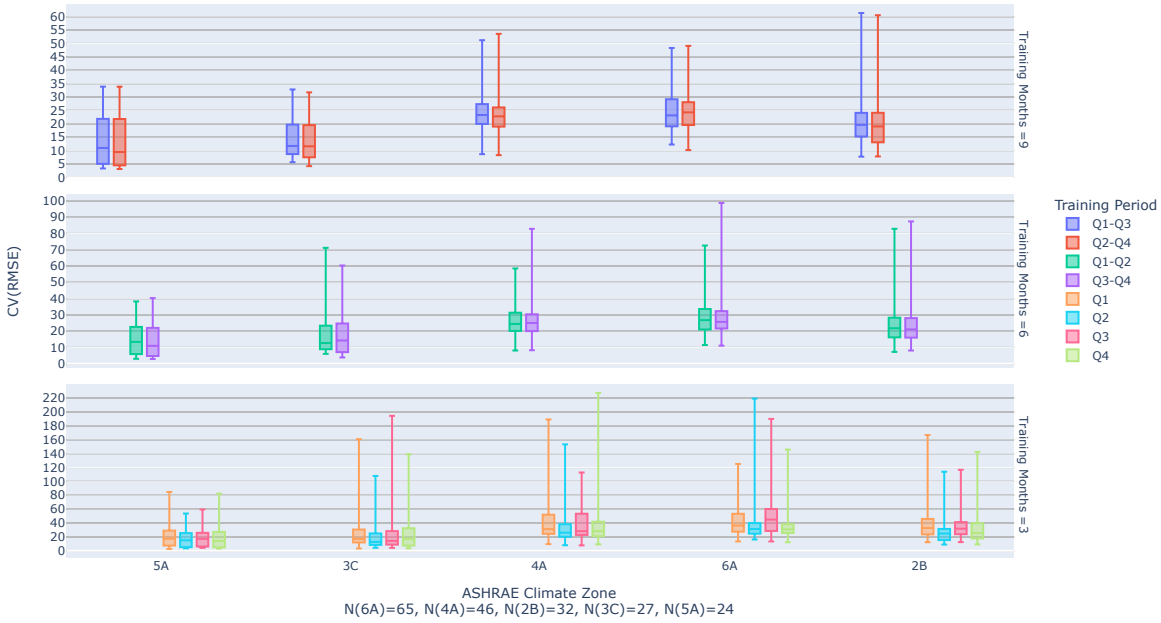


Figure 4.18: CV(RMSE) metric boxplots for various training periods expressed in quarters, and ASHRAE climate zones

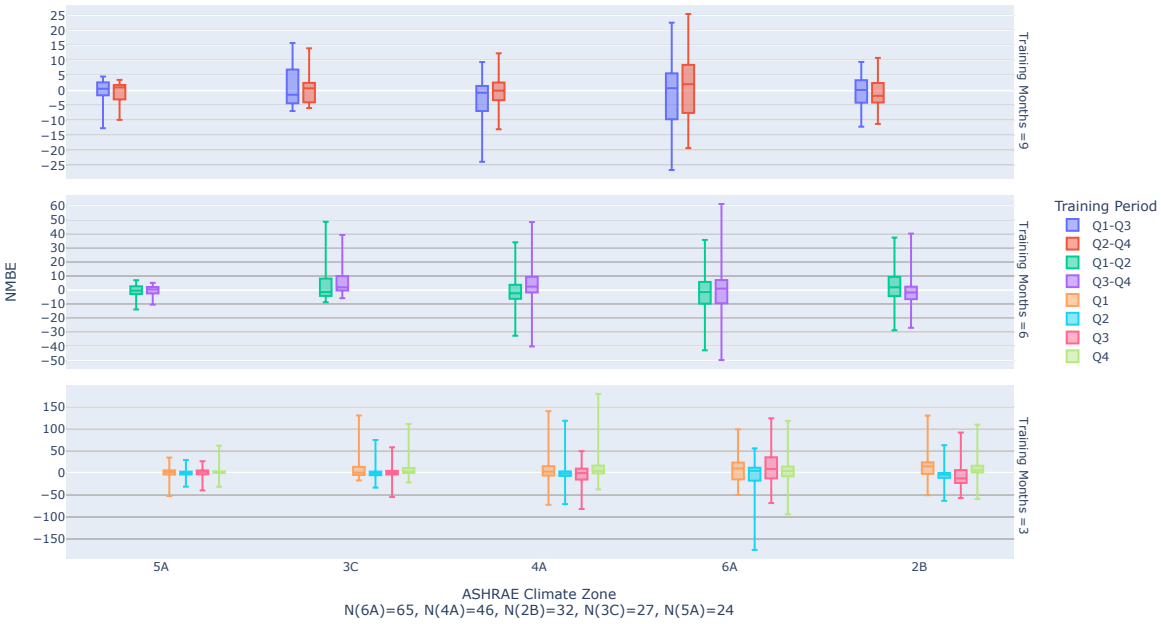


Figure 4.19: NMBE metric boxplots for various training periods expressed in quarters, and ASHRAE climate zones

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

To justify retrofitting buildings, M&V analysis is often needed to quantify the achieved energy savings and ultimately assess the cost-effectiveness of implemented energy efficiency measures. Data-driven modeling provides an effective approach to performing M&V analyses when compared to traditional deterministic modeling methods especially considering the growing availability of measured historical data specific to energy performance due to advancements in metering and monitoring of building energy systems.

This thesis presented a review of existing knowledge specific to M&V data-driven baseline modeling approaches, feature engineering, and current M&V analysis frameworks. It is clear from the presented review analysis that there is a need for the development of a general analysis tool to generate advanced data-driven models suitable for M&V analysis and capable of accurately estimating energy savings achieved by building retrofits. While the review has revealed some existing analysis tools, most of them are based mostly on simplified modeling LR-based methods. Several reported studies indicated that various modeling approaches deliver varying prediction accuracy levels with no clear guidelines for determining the best modeling approach for given M&V analysis project. Moreover, results from reported studies that developed data-driven modeling for M&V analysis have been tested only for specific case studies. Thus, their applications cannot be readily generalized to any building type, climate zone, and retrofit application.

As result, a new MVBEP tool is proposed to enhance the use of data-driven modeling for M&V applications to quantify achieved energy savings from retrofit projects. The MVBEP tool is segmented systematic tasks executed by specific modules including initialization, transformation, development, interpretation, and quantification. Each module is developed using methods that have been proven to be effective for train-

ing and testing data-driven baselining models through reported studies. Furthermore, the thesis performed several applications of the proposed MVBEP tool to explain its methodologies, tuning capabilities, development components, and testing results. The testing of the MVBEP tool has been carried out using datasets obtained from two data sources: synthetic dataset from simulation analysis of an office building in Boulder, CO and the BDG2 dataset.

5.1 Findings Summary

The developed MVBEP tool is first demonstrated using dataset for a single simulated office building in Boulder, CO to evaluate the performance of the data-driven models using clean dataset. The dataset from the office building demonstrated that MVBEP tool was able to develop models having low CV(RMSE) scores of at least 9% and NMBE values as low as 2.25%. The estimated annual energy savings or avoided energy use (AEU) estimated by the developed data-driven models using the MVBEP tool is found to be within 5.54% from the actual AEU. The relation of CV(RMSE) and NMBE with the percentage difference between the estimated and true AEU was investigated and an apparent correlation was observed between attaining a low percentage difference and obtaining better evaluation metrics. The better the obtained value for both CV(RMSE) and NMBE the closer the estimated AEU from the true AEU. To highlight the general performance of the develop MVBEP tool, a series of sensitivity analyses is carried out as summarized in the following sections.

5.1.1 Impact of Hyperparameter Tuning

The effects of hyperparameter tuning were assessed by analyzing the standard deviations in every performed grid search using dataset for a 10-building sample. The RMSE boxplots of grid search's model combinations with different hyperparameter values showed significant variations with some reaching a standard deviation of 50. Hence, the hyperparameter tuning is an important process to improve the prediction accuracy of the data-driven models considered by the MVBEP tool. To determine the optimal set of hyperparameters to set as default values for expedited model development within the MVBEP tool, the average difference from the best RMSE scores for specific hyperparameter values were obtained in all 10 buildings

considered in the sample used in the analysis. Boxplots were used to visualize the distributions of selecting values for specific hyperparameters utilized by each data-driven model. Table 5.1 shows a summary of the recommended values for the hyperparameters for various MVBEP models.

5.1.2 Impact of Modeling Approaches

The application of the MVBEP tool using datasets for multiple buildings encompassing different topologies and climate zones was performed using a set of optimal hyperparameters for various data-driven models. Firstly, only 601 out of 1,500 buildings met the data quality requirements defined in Section 3.2. For the sensitivity analyses, buildings with dataset having an average training set CV(RMSE) score more than 25% were not considered as detailed in Section 4.3. The final dataset used in the sensitivity analyses has 208 buildings with both hourly and daily frequencies. The analysis showed that both CV(RMSE) and NMBE scores varied with the modeling approach and data frequency. XGB-based models with daily frequency resulted in the lowest quantile values for CV(RMSE) distributions. The analysis of both training and testing prediction accuracy metrics demonstrated that complex models are more prone to overfitting when training with daily data. For the NMBE distributions, the impact of data frequency was not significant and various modeling approaches produced similar performance with hourly LR-based models producing the lowest absolute NMBE median values.

Table 5.1: Summary results of the hyperparameter tuning analysis

Modeling Approach	Hyperparameter	Grid-search values	Recommended value
RF	bootstrap	True, False	True
	min_samples_leaf	3, 5	3
	n_estimators	100, 200, 500	100
XGB	n_estimators	100, 200, 500	100
	eta	0.05, 0.3	0.05
	gamma	0, 0.4	0.4
SVR	C	0.1, 1, 10	0.1
	kernel	rbf, poly	rbf
SLP	learning_rate_init	0.0001, 0.0005, 0.001	0.00055
	hidden_layer_sizes	5, 7, 14, 18	13
	solver	adam, sgd	sgd

5.1.3 Impact of Building Typology and Climate

The climate effects did not suggest any patterns of deterioration or improvement of whether the data is retrieved from a cooling or a heating dominant climate. The building typology effect was limited to a set of building types that have more than 10 buildings after filtering the data with low training prediction accuracy levels. The resulting evaluation metrics' distributions for the available building typologies did not result in any significant variations in the performance of the data-driven models indicating that the building typology may not have a significant effect on the selection of the best models.

5.1.4 Impact of Training Period Selection

The effects of training period was analyzed with buildings that have at least 2-year dataset. The reduction of the training period from 12 to 9 months did not significantly reduce the prediction accuracy levels compared to reducing it from 9 to 6 months. Finally, the training period effects were evaluated for various climate zones and selection of continuous quarters for training. No consistent impacts were observed as similar climate zones produced different best quarters based on both CV(RMSE) and NMBE metrics.

5.2 Future Work

Although the findings of this thesis depict a promising performance of the data-droven models considered by the MVBEP, there are additional improvements before making the tool suitable for performing M&V analyses by the general public. The developed MVBEP tool so far can perform the tasks described by Chapter 3 including generating reports to inspect the initialization and development process along with interpretation and quantification. Some of improvements that can be considered as part of future work are described in the following sections.

5.2.1 Robust Generalization Test

The BDG2 database has one of the largest publicly available datasets for building energy consumption. However, as highlighted in Section 4.3, not all the buildings of the database have the required data completeness and quality to be used for testing the ability the MVBEP tool to develop data-driven models

for baselining energy consumption. Moreover and as discussed in Section 4.3, the impacts of various factors such as building topology, climate zone, and training period could not be assessed with statistically acceptable outcomes. Instead, differences between CV(RMSE) distributions were analyzed via plots and quantile values. To overcome such limitations, a large and better quality dataset with randomly selected buildings to account for variations in climate zone and building type should be identified and used.

5.2.2 Non-Routine Events Adjustment

The produced output in any data-driven model is highly dependent on its inputs. For accurate quantification, the model must exhibit satisfactory prediction accuracy which can be significantly affected by the presence of outliers. A specific type of outliers is the occurrence of non-routine events which if not detected, quantification results will be misleading. The existing state of the framework expects that the energy consumption patterns do not abruptly change in the post-retrofit period. By implementing a feature to process such outliers or at least detect them to be processed manually, inaccurate AEU estimation can be avoided.

5.2.3 Complex Models Quantification Uncertainity

The accuracy levels for estimating energy savings for retrofit projects using data-driven models can be assessed using multiple factors such through observing the predicted baseline visually or comparing the achieved CV(RMSE) and NMBE scores to determine the certainty levels in the predicted values. In addition to accurate predictions, data-driven models should provide accurate confidence intervals to better determine the range of possible energy savings from building retrofits. For simple models like those based on linear regression, such an interval can be easily estimated. However, for non-linear models (e.g. neural networks or ensemble models), confidence intervals are not easily determined. These confidence intervals can be estimated using complex techniques such bootstrapping resulting in computationally expensive processes. These techniques could be added to the MVBEP tool to improve the accuracy for the estimated energy savings from retrofit project.

BIBLIOGRAPHY

[1] ABC Global. Global status report for buildings and construction. Global Alliance for Buildings and Construction, 2020.

[2] US EIA. Energy information administration "international energy outlook". Technical report, report US Department of Energy, 2021.

[3] A. Allouhi, Y. El Fouih, T. Kousksou, A. Jamil, Y. Zeraouli, and Y. Mourad. Energy consumption and efficiency in buildings: current status and future trends. Journal of Cleaner Production, 109:118–130, 2015. Special Issue: Toward a Regenerative Sustainability Paradigm for the Built Environment: from vision to reality.

[4] Yang-Yang Guo. Revisiting the building energy consumption in china: Insights from a large-scale national survey. Energy for Sustainable Development, 68:76–93, 2022.

[5] Anita Thonipara, Petrik Runst, Christian Ochsner, and Kilian Bizer. Energy efficiency of residential buildings in the european union – an exploratory analysis of cross-country consumption patterns. Energy Policy, 129:1156–1167, 2019.

[6] Radwan A. Almasri and M.S. Alshitawi. Electricity consumption indicators and energy efficiency in residential buildings in gcc countries: Extensive review. Energy and Buildings, 255:111664, 2022.

[7] Andrew Satchwell, Mary Ann Piette, Aditya Khandekar, Jessica Granderson, Natalie Mims Frick, Ryan Hledik, Ahmad Faruqui, Long Lam, Stephanie Ross, Jesse Cohen, et al. A national roadmap for grid-interactive efficient buildings. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2021.

[8] Net Zero Strategy : Build Back Greener. HM Government, UK, 2021.

[9] Smriti Mallapaty. How china could be carbon neutral by mid-century. Nature, 586(7830):482–484, 2020.

[10] Simon Roberts. Altering existing buildings in the uk. Energy Policy, 36(12):4482–4486, 2008. Foresight Sustainable Energy Management and the Built Environment Project.

[11] Vaclav Hasik, Elizabeth Escott, Roderick Bates, Stephanie Carlisle, Billie Faircloth, and Melissa M. Bilec. Comparative whole-building life cycle assessment of renovation and new construction. Building and Environment, 161:106218, 2019.

[12] US EIA. Sustainable recovery. Technical report, report US Department of Energy, 2020.

[13] International Performance Measurement & Verification protocol. Efficiency Valuation Organization, 2016.

[14] ASHRAE Gudeline 14: Performance measurement protocols for commercial buildings. Guideline, American Society of Heating Refrigerating and Air-Conditioning Engineers, 2010.

[15] Zhenjun Ma, Paul Cooper, Daniel Daly, and Laia Ledo. Existing building retrofits: Methodology and state-of-the-art. Energy and Buildings, 55:889–902, 2012. Cool Roofs, Cool Pavements, Cool Cities, and Cool World.

[16] Stamatis Karnouskos, Orestis Terzidis, and Panagiotis Karnouskos. An advanced metering infrastructure for future energy networks. In New Technologies, Mobility and Security, pages 597–606. Springer, 2007.

[17] Jason Kupser, Sophia Francois, Joshua Rego, Peter Steele-Mosey, Toben Galvin, and Craig McDonald. M&v 2.0: Hype vs. reality. In Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings, 2016.

[18] Drury B. Crawley, Linda K. Lawrie, Frederick C. Winkelmann, W.F. Buhl, Y.Joe Huang, Curtis O. Pedersen, Richard K. Strand, Richard J. Liesen, Daniel E. Fisher, Michael J. Witte, and Jason Glazer. Energyplus: creating a new-generation building energy simulation program. Energy and Buildings, 33(4):319–331, 2001. Special Issue: BUILDING SIMULATION'99.

[19] University of Wisconsin-Madison. Solar Energy Laboratory. TRNSYS, a transient simulation program. Madison, Wis. : The Laboratory, 1975., 1975. Loose-leaf for updating.;March 31, 1975.;"This manual, and the TRNSYS program it describes, were developed under grants from the RANN program of the National Science Foundation (Grant GI 34029), and from the Energy Research and Development Administration (Contract E(11-1)-2588).

[20] F C Winkelmann, B E Birdsall, W F Buhl, K L Ellington, A E Erdem, J J Hirsch, and S Gates. Doe-2 supplement: Version 2.1e.

[21] DesignBuilder Software. Designbuilder.

[22] Dietmar Siegele, Eleonora Leonardi, and Fabian Ochs. A new matlab simulink toolbox for dynamic building simulation with bim and hardware in the loop compatibility. In Building Simulation, 2019.

[23] Michael Wetter, Wangda Zuo, Thierry S Nouidui, and Xiufeng Pang. Modelica buildings library. Journal of Building Performance Simulation, 7(4):253–270, 2014.

[24] Ming-Tsun Ke, Chia-Hung Yeh, and Jhong-Ting Jian. Analysis of building energy consumption parameters and energy savings measurement and verification by applying equest software. Energy and Buildings, 61:100–107, 2013.

[25] Alessandro Piccinini, Magdalena Hajdukiewicz, and Marcus M. Keane. A novel reduced order model technology framework to support the estimation of the energy savings in building retrofits. Energy and Buildings, 244:110896, 2021.

[26] A. Giretti, M. Vaccarini, M. Casals, M. Macarulla, A. Fuertes, and R.V. Jones. Reduced-order modeling for energy performance contracting. Energy and Buildings, 167:216–230, 2018.

[27] Yongbao Chen, Mingyue Guo, Zhisen Chen, Zhe Chen, and Ying Ji. Physical energy and data-driven models in building energy prediction: A review. Energy Reports, 8:2656–2671, 2022.

[28] Web of science. https://www.webofscience.com/wos/woscc/basic-search. Accessed: 2022-01-03, Clavirate Analytics London, UK.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[30] Jing Liang, Yueming Qiu, Timothy James, Benjamin L. Ruddell, Michael Dalrymple, Stevan Earl, and Alex Castelazo. Do energy retrofits work? evidence from commercial and residential buildings in phoenix. Journal of Environmental Economics and Management, 92:726–743, 2018.

[31] Zhaohua Wang, Qiang Liu, and Bin Zhang. What kinds of building energy-saving retrofit projects should be preferred? efficiency evaluation with three-stage data envelopment analysis (dea). Renewable and Sustainable Energy Reviews, 161:112392, 2022.

[32] Deborah L Brodt-Giles and Michael N Rossol. Open energy data initiative: Advancing analytics and research innovation through improved data access.

[33] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Linear model selection and regularization. In Springer Texts in Statistics, Springer texts in statistics, pages 203–264. Springer New York, New York, NY, 2013.

[34] Johanna L. Mathieu, Phillip N. Price, Sila Kiliccote, and Mary Ann Piette. Quantifying changes in building electricity use, with application to demand response. IEEE Transactions on Smart Grid, 2(3):507–518, 2011.

[35] Jessica Granderson, Samir Touzani, Claudine Custodio, Michael D. Sohn, David Jump, and Samuel Fernandes. Accuracy of automated measurement and verification (m andv) techniques for energy savings in commercial buildings. Applied Energy, 173:296–308, 7 2016.

[36] Moon Keun Kim, Yang Seon Kim, and Jelena Srebric. Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression. Sustainable Cities and Society, 62:102385, 11 2020.

[37] Haidong Wang, Yuantao Xue, and Yu Mu. Assessment of energy savings by mechanical system retrofit of existing buildings. Procedia Engineering, 205:2370–2377, 2017. 10th International Symposium on Heating, Ventilation and Air Conditioning, ISHVAC2017, 19-22 October 2017, Jinan, China.

[38] Suhaidi Aris, Nofri Dahlan, Mohd Nasrun Mohd Nawi, Tengku Nizam, and Mohamad Tahir. Quantifying energy savings for retrofit centralized hvac systems at selangor state secretary complex. Jurnal Teknologi, 77:93–100, 09 2015.

[39] Benedetto Grillone, Gerard Mor, Stoyan Danov, Jordi Cipriano, Florencia Lazzari, and Andreas Sumper. Baseline energy use modeling and characterization in tertiary buildings using an interpretable bayesian linear regression methodology. Energies, 14(17), 2021.

[40] Aaron Zeng, Hodde Ho, and Yao Yu. Prediction of building electricity usage using gaussian process regression. Journal of Building Engineering, 28:101054, 2020.

[41] Minjae Shin and Sung Lok Do. Prediction of cooling energy use in buildings using an enthalpy-based cooling degree days method in a hot and humid climate. Energy and Buildings, 110:57–70, 2016.

[42] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. Trees and rules. In Data Mining, pages 209–242. Elsevier, 2017.

[43] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. Ensemble learning. In Data Mining, pages 479–501. Elsevier, 2017.

[44] Robert E Schapire. Explaining adaboost. In Empirical inference, pages 37–52. Springer, 2013.

[45] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001.

[46] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4):1–4, 2015.

[47] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30, 2017.

[48] Alexey Malistov and Arseniy Trushin. Gradient boosted trees with extrapolation. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pages 783–789, 2019.

[49] Samir Touzani, Jessica Granderson, and Samuel Fernandes. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158:1533–1543, 1 2018.

[50] Zakia Afroz, H. Burak Gunay, William O'Brien, Guy Newsham, and Ian Wilton. An inquiry into the capabilities of baseline building energy modelling approaches to estimate energy savings. Energy and Buildings, 244:111054, 8 2021.

[51] Marc Agenis-Nevers, Yuqi Wang, Muriel Dugachard, Raphael Salvazet, Gwenaelle Becker, and Damien Chenu. Measurement and verification for multiple buildings: An innovative baseline model selection framework applied to real energy performance contracts. Energy and Buildings, 249:111183, 10 2021.

[52] Yang Liu, Hongyu Chen, Limao Zhang, and Zongbao Feng. Enhancing building energy efficiency using a random forest model: A hybrid prediction approach. Energy Reports, 7:5003–5012, 2021.

[53] Zeyu Wang, Yueren Wang, Ruochen Zeng, Ravi S. Srinivasan, and Sherry Ahrentzen. Random forest based hourly building energy prediction. Energy and Buildings, 171:11–25, 7 2018.

[54] Ran Wang, Shilei Lu, and Qiaoping Li. Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings. Sustainable Cities and Society, 49:101623, 8 2019.

[55] Zhenxiang Dong, Jiangyan Liu, Bin Liu, Kuining Li, and Xin Li. Hourly energy consumption prediction of an office building based on ensemble learning and energy consumption pattern classification. Energy and Buildings, 241:110929, 2021.

[56] Zeyu Wang, Yueren Wang, and Ravi S. Srinivasan. A novel ensemble learning approach to support building energy use prediction. Energy and Buildings, 159:109–122, 2018.

[57] Lingyan Cao, Yongkui Li, Jiansong Zhang, Yi Jiang, Yilong Han, and Jianjun Wei. Electrical load prediction of healthcare buildings through single and ensemble learning. Energy Reports, 6:2751–2767, 2020.

[58] Muhammad Waseem Ahmad, Monjur Mourshed, and Yacine Rezgui. Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption. Energy and Buildings, 147:77–89, 2017.

[59] Lu Yan and Meng Liu. A simplified prediction model for energy use of air conditioner in residential buildings based on monitoring data from the cloud platform. Sustainable Cities and Society, 60:102194, 2020.

[60] Yao Huang, Yue Yuan, Huanxin Chen, Jiangyu Wang, Yabin Guo, and Tanveer Ahmad. A novel energy demand prediction strategy for residential buildings based on ensemble learning. Energy Procedia, 158:3411–3416, 2019. Innovative Solutions for Energy Transitions.

[61] Ali S Al Bataineh. A gradient boosting regression based approach for energy consumption prediction in buildings. Advances in energy research, 6(2):91–101, 2019.

[62] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Support vector machines. In Springer Texts in Statistics, Springer texts in statistics, pages 337–372. Springer New York, New York, NY, 2013.

[63] Richard E. Edwards, Joshua New, and Lynne E. Parker. Predicting future hourly residential electrical consumption: A machine learning case study. Energy and Buildings, 49:591–603, 6 2012.

[64] K.P. Amber, R. Ahmad, M.W. Aslam, A. Kousar, M. Usman, and M.S. Khan. Intelligent techniques for forecasting electricity consumption of buildings. Energy, 157:886–893, 2018.

[65] Edward Y Chang. Psvm: Parallelizing support vector machines on distributed computers. In Foundations of Large-Scale Multimedia Information Management and Retrieval, pages 213–230. Springer, 2011.

[66] Hai Xiang Zhao and Frédéric Magoulès. Parallel support vector machines applied to the prediction of multiple buildings energy consumption. Journal of Algorithms & Computational Technology, 4(2):231–249, 2010.

[67] Bing Dong, Cheng Cao, and Siew Eang Lee. Applying support vector machines to predict building energy consumption in tropical region. Energy and Buildings, 37:545–553, 5 2005.

[68] Marek Borowski and Klaudia Zwolińska. Prediction of cooling energy consumption in hotel building using machine learning techniques. Energies, 13(23), 2020.

[69] Aaron Zeng, Sheng Liu, and Yao Yu. Comparative study of data driven methods in building electricity use prediction. Energy and Buildings, 194:289–300, 7 2019.

[70] Todorka Samardzioska, Valentina Zileska Pancovska, Silvana Petrushev, and Blagica Sekovska. Prediction of energy consumption in buildings using support vector machine. Tehnicki vjesnik - Technical Gazette, 28(2), April 2021.

[71] Minglei Shao, Xin Wang, Zhen Bu, Xiaobo Chen, and Yuqing Wang. Prediction of energy consumption in hotel buildings via support vector machines. Sustainable Cities and Society, 57:102128, 2020.

[72] Charu C Aggarwal. Training deep neural networks. In Neural Networks and Deep Learning, pages 105–167. Springer International Publishing, Cham, 2018.

[73] Chengdong Li, Zixiang Ding, Dongbin Zhao, Jianqiang Yi, and Guiqing Zhang. Building energy consumption prediction: An extreme deep learning approach. Energies, 10(10), 2017.

[74] Colm V. Gallagher, Kevin Leahy, Peter O'Donovan, Ken Bruton, and Dominic T.J. O'Sullivan. Development and application of a machine learning supported methodology for measurement and verification (m&v) 2.0. Energy and Buildings, 167:8–22, 2018.

[75] Yuna Zhang, Zheng O'Neill, Bing Dong, and Godfried Augenbroe. Comparisons of inverse modeling approaches for predicting building energy performance. Building and Environment, 86:177–190, 2015.

[76] Burak Gunay, Weiming Shen, and Guy Newsham. Inverse blackbox modeling of the heating and cooling load in office buildings. Energy and Buildings, 142:200–210, 2017.

[77] Iffat Ridwana, Nabil Nassif, and Wonchang Choi. Modeling of building energy consumption by integrating regression analysis and artificial neural network with data classification. Buildings, 10(11):198, 2020.

[78] Shalika Walker, Waqas Khan, Katarina Katic, Wim Maassen, and Wim Zeiler. Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. Energy and Buildings, 209:109705, 2020.

[79] Kwonsik Song, Nahyun Kwon, Kyle Anderson, Moonseo Park, Hyun-Soo Lee, and SangHyun Lee. Predicting hourly energy consumption in buildings using occupancy-related characteristics of end-user groups. Energy and Buildings, 156:121–133, 2017.

[80] Kangji Li, Chenglei Hu, Guohai Liu, and Wenping Xue. Building's electricity consumption prediction using optimized artificial neural networks and principal component analysis. Energy and Buildings, 108:106–113, 2015.

[81] Henrique Pombeiro, Rodolfo Santos, Paulo Carreira, Carlos Silva, and João M.C. Sousa. Comparative assessment of low-complexity models to predict electricity consumption in an institutional building: Linear regression vs. fuzzy modeling vs. neural networks. Energy and Buildings, 146:141–151, 7 2017.

[82] K.P. Amber, M.W. Aslam, and S.K. Hussain. Electricity consumption forecasting models for administration buildings of the uk higher education sector. Energy and Buildings, 90:127–136, 2015.

[83] Zhaoyang Ye and Moon Keun Kim. Predicting electricity consumption in a building using an optimized back-propagation and levenberg–marquardt back-propagation neural network: Case study of a shopping mall in china. Sustainable Cities and Society, 42:176–183, 2018.

[84] Frank E Harrell, Jr. Regression modeling strategies. Springer Series in Statistics. Springer International Publishing, Cham, Switzerland, October 2016.

[85] Oliver Kramer. Unsupervised k-nearest neighbor regression. arXiv preprint arXiv:1107.3600, 2011.

[86] Enno Mammen and James S Marron. Mass recentred kernel smoothers. Biometrika, 84(4):765–777, 1997.

[87] W.T. Ho and F.W. Yu. Chiller system optimization using k nearest neighbour regression. Journal of Cleaner Production, 303:127050, 2021.

[88] Colm V. Gallagher, Ken Bruton, Kevin Leahy, and Dominic T.J. O'Sullivan. The suitability of machine learning to minimise uncertainty in the measurement and verification of energy savings. Energy and Buildings, 158:647–655, 2018.

[89] Ran Wang, Shilei Lu, and Wei Feng. A novel improved model for building energy consumption prediction based on model integration. Applied Energy, 262:114561, 2020.

[90] W.T. Ho and F.W. Yu. Measurement and verification of energy performance for chiller system retrofit with k nearest neighbour regression. Journal of Building Engineering, 46:103845, 2022.

[91] Widyaning Chandramitasari, Bobby Kurniawan, and Shigeru Fujimura. Building deep neural network model for short term electricity consumption forecasting. In 2018 International Symposium on Advanced Intelligent Informatics (SAIN), pages 43–48. IEEE, 2018.

[92] Robert Henson. Meteorology today. CENGAGE Learning Custom Publishing, Mason, OH, 12 edition, January 2018.

[93] Prashant Anand, Chirag Deb, Ke Yan, Junjing Yang, David Cheong, and Chandra Sekhar. Occupancy-based energy consumption modelling using machine learning algorithms for institutional buildings. Energy and Buildings, 252:111478, 2021.

[94] Kangji Li, Jinxing Zhang, Xu Chen, and Wenping Xue. Building's hourly electrical load prediction based on data clustering and ensemble learning strategy. Energy and Buildings, 261:111943, 2022.

[95] Ruochen Lei and Jian Yin. Prediction method of energy consumption for high building based on lmbp neural network. Energy Reports, 8:1236–1248, 2022. 2021 International Conference on New Energy and Power Engineering.

[96] Yuan Gao and Yingjun Ruan. Interpretable deep learning model for building energy consumption prediction based on attention mechanism. Energy and Buildings, 252:111379, 2021.

[97] Peder Bacher, Henrik Madsen, Henrik Aalborg Nielsen, and Bengt Perers. Short-term heat load forecasting for single family houses. Energy and Buildings, 65:101–112, 2013.

[98] Julian J Faraway. Linear Models with R. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, Philadelphia, PA, 2 edition, July 2014.

[99] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In Joint European conference on machine learning and knowledge discovery in databases, pages 313–325. Springer, 2008.

[100] David Lindelöf, Mohammad Alisafaee, Pierluca Borsò, Christian Grigis, and Jean Viaene. Bayesian verification of an energy conservation measure. Energy and Buildings, 171:1–10, 2018.

[101] Benedetto Grillone, Gerard Mor, Stoyan Danov, Jordi Cipriano, and Andreas Sumper. A data-driven methodology for enhanced measurement and verification of energy efficiency savings in commercial buildings. Applied Energy, 301:117502, 11 2021.

[102] Chuan Zhang, Liwei Cao, and Alessandro Romagnoli. On the feature engineering of building energy data mining. Sustainable Cities and Society, 39:508–518, 2018.

[103] Cheng Fan, Yongjun Sun, Yang Zhao, Mengjie Song, and Jiayuan Wang. Deep learning-based feature engineering methods for improved building energy prediction. Applied Energy, 240:35–45, 2019.

[104] Ngo Phil. Openeemeter. https://github.com/openeemeter/eemeter, 2021.

[105] CalTRACK. Caltrack methods. https://www.caltrack.org/.

[106] SBW. Ecam (energy charting & metrics). https://sbwconsulting.com/ecam/, Jan 2022.

[107] KW Engineering. Nmecr (normalized metered energy consumption). https://github.com/kW-Labs/nmecr, 2022.

[108] LBNL. Rmv2.0 - lbnl m&v2.0 tool. https://github.com/LBNL-ETA/RMV2.0, 2020.

[109] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. A systematic literature review on applying crisp-dm process model. Procedia Computer Science, 181:526–534, 2021. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.

[110] holidays.

[111] Louis Owen. Hyperparameter Tuning with Python. Packt Publishing, Birmingham, England, July 2022.

[112] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery, 9(3):e1301, 2019.

[113] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415:295–316, 2020.

[114] M.Y Rafiq, G Bugmann, and D.J Easterbrook. Neural network design for engineering applications. Computers & Structures, 79(17):1541–1552, 2001.

[115] Aurelien Geron. Hands-on machine learning with scikit-learn and TensorFlow. O'Reilly Media, Sebastopol, CA, March 2017.

[116] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[117] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 4765–4774. Curran Associates, Inc., 2017.

[118] Parker Andrew Fontanini Anthony Present Elaina Reyna Janet Adhikari Rajendra Bianchi Carlo CaraDonna Christopher Dahlhausen Matthew Kim Janghyun LeBar Amy Liu Lixi Praprost Marlena Zhang Liang DeWitt Peter Merket Noel Speake Andrew Hong Tianzhen Li Han Mims Frick Natalie Wang Zhe Blair Aileen Horsey Henry Roberts David Trenbath Kim Adekanye Oluwatobi Bonnema Eric El Kontar Rawad Gonzalez Jonathan Horowitz Scott Jones Dalton Muehleisen Ralph Platthotam Siby Reynolds Matthew Robertson Joseph Sayers Kevin Wilson, Eric and Qu. Li. End-use load profiles for the u.s. building stock.

[119] Rob Guglielmetti, Dan Macumber, and Nicholas Long. Openstudio: an open source integrated analysis platform. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States), 2011.

[120] T Agami Reddy, Itzhak Maor, and Chanin Panjapornpon. Calibrating detailed building energy simulation programs with measured data—part i: General methodology (rp-1051). Hvac&R Research, 13(2):221–241, 2007.

[121] Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W Hobson, Zixiao Shi, and Forrest Meggers. The building data genome project 2, energy meter data from the ashrae great energy predictor iii competition. Scientific data, 7(1):1–13, 2020.

[122] ANSI/ASHRAE/IES Standard 90.1: Energy Standard for Buildings Except Low-Rise Residential Buildings. Standard, American Society of Heating Refrigerating and Air-Conditioning Engineers, 2019.

Acronyms

AIC	Akaike Information Criterion
AMI	Advanced Metering Infrastructure
ANN	Artificial Neural Network
ASHRAE	American Society of Heating Ventilation Refrigeration and Air-conditioning Engineers
BIC	Bayesian Information Criterion
BMS	Building Management System
CDD	Cooling Degree Days
DT	Decision Tree
ECM	Energy Conservation Measure
EIA	Energy Information Administration
ELM	Extreme Learning Machine
FFNN	Feed Forward Neural Network
GBM	Gradient Boosting Machine
GOF	Goodness-of-Fit
HDD	Heating Degree Days
HVAC	Heating Ventilation and Air Conditioning
IPMVP	International Performance Measurement and Verification Protocol
KNN	K-Nearest Neighbor
LCA	Life Cycle Assessment
LGBM	Light Gradient Boosting Machine
LR	Linear Regression
LS-SVM	Least Squares Support Vector Machine
M&V	Measurement and Verification
MBE	Mean Bias Error
MLP	Multilayer Percepton

NARX	Nonlinear Autoregressive with Exogenous inputs
NMBE	Normalized Mean Bias Error
OECD	Organization for Economic Cooperation and Development
OEDI	Open Energy Data Initiative
OLS	Ordinary Least Square
PI-SVM	Parallel Implemented Support Vector Machine
RBFNN	Radial Basis Function Neural Network
RC	Resistance and Capacitance
RF	Random Forest
ReLU	Rectified Linear Unit
RSS	Residual Sum of Squares
SLP	Single Layer Perceptron
SVM	Support Vector Machine
TOWT	Time of Week and Temperature
VIF	Variance Inflation Factor
WLS	Weighted Least Squares
XGB	Extreme Gradient Boosting