Identification of thermal building properties using gray box and deep learning methods

by

Gaby Baasch
B.Sc. University of British Columbia, 2015

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Civil Engineering

We acknowledge with respect the Lekwungen peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

Identification of thermal building properties using gray box and deep learning methods

by

Gaby Baasch
B.Sc. University of British Columbia, 2015

Supervisory Committee

Dr. Ralph Evins, Supervisor
(Department of Civil Engineering)

Dr. Tom Gleeson, Departmental Member
(Department of Civil Engineering)

Abstract

Enterprising technologies and policies that focus on energy reduction in buildings are paramount to achieving global carbon emissions targets. Energy retrofits, building stock modelling, heating, ventilation, and air conditioning (HVAC) upgrades and demand side management all present high leverage opportunities in this regard. Advances in computing, data science and machine learning can be leveraged to enhance these methods and thus to expedite energy reduction in buildings but challenges such as lack of data, limited model generalizability and reliability and un-reproducible studies have resulted in restricted industry adoption [44]. In this thesis, rigorous and reproducible studies are designed to evaluate the benefits and limitations of state-of-the-art machine learning and statistical techniques for high-impact applications, with an emphasis on addressing the challenges listed above.

The scope of this work includes calibration of physics-based building models and supervised deep learning, both of which are used to estimate building properties from real and synthetic data.

- Original grey-box methods are developed to characterize physical thermal properties (*RC* and *RK*)from real-world measurement data.
- The novel application of supervised deep learning for thermal property estimation and HVAC systems identification is shown to achieve state-of-the-art performance (root mean squared error of 0.089 and 87% validation accuracy, respectively).
- A rigorous empirical review is conducted to assess which types of gray and black box models are most suitable for practical application. The scope of the review is wider than previous studies, and the conclusions suggest a re-framing of research priorities for future work.
- Modern interpretability techniques are used to provide unique insight into the learning behaviour of the black box methods.

Overall, this body of work provides a critical appraisal of new and existing data-driven approaches for thermal property estimation in buildings. It provides valuable and novel insight into barriers to widespread adoption of these techniques and suggests pathways forward. Performance benchmarks, open-source model code and a parametrically generated, synthetic dataset are provided to support further research and to encourage industry adoption of the approaches. This lays the necessary groundwork for the accelerated adoption of data-driven models for thermal property identification in buildings.

Table of Contents

List of Tables

List of Figures

Author Contributions

This is an interdisciplinary thesis spanning both civil engineering and computer science. Journal papers are the preferred publishing venue for the former, while conferences are preferred for the latter. The publications that comprise this work were submitted to the peer reviewed venues for both disciplines. Chapter 2 was published at the ACM BuildSys Conference (acceptance rate 31%), Chapter 3 was published at the Climate Change AI workshop (less 50% acceptance) at NeurIPS, the largest machine learning conference in the world. Chapter 4 has been submitted for publication in the Energy and Buildings journal. Chapter 5 is prepared for submission to ACM e-Energy 2021. Full citations and author contribution details for each work are given below.

Chapter 2: Baasch G., Wicikowski A., Faure G., Evins R. Comparing Gray Box Methods to Derive Building Properties from Smart Thermostat Data. *6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19). ACM, New York, NY, USA*

GB implemented the methods, preformed the analysis, wrote the manuscript and contributed to developing the methodology. WA contributed to implementing the methods, developing the methodology and formatting the results. GF contributed to developing the methodology and writing the manuscript. RE supervised the work, contributed to the methodology and contributed to writing the manuscript.

Chapter 3: Baasch G., Evins R. Targeting Buildings for Energy Retrofit Using Recurrent Neural Networks with Multivariate Time Series *Climate Change AI workshop at the 33rd Conference on Neural Information Processing Systems (NeurIPS '20). Vancouver, BC, CA*

GB implemented the methods, performed the analysis and wrote the manuscript. RE supervised the work and edited the manuscript.

Chapter 4: Baasch G., Westermann P., Evins R. Identifying Whole-Building Heat Loss Coefficient from Heterogeneous Sensor Data: An Empirical Survey of Gray and Black Box Approaches *Submitted to the Energy and Buildings journal and undergoing revisions.*

GB developed the methodology, implemented and analyzed the Energy Signatures, the RC Network models and the deep learning methods. PW contributed to developing the methodology and implemented and analyzed the surrogate calibration methods. GB and PW co-wrote the manuscript. RE supervised the work and edited the manuscript.

Chapter 5: Baasch G., Evins R. Visual Explanations from Neural Networks Trained on Simulated Building Sensor Data *Prepared for submission to ACM e-Energy 2021.*

GB wrote the code, performed the analysis and wrote the manuscript. RE supervised the work, provided analysis suggestions and edited the manuscript.

Acknowledgements

Thank you to my supervisor, Dr. Ralph Evins, for always encouraging me to live a balanced lifestyle and for teaching me to be ambitious.

I would also like to thank all of my colleagues at the Energy in Cities group for the brainstorming sessions, for promptly answering my buildings questions and for the cocktail camaraderie. Special thank you to Paul for helping me to stay sane while writing and for teaching me the importance of literature review. Also, thank you to the IESVic and Civil Engineering staff who have provided relentless support, and to the external examiner, Dr. Margaret Storey.

I could not have done this without my family and friends; I am forever grateful.

Finally, thank you Lexi, for inspiring me, and for all of our adventures together.

Chapter 1

Introduction

At the time of writing, over 1,800 jurisdictions spanning 31 countries have declared a climate emergency,[1] and governments are scrambling to develop evidence-based carbon emissions reduction strategies. Upgrading existing buildings is a key component in the climate action plans of federal [3], provincial [2] and municipal [100] governments in Canada, where heating, cooling and electricity use in existing buildings accounts for 17% of national GHG emissions [3]. Further, decarbonizing the building stock has invaluable co-benefits including reduced consumer energy costs, job creation and improved occupant health and comfort [59] [58].

High-impact applications for carbon reduction in buildings, including retrofit analysis, stock modelling and demand-side management, are supported by the identification of building properties. Building retrofits entail upgrades that result in reduced energy use. For example, heating, ventilation and air conditioning (HVAC) systems might be replaced by more efficient alternatives, or the quality of the building constructions (i.e. the building envelope) might be improved to reduce heat loss [51]. The ability to perform mechanical systems and thermal envelope property diagnostics (respectively) is essential to the development of effective retrofit programs. Stock-level modelling of building properties allow

[1] https://www.cedamia.org/global/

stakeholders, such as municipalities, to deliver evidence-based carbon reduction plans and targets. Demand-side management schemes, which aim to control building heating systems while providing flexibility to the heating grid, also require stock-level building analysis.

To achieve global emissions reductions targets, these types of strategies must be implemented on a massive scale [26] [63]. In practice, methods for building property characterization rely on walk-throughs, surveys or the collection of in-situ measurements [72], which are neither scaleable nor cost-effective. Efficient and reliable methods that extract properties from large datasets are therefore required. Meanwhile, the rate of data collection in buildings is extraordinary; worldwide, more than one billion smart metering devices will be installed by the end of 2020 [83], and large construction markets have or will have (e.g. Canada [42]) nationwide coverage. These observations indicate that data-driven, statistical and machine learning approaches will be integral for decarbonizing the building stock.

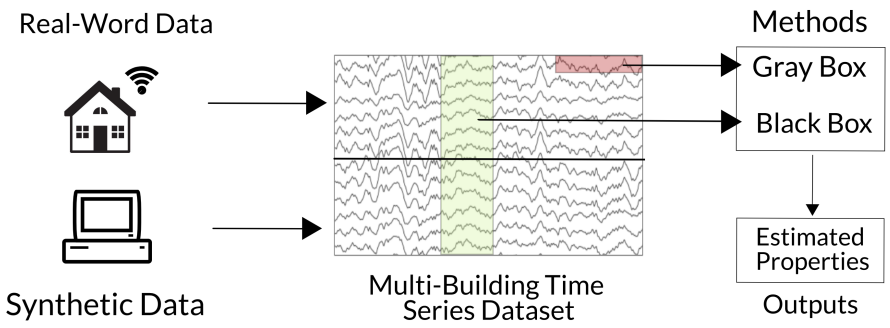


Figure 1.1: Workflow for (1) gray and (2) black box methods. (1) require both a physics-based model of the building and fitting to measurement data. They estimate properties by calibrating parameters to measurement data from a single building at a time. (2) are purely data-driven; they build statistical representations from large amounts of data to predict on unseen examples. Supervised deep learning is a popular approach to black box modelling.

Two paradigms[2] for data-driven building diagnostics exist: gray box and black box

[2] A third modelling paradigm, known as white box modelling, also exists. These methods are purely physics-based and not data-driven, so they cannot be used to estimate building properties from sensor measurements. White box models are used in this work to generate pramaterized synthetic datasets.

methods (see Figure 1.1). Together they provide comprehensive coverage of methods for data-driven estimation of building properties from large datasets, but neither have seen wide-spread adoption by the buildings industry [44]. For black box models this can be attributed in part to a lack of relevant labels. Smart meter data, for instance, increasingly provide nationwide coverage, but do not include detailed information about the building characteristics or energy loads [102]. Existing gray box methods do not require labelled data, but they have not been validated for use on large, heterogenous building sensor datasets. In general, benchmarking of both black and gray box approaches is limited, and it is often unclear whether the approaches are scalable and robust to diverse building properties.

The major contribution of this thesis is to support industry adoption of novel and existing gray box and black box methods for practical, big-data applications such as retrofit analysis and building stock modelling. This is done via rigorous, empirical validation. Two primary objectives arise:

- To assess novel and existing gray and black box methods for thermal property estimation in buildings. Both real-world data and synthetic data with ground-truth labels are used.
- To orient future research in terms of challenges that restrict industry adoption of data-driven modelling research, including data availability, reproducibility and model reliability, generalizability and transferability [45]. To support reproducibility, all the work completed over the course of this thesis is open-sourced.[3]

The following research questions are addressed in this work.

1. Can gray box models derive useful building properties from real-world datasets, in spite of limited information? (Chapter 2)

[3] Available at https://gitlab.com/energyincities

The first chapter in this thesis presents and compares three gray box methods for assessing heating characteristics of households using a real-world, smart thermostat dataset that does not contain ground truth or heating power measurements. The three methods are based on: (1) balance point plots, (2) the extraction of indoor temperature decay curves, and (3) the classic differential equation for indoor temperature. The result indicates that the methods can be used in a real-world context to ascertain relative values for the thermal characteristics of a building.

2. Can black box models predict thermal building properties using time series sensor data? (Chapter 3)

   This chapter serves as a novel showcase for how multivariate time series analysis with Gated Recurrent Units can be applied to targeted retrofit analysis via two case studies: (1) classification of building heating system type and (2) prediction of the numerical physical property that determines the rate of heat lost through the building envelope.

3. Is gray box calibration or black box learning more reliable for application on large, heterogenous building datasets? (Chapter 4)

   Seven different gray box and black box approaches for characterization of the whole-building heat loss coefficient are compared in this chapter. To do so, a synthetic dataset of 16,000 simulated buildings is created. The models are benchmarked according to four criteria: (1) data and infrastructure requirements, (2) scalability to larger datasets, (3) model validation and (4) comparison to ground truth, including an assessment of robustness to climate, construction materials, air-infiltration rate and occupant behaviour. It is shown for the first time that the deep learning methods outperform other approaches in terms of accuracy and robustness, but that all of the approaches have limitations that restrain their practical usage.

4. Can black box models that are trained on synthetic data be transferred to real world

data? (Chapter 5)

Gradient-based activation maps are used in interpretable machine learning research to highlight the important features of a datum for a specified prediction task. In this chapter activation maps are applied to illuminate how deep neural networks trained on time series inputs predict a building's heat loss coefficient (HLC). Several networks are trained on different sets of inputs, and the resulting activation maps are compared. The results indicate that the networks learn physically meaningful features from synthetic data, which in the long term might mean that pre-trained networks could be used to reduce real-world data requirements through transfer [98] or self-supervised [70] learning. This is one of the first applications of activation maps for both time series and building data.

Table 1.1 lists all of the models that are evaluated in this thesis, alongside their inputs and outputs. The 'Introduction' sections in Chapters 2-5 contain the relevant background, so to avoid repetition an additional literature review section is not included.

Chp.	Model	Data Source	Input Variables	Granularity	# Buildings	Output
2	Balance Point	smart thermostat	outdoor temp., heating system (on/off)	day	4,646	RK
2	Decay Curves	smart thermostat	indoor temp., outdoor temp.	5 min.	4,646	RC
2	Energy Balance	smart thermostat	indoor temp., outdoor temp., heating system (on/off)	5 min.	4,646	RK, RC
3	Recurrent Neural Network	smart thermostat	outdoor temp., indoor temp., heating power	5 min.	602 (train), 182 (test)	heating system type
3	Recurrent Neural Network	synthetic	outdoor temp., indoor temp., heating power	10 min.	773 (train), 193 (test)	R
4	Energy Signature (i.e. Balance Point)	synthetic	outdoor temp., indoor temp., heating power, solar gains	day	3,200	HLC (i.e. 1/R)
4	First Order Lumped Capacitance (i.e. RC) model	synthetic	outdoor temp., indoor temp., heating power, solar gains	5 min.	3,200	HLC (i.e. 1/R)
4	Second Order Lumped Capacitance (i.e. RC) model	synthetic	outdoor temp., indoor temp., heating power, solar gains	5 min.	3,200	HLC (i.e. 1/R)
4	BES* Calibration with Genetic Algorithms	synthetic	outdoor temp., indoor temp., heating power, solar gains	5 min.	12,8000 (train), 3,200 (test)	HLC (i.e. 1/R)
4	BES* Calibration with Bayesian Optimization	synthetic	outdoor temp., indoor temp., heating power, solar gains	5 min.	12,8000 (train), 3,200 (test)	HLC (i.e. 1/R)
4	Recurrent Neural Network	synthetic	outdoor temp., indoor temp., heating power, solar gains	5 min.	12,8000 (train), 3,200 (test)	HLC (i.e. 1/R)
4	Convolutional Neural Network	synthetic	outdoor temp., indoor temp., heating power, solar gains	5 min.	12,8000 (train), 3,200 (test)	HLC (i.e. 1/R)

Table 1.1: The inputs, outputs and data features for the models studied in this work. The outputs are explained in more detail in the relevant chapters. *BES refers to building energy simulation, which is a high-fidelity, white box representation of building.

Chapter 2

Comparing Gray Box Methods to Derive Building Properties from Smart Thermostat Data

2.1 Introduction

2.1.1 Motivation

Retrofitting the existing building stock is one of the primary means by which we can reduce building energy consumption and reach energy efficiency targets globally to mitigate climate change [63]. Existing buildings account for 32% of global energy demand and 30% of global carbon emissions [107]. Of that, approximately 60% of the energy required by residential buildings are for thermal uses [107]. It follows that many studies highlight the environmental necessity of retrofits. For instance a study by Deconinck and Roels et al. determined that 2050 energy reduction targets for two case studies can not be achieved without a retrofit rate of at least 2% of buildings per year [26]. In a 2011 study, Mills et al. show that retrofits in the US can result in a median of 16% whole energy building savings, with a payback period

of only 4.2 years [69]. Another area in which stock-level analysis of building properties would be valuable is in assessing the potential of demand-side management (DSM) schemes which seek to control building heating systems to provide flexibility to the electricity grid without discomforting building occupants. A quantitative and scalable approach to filtering viable building candidates would be beneficial in targeting retrofit and DSM measures and assessing the applicability of such measures to whole building stocks. This work explores methods to provide stock-level overviews of building characteristics as needed to assess the potential for such measures.

2.1.2 Research Background

In order to target buildings for envelope upgrades and to tailor appropriate construction strategies, building performance evaluation is required. Currently, techniques for evaluating thermal building characteristics require onsite measurements and performance appraisal, often combined with complex building simulations [72]. Basic energy audits include walk-through assessments and survey analysis [63], while more complex analysis involves advanced computational techniques and sensor networks. For example, Biddulph et al. and Gori et al. use Bayesian techniques take advantage of rich in-situ measurements and time series data to predict the thermophysical properties of buildings [38][12], Aznar et al. use the data from in-wall sensors to train a deep-learning model that measures and predicts heat transfer [8] and Nagy et al. implement a low cost sensor network to estimate the thermal transmittance of the building [72].

As illustrated by the examples above, there has been a lot of progress towards the energy evaluation of a single building. While indispensable, this type of analysis is not scalable and cannot filter for viable retrofit candidates at a district scale. With the advent of smart sensing technology and the internet of things (IoT) unprecedented amounts of building data are becoming available. This provides an opportunity to address the scalability issues of

building energy performance assessment. More recently, researchers are starting to take advantage of this new resource. Studies by Tabatabaei et al. and Van der Ham et al. use thermostat data to estimate the thermal characteristics of houses. The former evaluates 99 Dutch households while the latter uses 67 households [97][99]. Ghiaus uses aggregate data to predict the energy consumption and heat loss coefficient for a single building in several locations [35]. In perhaps the most large-scale study to date, Iyengar et al. use Bayesian inference over more than 10,000 buildings to create a partial ordering of buildings based on their efficiency [50]. All of the papers cited above use temperature and heating load data.

Research into the use of "big data" from smart-sensing and IoT devices to predict the thermal characteristics of buildings is still a relatively young field. The aforementioned studies provide a valuable starting point, but much work remains to be done in this area. It is still unclear what types of thermal characteristics can be estimated using big data, how reliable these estimates are, what types of data are required and what are the limitations.

2.1.3 Contribution

This paper addresses these open research questions by comparing three gray box methods which predict the thermal characteristics of buildings: balance point plots, decay curves and numerical integration of the energy balance equation. A summary of these three methods and the required data can be seen in Table 2.1, and they are described in detail in section 2.2.2. The methods in this paper are novel and differ from the aforementioned studies in several ways. First, all of the above studies use energy load data, which are not always available. In cases where energy data is available, as with smart meter data, there are problems with working backwards from aggregated loads to identify just the heating or cooling-related energy use [60]. As smart thermostats become more common[1], it is important to develop methods that can derive building properties directly from temperature data. For these reasons

[1] In the US in 2015, 10 million thermostats were purchased, 40% of which were connected [49].

Table 2.1: Model summary

Method	Parameters	Required Data	Fitting Method
Balance point	RK	Heating system duty cycle External temperature	Linear regression
Decay curves	RC	Internal temperature External temperature	Non-linear least squares
Energy balance	RK, RC	Heating system duty cycle Internal temperature	Non-linear least squares

no load profiles are used in this paper[2]. Second, this paper uses a larger dataset than most of the previous studies, with over 4,000 buildings. Finally, it is important to capture the dynamic aspects of building energy performance such as thermal mass by using the rich and informative time series data which is becoming available. Two of the three methods in this study use the sequential time series data rather than an aggregated form.

The paper is organized as follows: a description of the data and models; a results section describing model performance; and a discussion of the merits of each method.

2.2 Methodology

To compare possible data analysis techniques for deriving the thermal characteristics of buildings from temperature time series, three gray box models were implemented:

(1) Balance point plots of daily heating demand against outdoor temperature.

(2) Exponential decay curves of indoor temperature following heating setpoint drops.

(3) Numerical integration of the 1D heat conduction differential equation from a known initial value.

Each model is fitted to each building for which suitable data are available to estimate the thermal parameters of that building.

A general analysis pipeline was created to compare each of the three models. Code for

[2]Heating system duty cycle is used as a proxy for an energy use in one of the three methods.

the pipeline can be found at https://gitlab.com/energyincities/besos-public/publications. An important part of the process involved creating appropriate filters for the time series data. Each time new filters were created, the analysis pipeline was rerun and the new results were compared. Initially, to reduce the computational cost and to avoid overfitting the filters, only a small subset (20%) of the data was evaluated. The entire dataset was analyzed only after the filters were finalized.

In this section each method is discussed in detail, along with the data, the final filters used in the preprocessing phase and the metrics used for the comparison of the models. The limitations of each method are also discussed.

2.2.1 Data

The Dataset

The dataset for this research was acquired through the ecobee Smart Thermostat Donate Your Data program[3]. The original dataset consists of over a terabyte of anonymized smart thermostat data from 76,000 households worldwide. The data for each household consists of both time series data and metadata. The time series data spans from 2015 to 2018 with a 5 minute granularity and includes indoor and outdoor temperatures, heating and cooling system duty cycles, occupant schedules and heating and cooling setpoints. The inside temperature is typically measured at a single thermostat and the outdoor temperature data is acquired from the nearest available weather station for each building. The metadata includes building characteristics such as size, age, heating system type, location, and occupancy. A more detailed description of the dataset can be found in "A longitudinal study of thermostat behaviours based on climate" [49].

[3] https://www.ecobee.com/donateyourdata/

Preprocessing

The methods proposed in this study aim to predict the thermal characteristics of buildings in cold climates where heating is required to maintain internal temperatures, though the methods could all be inverted to work in cooling-dominated climates. The dataset was reduced to include only buildings in the cold climates of Ontario, Canada and New York, USA. To further reduce potential confounding factors, only homes without auxiliary heating systems were evaluated. After this filters was applied, the dataset included 4,646 buildings.

In addition to the metadata filtering described above, the time periods over which the models are trained was limited to times when the outdoor temperature is lower than the indoor temperature and there are no solar gains. We therefore look only at time periods that are during the night (08:00 PM - 05:00 AM) in the winter months (November - February).

Specific methods also required particular filtering to obtain time periods that were consistent with the assumptions of that method. The way data was filtered had a significant effect on the performance of the methods used; understanding which filters were used to obtain the final result and the reason why will therefore be important for additional research in this domain. Details of the preprocessing filters applied for each method are given in the relevant sections below.

2.2.2 Models

The methods presented in this paper use a gray box approach, that is they combine the data-driven nature of black box modelling with the use of explicit domain knowledge and the physics equations of white box models. Black box models, which are commonly used by data scientists and machine learning practitioners, are statistical models used to extract and predict interesting information from large data sets. These types of models are seeing increasing uptake and application in a wide variety of fields; they require minimal domain

knowledge and can be used in spite of limited information. White box models, on the other hand, model the detailed physical behaviour of a system. They are more difficult to implement and often require highly specialized domain expertise, however they are easier to interpret and can be more reliable than black-box methods. Gray box methods marry these two approaches by formulating a statistical model according to a-priori physical knowledge. In this way, the physical parameters of a system can be reliably described, predicted and estimated, even in lieu of missing information.

The physics on which the models in this paper are based is derived from the thermal energy balance of a building, as described in the section below. This is followed by a description of each the methods and their associated fitting process.

Thermal Energy Balance in a Building

The thermal energy balance in a building can be expressed by equation 2.1, where T_{in} is the indoor temperature, T_{ext} is the outdoor temperature, $\dot{Q}_{in}$ is the internal heat gains, $\dot{Q}_h$ is the heat flow supplied by the heating system, $\dot{Q}_{sol}$ is the solar radiation gains, $\dot{Q}_{ven}$ is the heat flow due to ventilation, C is the lumped building capacitance and R the lumped building thermal resistance [19].

$$C\frac{dT_{in}}{dt}(t) = \dot{Q}_{in}(t) + \dot{Q}_h(t) + \dot{Q}_{sol}(t) - \frac{1}{R}(T_{in}(t) - T_{ext}(t)) - \dot{Q}_{ven}(t) \tag{2.1}$$

This equation includes lumped parameters for capacitance and thermal resistance. It therefore assumes that the different parts of the building cool or warm uniformly. The thermal resistance R ($\frac{^\circ K}{W}$) represents the global insulation of the building; the higher the R value, the better insulated the building. The thermal capacitance C ($\frac{J}{^\circ K}$) describes the ability of the building fabric to store energy and therefore its inertia; buildings with high thermal mass, for example built from concrete, have high values for C.

For this work, we assume that:

- the dominating heat flows are the heat flow supplied by heating system $\dot{Q}_h$ and the heat flow due to the indoor and outdoor temperature difference.

- the heat flow supplied by heating system $\dot{Q}_h$ can be rewritten as:

$$\dot{Q}_h(t) = \delta_{on}(t) \times K$$

  where K is the heating power, assumed constant, and δ_{on} is the proportion of time that the heating system is on during a particular time interval.

The thermal energy balance therefore becomes:

$$C\frac{dT_{in}}{dt}(t) = \frac{1}{R}(T_{ext}(t) - T_{in}(t)) + \delta_{on}(t)K \tag{2.2}$$

The objective of the models in this paper is to determine the parameters R and C, to characterise the building fabric; to fit the equation when the heating system power is unknown, it may also be necessary to estimate K. Equation 2.2 must therefore be rewritten so that it can be parameterized and optimized:

$$\frac{dT_{in}}{dt}(t) = \frac{1}{RC}((T_{ext}(t) - T_{in}(t)) + \delta_{on}(t)RK) \tag{2.3}$$

$$\frac{dT_{in}}{dt}(t) = \alpha((T_{ext}(t) - T_{in}(t)) + \beta\delta_{on}(t)) \tag{2.4}$$

The parameters now become $(RC)^{-1}$ (α) and RK (β), both of which can be determined through the fitting of statistical models by various methods, as outlined below. The RC value is easy to interpret as the "time constant" of the building,[4] while RK is more abstract. In this paper RK primarily provides an additional means to compare the results of two

[4]The units of $\frac{^\circ K}{W}$ and $\frac{J}{^\circ K}$ cancel to s (*seconds*), since $1J = 1Ws$. This requires a conversion in the energy balance equation when applied to the ecobee data, which has a time resolution of 5 minutes. Values are reported in hours.

of the models, further justifying the observed correlation between the methods. It should be noted that it is impossible to derive the individual factors R, C and K independently from equation 2.2; based on temperature data alone, there is no way to distinguish a poorly insulated building with high thermal mass from a well-insulated but thermally lightweight building. This is a further justification for the use of multiple methods: if RC and RK can be estimated with reasonable accuracy, it may be possible to use assumptions about K to determine specific values of R and C.

Model 1: Balance Point

"Energy signature" methods have long been used as a tool for estimating building energy performance [39]. In these models, the energy use of a single building is plotted as a function of the outdoor temperature. Each point represents the heating load and temperature, typically aggregated by taking the mean outdoor temperature ($\overline{T}_{ext}$) and the total heating load ($\dot{Q}_{h,d}$) on a daily basis. Usually the plot shows two distinctive sections on either side of a particular value of the outside temperature called the "balance point". A linear correlation is visible below this point and, above this point, the temperature has no impact on the heating load [39]. By applying a linear regression on the portion below the balance point we can find the gradient, which represents the R value in equation 2.5 where $y = \dot{Q}_{h,d}$ and $x = \overline{T}_{ext}$. This equation is derived from equation 2.3 by assuming that on average the indoor temperature does not change during the day ($mean(\frac{dT_{in}}{dt}) = 0$). The balance point method cannot predict values for C, since the points are evaluated independently of time so no dynamic behaviour can be captured.

$$y(x) = \frac{1}{R}(\overline{T}_{in} - x) \tag{2.5}$$

The ecobee data does not provide any information about specific heating or energy load $\dot{Q}_h$, and the metadata does not contain any information on the system capacity, so the duty

cycle of the heating system δ_{on} is used instead. A daily unitless heating runtime fraction $F_{h,d}$ is derived from δ_{on} as the fraction of time periods where $\delta_{on,i} = 1$. As mentioned in section 2.2.1, to limit the impact of solar and internal gains, the fractions are computed only over the night periods. The shape of the signature produced using this heating runtime fraction $F_{h,d}$ is similar to the shapes seen in typical energy signatures (see Figure 2.1), indicating that this is a reasonable proxy. Note that Figure 2.1 shows only the data with a linear dependance on outdoor temperature, well below the "balance point", since we filter for only winter nighttime values.

Solving equation 2.2 with $y = F_{h,d}$ and $x = \overline{T}_{ext}$ we derive:

$$y(x) = \frac{1}{RK}(\overline{T}_{in} - x) \tag{2.6}$$

We can see from this equation that the slope of the line of best fit for the balance point plot is now represented by $-(RK)^{-1}$. Therefore, a simple linear regression can be used to solve for RK by finding the slope of the line of best fit.[5] Outliers, which for this model are simply defined as any points that lie more than one standard deviation away from the mean, are excluded. This is illustrated in Figure 2.1. This approach has been applied previously by in other works, where regression was used to estimate building energy performance from heating load and outdoor temperature [35].

Though finding the slope of the balance point plot is a relatively common approach to estimate the thermal characteristics of buildings, this method is subject to certain limitations. First, for this work it is assumed that the form of the scatter plot is linear, but in reality a typical building always exhibits some dynamic, non-linear behaviour [39]. Second, linear regression is highly susceptible to outliers. The way in which we perform outlier detection and removal in this study is rather crude and may accidentally remove valuable information.

[5]The *scipy.stats.linregress* python module is used to perform the linear regression, see https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html.

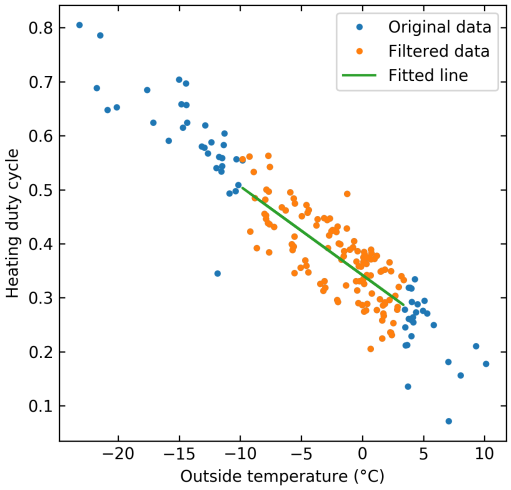


Figure 2.1: Example balance point plot showing daily sampled data (filtered for winter nighttimes), remaining data after outlier removal, and the final linear regression whose slope gives RK.

A more robust method for outlier detection should be implemented, perhaps by iteratively removing outliers over a series of regressions to obtain a minimum error.

Model 2: Decay Curves

Unlike balance point plots, which use daily aggregated values, decay curve analysis takes advantage of the rich time series data available. A typical decay curve occurs when there is no heat input into the system and the outdoor temperature is much lower than the initial indoor temperature. According to equation 2.2, at these times there will be an exponential rate of decay of the indoor temperature towards the outdoor temperature. An example is shown in Figure 2.4. With no heating and constant outdoor temperature, equation 2.7 is a specific analytical solution to the general equation 2.2. Note that since the decay curves describe the behaviour of the building when no heating is present, this method cannot predict values of K, which in any case are not relevant in assessing the building envelope.

$$\theta(t) = \theta_0 e^{\frac{-t}{RC}} \tag{2.7}$$

where $\theta(t) = T_{in}(t) - T_{ext}$.

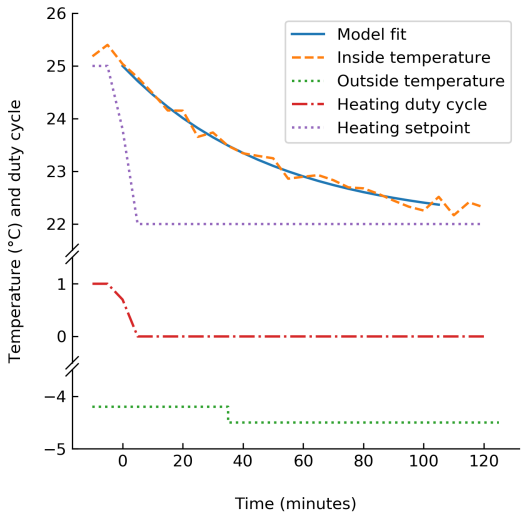


Figure 2.2: A decay curve fit for indoor temperature decrease following a setpoint drop and heating duty cycle decrease. Though plotted on the same axes, the heating duty cycle is not in temperature units; it is a unitless, proportional value.

One significant limitation of this method is the necessary assumption that the outdoor temperature is constant. To account for this we filter for periods of time across which the mean outdoor temperature remains relatively stable. We assume that small variations in outdoor temperature should not have a huge effect and therefore do not filter for complete stationarity of this value. There is a trade-off between over-filtering the data in the search for periods which are closest to the ideal and retaining many periods over which we can average the values obtained. The full set of filters used to preprocess the data for this method are shown in Table 2.2.[6]

For each building, multiple decay curves are extracted, one for each subset of the time

[6]An additional input for this and the following method was an initial guess for the parameters being predicted. This initial guess does not constitute a filter, but rather a hyperparameter, but is also given in the relevant table.

Table 2.2: Decay curve filters

Filter	Value
Stationarity of Mean in Outdoor Temperature	1.0
Minimum Indoor-Outdoor Temperature Difference	5°
Maximum Proportion of Time Heating is Added	0.1
Maximum Time Period	6 hours
Minimum Time Period	10 minutes

series which meets the filtering criteria described above. This extraction is deterministic; as long as the filters are the same, the same decay curves will always be extracted for a given time series. The number of decay curves that were found for each building can be seen in Figure 2.3.

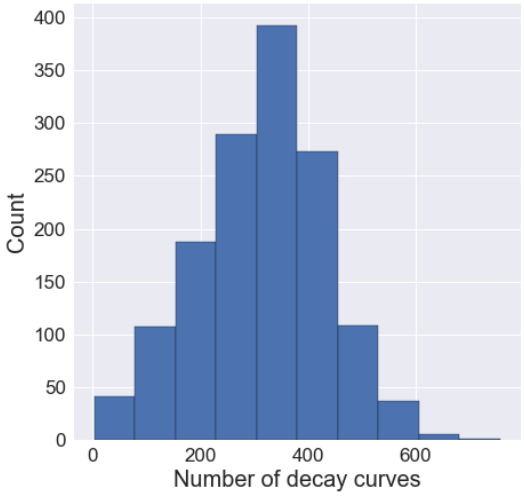


Figure 2.3: Histogram of the number of decay curves per building.

After the decay curves are extracted, the parameters in equation 2.7 are determined using a non-linear least-squares curve fitting method[7] with two parameters, RC and T_0. Using this approach, an RC value is derived for each of the decay curves available for a given building,

[7]The *scipy.optimize.curve_fit* module is used, see https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.curve_fit.html.

and the mean, median and standard deviations of these values are examined. *RC* values with a small standard deviation will be the most reliable and most likely obtained for a building for which there are many decay curves available.

Model 3: Energy Balance

This method was introduced to overcome the restriction on the decay curve method that outdoor temperature must be constant and no heating can take place. It requires fewer restrictions than the decay curve fitting as it can be applied to periods with heating and unsteady outdoor temperatures, but it is not without limitations. Filters are required to ensure that significant heating and sufficient variation of the indoor temperature occurs in a given time period.[8] The full set of filters used to preprocess the data for this method are shown in Table 2.3.

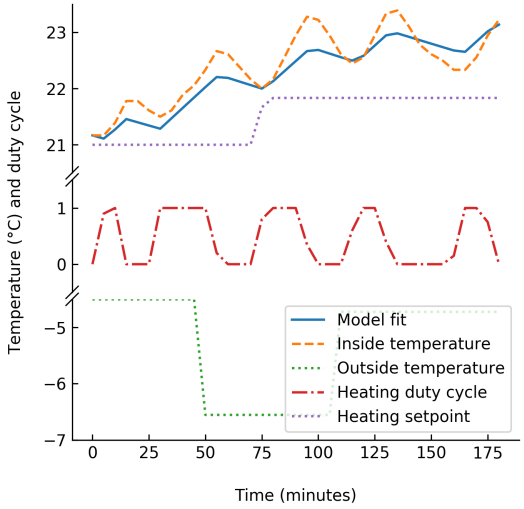


Figure 2.4: A typical energy balance fit showing how the inside temperature output changes depending on the heating duty cycle and outside temperature.

This method involves solving the differential equation in equation 4 and using Euler's

[8] If there is not enough heating over a given time period the value for *RK* cannot be predicted. If there is not enough variation in indoor temperature the rate of change is always 0 and *RK* can take on any value.

Table 2.3: Energy balance filters & other parameters

Filter	Value
Time Periods Used for Fit	30
Duration of Time Periods Used for Fit	3 hours
Fits Attempted for Each Building	10
Heating Duty Cycle Maximum	0.8
Heating Duty Cycle Minimum	0.05
Minimum Variance in Indoor Temperature	0.2

method of numerical integration. Euler's method approximates the solution to a first-order differential equation given an initial value. Using this method a new model parameterized by RC and RK can be derived:

$$T_{in,i+1} = T_{in,i} + \Delta t[\alpha((T_{ext,i} - T_{in,i}) + \beta\delta_{on,i})] \tag{2.8}$$

where Δt is a 5 minute timestep.

The equation above can be used to create a dataset for curve fitting, over which the parameters α and β are optimized. The dataset is created as follows:

1. Create random initial guess for α and β.

2. Let $i = 0$. Use the real value of T_i, alongside the initial guess from step (1) to solve for T_{i+1}.

3. Now let $i = i + 1$. Use the estimated value of T_i to solve T_{i+1}. Repeat for all n timesteps.

To avoid overfitting and to prevent error propagation through time, the steps above are repeated a number of times for different intervals in the data, specified by the variable *Time Periods Used for Fit*. This process creates a dataset of temperature values that can be compared to the real data. A non-linear curve fitting algorithm is then used to minimize the

Table 2.4: Pros and cons of each method

Model	Pros	Cons
Balance point	1. Easy to implement 2. Works with aggregated data so its applicable to many data sources	1. Aggregated so cannot describe variance in data 2. Produced many outlier results
Decay curves	1. Produced few outlier results	1. Requires a lot of data filtering 2. Assumes outdoor temperature is constant
Energy balance	1. Less filtering than the decay curves 2. Accounts for changes in outdoor temperature	2. Produced many outlier results

difference between the generated data and the real data by changing α and β. In this process a number of time periods of a certain length are selected at random from the filtered data. Various period lengths were trialled, with 3 hours giving the best results.

Similar to the decay curve method, the effectiveness of this model can be significantly reduced by disturbances such as internal gains, solar gains and ventilation, which affect the RC and RK. Additionally, lag between heating runtime and actual heating in certain heating systems can confuse the model.

2.2.3 Metrics for Model Comparison

Not all methods predict all characteristics, as shown in Table 2.1. For the decay curve and energy balance methods, multiple values are generated and averaged, giving the opportunity to judge individual model performance based on the spread of these values.

In order to compare the performance of the models in the absence of ground-truth data, several basic statistical tests are implemented to assess the similarity of the characteristics predicted for a given building. First, the final results for RK from the balance point method and the energy balance method are plotted against each other, as well as the results for RC

from the decay curve and energy balance methods. In both cases, a perfect result would be when the values are identical, that is, when all points fall exactly on the same line. By measuring the correlations of these values, we can determine whether the results are correct, relative to one another. Second, the standard deviation of the results from the decay curve and energy balance methods are evaluated. Third, a statistical t-test is used to compare population means and determine whether the absolute values produced from each method are similar. Fourth, the relative differences in the results from each method are evaluated using quantiles and plotted with a violin plot. These quantiles describe the percentages of results that are similar.

2.3 Results

2.3.1 Individual Model Performance

Before comparing methods, the fitting error of the optimization functions and the standard deviations for each method are evaluated to help us to understand how well the models performed. The model fitting for each of the three methods has an associated error which is discussed in sections 2.3.1 - 2.3.1. Standard deviation can be evaluated for the decay curve and energy balance methods, since in each of these two methods multiple values are returned for each building. The standard deviations of these values represent the reliability of the result for a particular building.

Cases where the models perform poorly, as defined by the fitting error and the standard deviations, are removed from the reported results. A summary of these values can be seen in Table 2.5. In general, the thresholds were set to reduce the number of outliers while also retaining a reasonable amount of buildings. After outlier removal there were 1443 remaining buildings, that is, 31% of the buildings produced reliable results — as defined by the threshold values in Table 2.5.

Performance of the Balance Point Method

Linear regression returns metrics that represent the goodness of fit: p-value, r-value and standard error. The null hypothesis for the p-value is that the slope of the line is zero. A scatter plot with a slope of zero indicates that there is no correlation between the variables. A result with a small p-value (below 0.05) has a statistically significant slope, meaning that the dependent variable (heating load) is affected directly by the independent variable (outdoor temperature). The standard error is measured in the units of the dependent variable, and measures the standard deviation of the errors. The r-value is the correlation coefficient, which quantifies the linear relationship between two variables. Together, these metrics can be used to evaluate the quality of the estimated gradient.

In general, the balance point method returned results that had a relatively high standard error and an r-value that did not represent a strong linear correlation. Of the three methods, the balance points seemed to be the most unreliable.

Performance of the Decay Curves Method

For the decay curve method, the standard deviation and the fitting error represent the quantitative performance of the model. In Figure 5 (d) these values are plotted against each other to represent their relationship. Higher density areas represent values for standard deviation and fitting error that were obtained for many buildings. It can be seen that there is a general positive linear trend between these two values. This indicates that as the fitting error increases, the standard deviation of the values for RC within a building also increases. Buildings with higher fitting error and standard deviations are the least reliable.

Figure 5 (d) also gives the distribution of the standard deviations (right) and fitting errors (top) for each building. This shows that the standard deviations are relatively tightly clustered around $30h$ (compared to predicted values of RC of 50 to $200h$) and moderate fitting errors of around 50 to $100°K^2$.

Performance of the Energy Balance Method

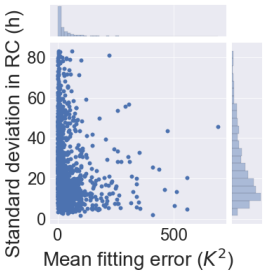
As with decay curves, the fitting error and the standard deviation are interesting quantitative indicators of model performance. The relationships between each of these measures can be seen in Figure 5. Unsurprisingly, there is a positive linear correlation between the standard deviation of RC and RK. This indicates that there is consistency in these results. On the other hand, neither the standard deviations for RC or RK have a linear relationship with the fitting error. This is an unexpected result that likely indicates that the fitting cost function is not fully expressing the goodness of fit. This could be because there is information missing from the model.

Figure 5 also shows the distribution of the standard deviations and fitting errors for each building to the right and top of each plot respectively. This shows that the fitting errors are very low, clustering very near to zero with almost all values below $50°K^2$. Standard deviations for RC are mostly clustered between between around $5h$ and $20h$, which is slightly lower than for the decay curve method. Standard deviations for RK are clustered around 5 to $15°K$ (compared to predicted values of RK of 20 to $200°K$).

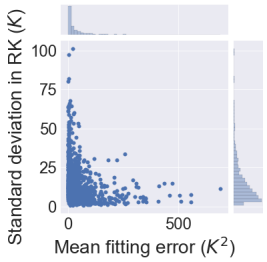
2.3.2 Model Comparison

A comparison of the three methods applied in this paper shows that there is a positive linear correlation between: (1) the model fitting method and the balance point method (used to solve for RK) and (2) the model fitting method and the decay curve method (used to solve for RC), as shown in Figure 2.6 and in the correlation values in Figure 2.8. By examining Figure 2.6, one can see that the balance point method results in some outlier values that were not caught by the parameters presented in Table 2.5 and that the energy balance method overpredicts RC compared to the decay curve method.

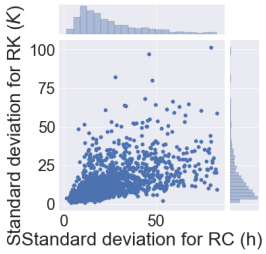
A t-test was performed to determine if the population means from the methods are



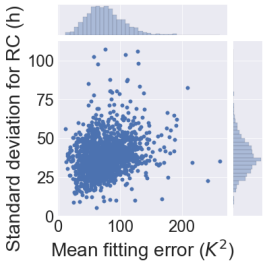
(a) Energy Balance (RC vs. cost)



(b) Energy Balance (RK vs. cost)

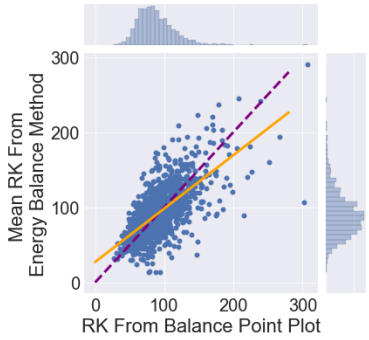


(c) Energy Balance (RK vs. RC)



(d) Decay Curves (RC vs. cost)

Figure 2.5: The performance of the energy balance model fitting and the decay curve model fitting. The scatter plots shows the correlation between the fitting costs and the standard deviation of *RC* and *RK* predictions. The histograms on the axes show the frequency distributions.



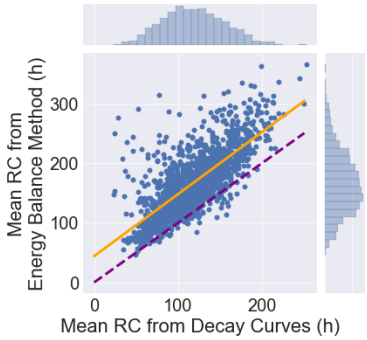


Figure 2.6: Comparison of the results for *RC* (energy balance and decay curve methods) and *RK* (energy balance and balance plot methods), with lines of perfect agreement (dashed) and actual fit (solid).

statistically similar. For *RK* the p-value is 0.25, indicating that the methods produce a population of values with a similar mean. For *RC*, on the other hand, the p-value is far below 0.05, which follows since the energy balance method systematically over predicts

Table 2.5: Parameters for removal of unreliable results.

Type	Model	Threshold	% of total
Standard error	Balance point	0.0015	43.93
r-value	Balance point	-0.75	41.22
RK Standard deviation	Energy balance	80	22.00
RC Standard deviation	Energy balance	125	35.67
Fitting cost	Energy balance	700	22.49
Intervals found	Energy balance	4	31.83

Note that these values are not mutual exclusive. The % of total represents the amount of outlier values of the given type.

when compared to the decay curve method.

The proportional differences between the methods for each building were examined (see Figure 2.7). This proportional difference was obtained by dividing the difference in result for each building by the result from the energy balance method. For example, a value of 20% for *RC* means that the energy balance method overestimated by 20% compared to the decay curve method. The median proportional difference for *RK* sits around zero, further indicating that the balance point method and the energy balance method produce a population with similar means. On the other hand, the energy balance method over-predicts *RC* by a median value of around 20%. This may be explained by intermittent internal gains or ventilation losses that are not taken account in equation 2.2 but are partially captured by *RC*.

Overall, we found that there is a strong positive statistical correlation between the three methods. The absolute values obtained for *RK* are similar, but the absolute values for *RC* vary significantly between methods. The statistical correlation indicates that these approaches may be viable in assessing the relative values for thermal characteristics of homes, even if the estimates do not represent the ground truth. If the end goal is to build a crude filter that can reasonably target potential retrofit candidates from a large dataset then the methods do not need to have a very high accuracy; rather, they should be internally

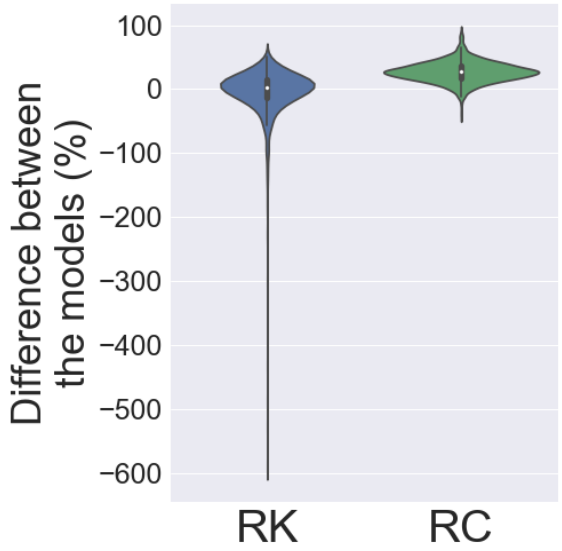


Figure 2.7: Distributions of the proportional difference between methods for a given building.

consistent. These approaches may therefore be viable in assessing the thermal characteristics of homes, and thus help with filtering for envelope retrofit. Additionally, having shown that the methods return results that have relative significance, if not absolute accuracy, we can conclude there is a huge range in the thermal quality of the buildings in this study.

2.3.3 Results obtained for RC and RK

The decay curve method predicts RC values that range from $23h$ to $252h$ with a mean of $119h$, while the full energy balance method predicts values from $55h$ to $365h$ with a mean of $170h$. For RK, the balance point method predicts values that range from $37°K$ to $306°K$ with a mean of $99°K$, while the full energy balance method predicts values that range from $13°K$ to $290°K$ with an average of $99°K$. The full distributions of both parameters for both methods are shown in Figure 2.6. It is notable that the RK distributions are tighter than those for RC.

As a speculative exercise, dividing the RC values obtained by very approximate values for C ranging from 10,000 to 20,000Wh/K (lightweight to heavyweight construction) and

multiplying by a surface area of 1,000m^2 gives area-averaged R values that range from 2 to 38 m^2K/W with the mean values equating to around 7 m^2K/W. Applying a similarly broad assumption of a heating power $K = 25,000W$ gives area-averaged R values that range from 0.1 to 14 m^2K/W with the mean values equating to around 3 m^2K/W. While there is clearly significant variation between the methods, these values are all within the realm of possibility.

The results for RC and RK obtained from these methods were compared with the building metadata. We hypothesised that there would be a strong correlation with the age and size of the building, but only weak correlations were found (see Figure 2.8). Further evaluation revealed that the correlations remain weak even for cases with high similarity between all three methods. This result is consistent with past studies; Tabatabaei et al. similarly did not find a strong correlation between the age of the home and the R value [97].

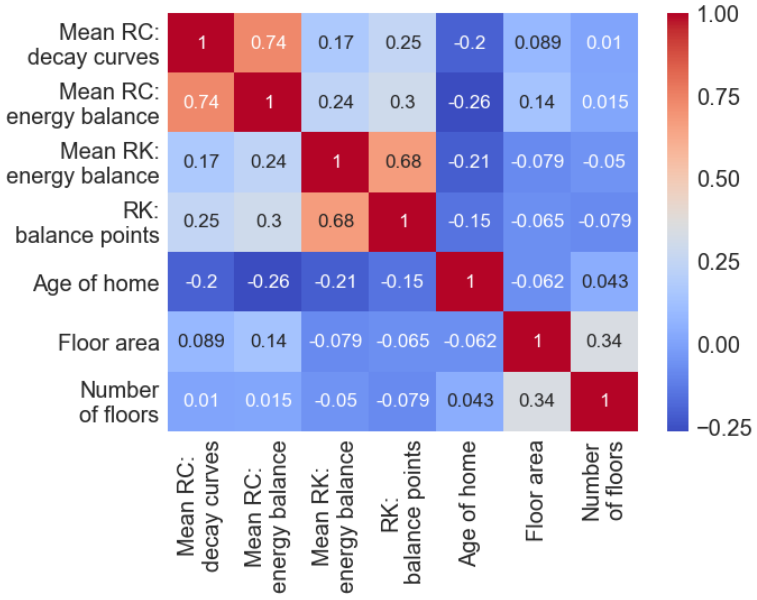


Figure 2.8: Heatmap showing the correlations between parameters determined in the analysis and building metadata.

2.4 Discussion

The results from the previous section were acquired by applying the methods described in section 2.2.2 and filtering out buildings that were not able to fit successfully. It is important to recognize the quality of the results is highly dependant on this outlier filtering. For example, a maximum value of 0.0015 is used as a threshold for the balance point method. If that value is raised to 0.0025 the population means for RK are no longer statistically similar, but less buildings are rejected from the final results. Clearly the outlier rejection results in a tradeoff between more statistically significant results and the amount of data that is retained. The chosen threshold values are up to the discretion of the user and will likely change depending on the use case.

Each of the proposed methods exhibits its own strengths and weaknesses which are summarized in Table 2.4. The balance point method is easy to implement and it can work with aggregated data sources, but this means that there may be important information that is not captured by the model. It follows that many buildings were rejected as outliers because the balance point plots do not have a strong statistical linear relationship. On the other hand, both the decay curve method and the energy balance method use detailed time series information that is more descriptive than aggregated values, so they should be able to better model building behaviour. The decay curve method requires a lot of filtering and assumes that the outdoor temperature is constant, but it returns stable results with few outliers. The energy balance method can be applied with less restrictions and it accounts for changes in outdoor temperature, but there were more outlier values than with the decay curve method.

Of the three, the decay curve method appears to be the most stable, based on the small proportion of outliers in the final results. Interestingly, though, its mean population is not statistically similar to the energy balance method for RC, although the population means are similar for RK from the balance point method and the energy balance method. This may

be because both use less filtering than the decay curve method and unintentionally capture extra heating and cooling behaviour from what is expected in the models.

We found that the energy balance method performed poorly for time periods where a disproportionate amount of internal losses or gains could not be captured by equation 2.2. Upon manual inspection of the time series data, we concluded that these periods seems to contain events such as windows and doors opening. Further research into this area could yield very interesting results.

One major limitation of all three methods is it is impossible to determine the parameters R, C, and K independently. If information on the power of the heating system for a household were available, R and C values could be isolated by determining RK via the balance point or X methods and dividing by the known K, then finding RC values using the decay curve or energy balance methods and dividing by R. We did not have sizing information in this dataset, but this is a potential area for future research.

In general, the presented methods should be useful in large scale retrofit analysis, but no method should be used independently. When using these methods, buildings should always be evaluated relative to one another. Past studies in this area commonly evaluate only a few buildings or only use a single method, so it is difficult to understand how useful they are for wide-scale retrofit analysis [35][97][96][99]. To help other researchers reproduce this work, the code is provided at https://gitlab.com/energyincities/besos-public/publications.

2.5 Conclusion and Future Work

The purpose of this study was to explore how big data may be used to estimate the thermal characteristics of homes. Naturally, real world data is messy, noisy and requires a lot of filtering and preprocessing to be useable. Even so, it was determined that by using gray box

models a reasonable estimate for relative values of *RC* and *RK* can be found, but absolute values are harder to determine. A high degree of accuracy is not required to filter retrofit candidates, so the three methods presented are likely sufficient for this purpose. Our methods differ from past studies in several ways: we use temperature data rather than energy loads, evaluate a large dataset and use granular time series data.

There are many interesting avenues of investigation still remaining. Better filtering could allow an hourly-resolution energy balance plot to give meaningful results. More investigation of the filtering trade-offs could improve the decay curve method. Resampling the data at a coarser time resolution could be beneficial for the energy balance method, since it will allow longer time periods to be assessed without compounding errors in the projection of the equation. More comprehensive cross-validation for all three methods could identify areas where they perform poorly or well. A dataset with more comprehensive metadata on building envelope and system parameters would allow the models to be fully validated against a known ground truth.

For the purpose of this study outlier values were rejected from the final results, but a detailed analysis should be done to better understand what causes a building to produce bad predictions for *RC* and *RK*. Research into this area could potentially result in data quality control or fault detection strategies.

With the help of more detailed weather data, the energy balance model could be expanded to consider solar gains via a solar susceptibility parameter. In conjunction with cooling data available from ecobee and a term for cooling power, it would be possible to apply the model to daytime and summer periods. If such a model proved more robust, it would be worthwhile to expand the studied area to climates with more moderate winter temperatures. Methodologically it would be particularly interesting to compare these gray-box methods and a numerically-calibrated white-box approach, for example by using an optimization algorithm to calibrate an EnergyPlus simulation model.

Finally, it would be relatively easy to commission detailed energy audits for a tiny sample of the buildings using traditional methods. These known datapoints would allow much greater accuracy in the methods used here, through the improvements in filtering, outlier identification, and model refinement.

This paper clearly demonstrates several distinct ways in which big data from existing sources can provide meaningful insights into the state of the building stock. The lessons learnt provide a valuable step towards understanding how big data may be used to derive the thermal characteristics of buildings. It explores the types of problems that can and cannot be addressed with existing datasets that do not include heating system power or ground-truth data for calibration. Hopefully this will serve as an incentive to policy-makers and analysts to deliver better sources of data so that the full potential of such methods can be realised.

Chapter 3

Targeting Buildings for Energy Retrofit Using Recurrent Neural Networks with Multivariate Time Series

3.1 Introduction

A growing body of research confirms that retrofitting residential buildings provides a net reduction in carbon and energy use, as well as monetary savings [26][69][63][103]. The findings of these studies are reflected in international policies regarding building retrofits [63]. The development of large-scale computational approaches to building performance analysis are essential to the success of such retrofitting programs. Modern techniques for building assessment often involve expensive in-situ measurements and on-site appraisal [72][38][12][8], but researchers have started investigating the use of big data to scale this process [97][99][35][50]. Supervised machine learning methods, however, are not typically applied to building retrofit analysis, in part because there is a lack of data with useful labels. Sensing technologies such as smart meters and thermostats are becoming increasingly ubiquitous, but they are most commonly used for time series forecasting, load profile

analysis or benchmarking, rather than prediction of particular building properties [74]. It is not clear what types of building characteristics can be predicted based only on time series measurement data.

With all of these considerations in mind, the contributions of this paper are threefold:

- The introduction of a deep learning approach that targets residential buildings for retrofit.
- A showcase of the types of building metadata that can be derived from multivariate time series data.
- Helping to overcome the data scarcity in the Civil Engineering domain by introducing a novel methodology for dataset generation.

To accomplish these objectives two case studies will be presented - heat pump classification and R-value prediction. Each of these cases focuses on a particular retrofitting strategy that will be discussed in more detail in the following section. The remainder of this paper includes a description of the deep learning methods and model architecture, preliminary findings and a discussion of next steps.

3.2 Methodology

3.2.1 Case Studies

Heat Pump Classification: Load reduction measures in building retrofits involve upgrading mechanical equipment such as appliances and HVAC systems [51]. Heat pumps are a particularly efficient HVAC technology, and government programs already exist to encourage system upgrades [1]. The ability to target homes that do not have heat pumps would be highly beneficial to these types of programs.

R-Value Prediction: Thermal resistance, R ($\frac{^\circ K}{W}$) is a material property that describes the effectiveness of insulation; the higher the R-value, the more effective the insulator. The area weighted average of R-values for all external surface provides a proxy for the quality of the building envelope. Envelope measures in building retrofits aim to increase the R-value by improving the constructions. An effective program should target buildings with relatively low values, but quantifying R is not trivial and the results can be difficult to experimentally validate [72].

In this paper we propose a novel approach for predicting R using whole building simulation software. In our approach, computational physical modelling is used to simulate building behaviour based on inputs such as geometry and construction definitions. Unlike typical building assessment methods which use measurement data to deduce quantities about a building, our method uses building simulations to generate synthetic time series data. We postulate that one could build a predictor for R by training a deep learning model on this synthetic data. The model could then be used predict the R-value for a real building from measured data. The work in this paper focuses on the creation of the synthetic dataset and the model training; future work will validate the use of this approach on real buildings.

3.2.2 Data

The dataset used for heating system classification is acquired from ecobee's Donate Your Data program[1]. This dataset consists of smart thermostat time series data measured at 5 minute increments as well as metadata describing household attributes. A detailed description of this dataset can be found at [49]. For the problem at hand indoor temperature, outdoor temperature and heating system runtimes are the input variables and presence of heat pump is the output variable. For the purpose of this study, only homes in Ontario and New York were considered. Of this subset there was a disproportionate number of homes

[1] https://www.ecobee.com/donateyourdata/

with heat pumps. The dataset size was therefore reduced further to stratify the presence of heat pump so there was an even split in both the test set and the training set. The resulting data had 602 homes in the training set and 182 homes in the test set.

To predict R, a multivariate time series dataset with 10 minute granularity and 966 data points was created: 773 in the training set and 193 in the test set. The input variables consist of indoor temperature, outdoor temperature and heating power. Though the creation of this dataset is a significant contribution of this work, a full explanation is reserved for the Appendix.

For each case study the sequence length was limited to 2000 consecutive time steps per building[2], and mean imputation was used to handle missing values.

3.2.3 Model Definition, Optimization and Training

Given that the data structure for both of the above use cases is multivariate time series, the Recurrent Neural Network (RNN) is a natural choice of architecture. Gated Recurrent Units (GRUs) and Long-Short Term Memory Units (LSTMs) are extensions to the RNN that help to overcome the vanishing gradient problem and make them more suitable for learning long-term dependencies [43][18]. Both GRU and LSTM would be suitable for the work presented in this paper, however GRU was chosen because it has been shown to occasionally outperform LSTM in terms of convergence time and generalization [17]. Future work should also consider LSTM, as well as other architectures such as 2-dimensional Convolutional Neural Networks.

The same model architecture and optimization algorithm was used for both case studies. The model consisted of 3 stacked GRU layers with 80 feature units in each hidden state. As proposed by Cooijmans et al., batch normalization was included on each of the hidden-to-hidden transitions [21]. Cyclical learning rates, introduced by Leslie N. Smith, were used

[2]2000 time steps equates to one week for heat pump classification and two weeks for regression over R-value.

for training [92]; heat pump classification used a minimum rate of 1e-3, while prediction of R used a minimum rate of 1e-2. A weight decay of 1e-2 was used.[3] Finally, the training loop for the former case study used binary cross entropy loss while the latter used mean squared error loss.

3.3 Results

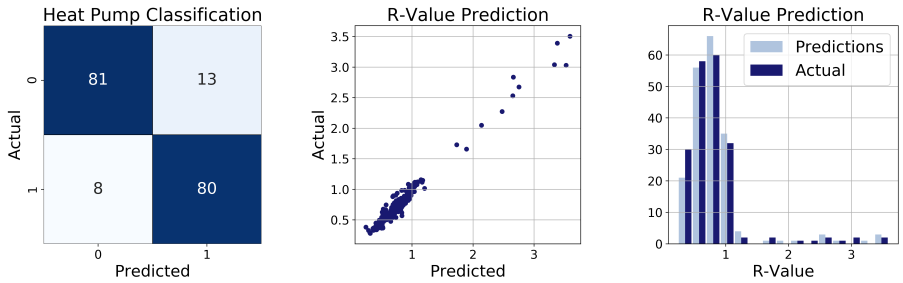


Figure 3.1: (a) The confusion matrix for heat pump classification. (b) Performance of R-value predictor. (c) Distribution of R-value predictions and actual values.

For heat pump classification, a validation accuracy of 0.87 was achieved on the test set, while the root mean squared error for prediction of R was 0.089 on the test set. The training for heat pump classification took 100 epochs while the training for prediction of R took 150. In both cases this is a relatively high level of performance with a relatively short training time.

A more comprehensive understanding of the results can be seen in Figure 1. The confusion matrix illustrates the precision-recall tradeoff in the heat pump classification problem, with a precision 0.86 and a recall of 0.91. The scatter plot shows the linear relationship between the predicted and actual values and the histogram represents the spread of values for R. The majority of values lie between zero and one[4]. With respect to this

[3]The values for weight decay were chosen according to the defaults in the fastai library [48]. The learning rates were chosen using a learning rate finder, also provided by fastai. Dropout was also tried but the accuracy suffered.

[4]All of the values greater than one are from a building model with the same initial definition whose values for R are quite different than the other building definitions

distribution, one can see that an RMSE of 0.089 is relatively low.

These findings should be considered preliminary; while they do indicate the usefulness of deep learning to building retrofit analysis, more work is required to improve accuracy and ensure generalizability.

3.4 Discussion & Conclusion

The ability to easily and accurately identify homes for retrofit is essential to inform international strategies for global energy and carbon reduction. Deep learning models in particular are affordable, scalable and reusable, and their successful application could prove invaluable in the building performance assessment industry. The findings in this paper are preliminary, but they show potential for the use of deep learning in targeted retrofit analysis. Future work should focus on continued data collection and model development in order to improve accuracy and ensure generalizability of results.

Chapter 4

Identifying Whole-Building Heat Loss Coefficient from Heterogeneous Sensor Data: An Empirical Survey of Gray and Black Box Approaches

4.1 Introduction

Digitization is transforming our understanding of a building from a passive, voiceless space into a constantly communicating, active service provider for healthy and sustainable living. At the core of this transformation is sensor data, which provides a continuous stream of information on indoor comfort conditions and energy performance. Worldwide, more than one billion smart metering devices will be installed by the end of 2020 [83], and large construction markets have or will have (e.g. Canada [42]) nationwide coverage. This provides viable source of information for building diagnostics and analysis, e.g. for identifying thermal properties which enable more effective energy retrofits [72], for deriving accurate building stock models to predict future energy behaviour [78], and for demand

response targeting [102].

Meanwhile, high-impact changes to the built environment are required to meet global emissions reduction targets. To this end, the ability to quantify building envelope performance on a massive scale is imperative [26] [63]. Traditional approaches that rely on walk-throughs, complex in-situ measurements or expert knowledge are not cost-effective or scalable. The development and deployment of reliable and efficient data-driven methods for building property characterization will therefore be integral in decarbonizing the building stock.

Two paradigms for identifying thermal properties from building data exist: gray and black box modelling.[1] The former estimate building properties via incremental updates in an optimization loop that reduces the error between the real time series data and data that is simulated by the physical model (see Figure 4.2). The latter use generalized models without domain knowledge and instead derive meaningful representations from large amounts of training data [57]. In the buildings domain, the quantity of research into data-driven modelling is astounding - for instance, in 2020 Hong et al. cite over 9,576 studies on machine learning in buildings (including both gray and black box methods) [46]. However, they note that the rate of industry adoption of these approaches is almost 0. Further, Roels [80] and Yilmaz et al. [106] assert that established, cost-effective and scalable methods for thermal performance characterization in buildings are lacking [15]. This is largely due to lack of (1) large scale labelled data and (2) model transferability, reliability and robustness [46].

Existing literature for building property characterization suffers from the problems identified above. High-fidelity labels are rare [27] so ground-truth is not typically used to evaluate the predictions. Rather, models are assessed according to the quality of the

[1] White box models are purely physics-based and therefore cannot identify thermal properties from data. In this work, white box models are used to generate the synthetic dataset.

calibrated model.[2] Novel techniques that estimate thermal properties from large amounts of sensor data are forced to use proxy-measurements, such as age or size of the buildings, for validation [9] [97] [36]. Generally, a lack of ground-truth assessment inhibits an honest appraisal of both new and existing methods.

It is also common that studies are conducted in isolation; a single approach is applied on an undisclosed and small subset of building data [75] [38] [88] [72] [41] [33] [11] [12] [27] [55]. This makes it difficult to compare approaches to one another and erodes industry trustability in upcoming methods. A few studies are beginning to benchmark model performance [79] [26] [87], but much work is still required. The existing studies do not consider robustness of approaches to heterogeneous building properties, the data and code is not open sourced so the works are difficult to extend, and the scope of the compared methods is limited.

More research is required to assess the efficacy of existing approaches for thermal property prediction from large scale, heterogenous building sensor data. Recently, black box approaches have shown promise in this domain [10], so they should also be benchmarked against existing work. This study therefore focuses on ground-truth assessment of gray and black box models for practical application cases such as large-scale retrofit analysis, building stock modelling and demand-response targeting. Whole-building heat loss coefficient (HLC_{wb}), which measures the rate at which heat is lost through the building envelope, was chosen for identification in this study due to its practical applicability and prevalence in the literature. The specific contributions of this work are as follows:

- A concise overview of the barriers that prevent the deployment of gray and black modelling paradigms in practice.
- Ground-truth benchmarking for seven models (including deep learning approaches

[2]The US Department of Energy, for example, outlines acceptable calibration tolerances for the monthly mean bias error (MBE) and the normalized error of variability (Cv(RMSE)) between the measured and predicted data [32].

that are novel to the domain) that estimate HLC_{wb} from whole-building, sub-hourly metered data[3], including a qualification of robustness towards heterogeneity in the building stock (i.e. climate, construction materials, air-infiltration and stochastic occupant behaviour). The analyzed methods include Energy Signatures (ES) ie. balance point plots, 1st and 2nd order resistance-capacitor networks (RC1 and RC2), and surrogate-based building energy simulation using genetic algorithms (GA-BES) and Bayesian optimization (B-BES), a recurrent neural network (RNN) and a residual convolutional neural network (CNN) (Figure 4.1).

- An open-source, extensible, synthetically generated dataset of 16,000 buildings with ground-truth labels[4]. All of the model code is also provided.

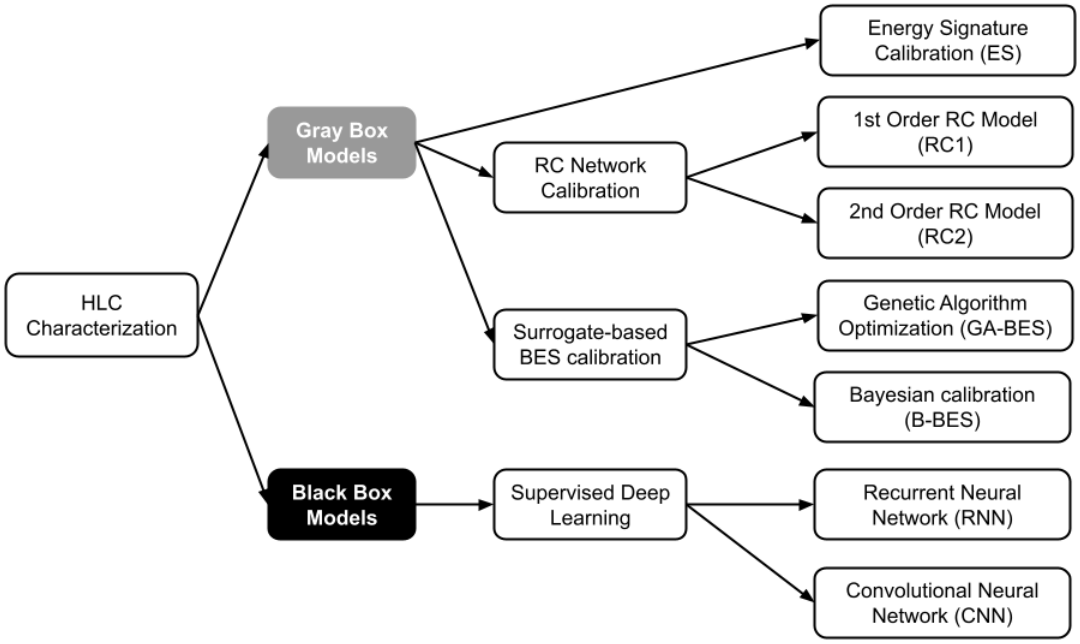


Figure 4.1: The investigated research paradigms and model implementations.

The following section provides a brief overview of thermal property characterization using gray and black box modelling approaches, with a particular focus on practical application

[3]Measurement variables include indoor and outdoor temperature, solar radiation and heating system power.
[4]https://gitlab.com/energyincities/bp-benchmarker

barriers.

4.2 Background

The bandwidth of thermal characteristics that can be extracted from buildings is large and includes both categorical information, like the installed heating system type [104][13], and quantitative information, like heating system efficiency [20] or the HLC_{wb}. In this paper we focus on the prediction of HLC_{wb} from sub-hourly measurement data using both gray and black box approaches. Figure 4.2 illustrates the workflows for each of these two modelling paradigms, while Table 4.1 and the remainder of this section describe the key differences between them. Within each paradigm the model complexities and optimization process differ, as well as the shape of the input data. These details are withheld until see Sections 4.3.1 and 4.3.2, respectively.

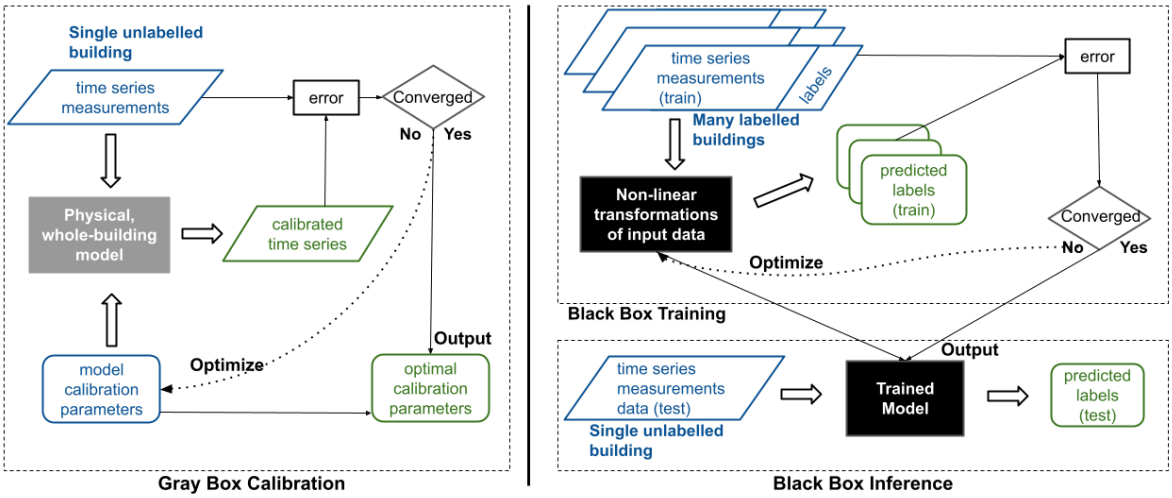


Figure 4.2: Flow diagrams describing the calibration process for gray box models and the training and inference procedures for black box models. Blue represents model inputs and green represents model outputs.

In the literature, the dominant approach for retrieving HLC_{wb} is using physics-based whole building models, whose parameters are calibrated using measurement data for a single

building at a time. These gray box models are commonly applied in the context of building control [64] [6] [76], but they are also used specifically to find building characteristics [41] [12] [73] [75] [38]. The complexity of the underlying physical model ranges from the 1 or 2 parameter ES models [35], to RC network models of various orders of complexity [11], and to complex BES model calibration approaches [20]. RC network and BES model calibration are highly dependent on the representativeness of the building model; RC network modellers use residual analysis to determine the appropriate model order for a given building [11] [27] [26] [79], and segmenting a building stock into groups of similar buildings (archetype classification) and deriving a suitable building energy model (architecture characterization) are decisive steps in BES calibration processes [93][56][52]. In practice, the use of gray box methods on large data depends on the scalability of the model identification process, which is currently expensive and requires expert intervention [77].

In contrast to gray box calibration, black box prediction (commonly with neural networks) is desirable because the models require no prior knowledge of system dynamics. They create a domain-agnostic mapping from inputs to the quantity of interest [57] which is then used to predict the quantity from unseen examples. However, supervised deep learning traditionally requires huge amounts of data[5] with high-fidelity labels which are not always available in this domain. Compared with calibration, applications of supervised deep learning to thermal property estimation, and HLC_{wb} identification in particular, are therefore fairly limited [62] [82] [90] [10].

Even though datasets with high-fidelity HLC_{wb} labels are rare, it was chosen for study because (1) of its relevance for building decarbonization, (2) it is estimated natively by ES and RC model calibration - both of which are highly popular in the literature - but their robustness has not been well verified and (3) state-of-the-art machine learning research helps reduce the stringency of data requirements. The application barriers presented above are

[5]Models that classify images with 95% accuracy are trained on more than 14 million labelled images [28].

Gray Box	Black Box
Depend on a representative building model.	Model is generic and agnostic to the domain.
Calibrated on a single building at a time.	Trained on many buildings. Predict on a single building at a time.*
No labels required.	Require high-fidelity labels for training.
Well researched for HLC_{wd} identification.	Not well researched for HLC_{wd} identification.
Robustness not established.	

Table 4.1: Key, practical differences between the gray and black box paradigms. See [20] for further information. *Although black box methods predict on a single building at a time, the prediction time is very fast, especially compared with building-by-building calibration.

therefore surmountable, but we propose here that reliable ground-truth performance must be asserted first. The focus of this paper is therefore on benchmarking a broad set of methods, rather than diving deeply into a single approach. The methodology and results section in this paper therefore focus on this benchmarking.

4.3 Methodology

The flow diagram in Figure 4.3 illustrates the methodology of this work, and this section provides the technical information of the seven models studied including: the model identification procedures for the RC network calibration and the BES surrogate training[6], the data requirements for each of the models (Section 4.3.2), the performance metrics that are used to assess the models (Section 4.3.3), and the synthetic dataset design and creation (Section 4.3.4).

4.3.1 Models

Figure 4.2 illustrates that, within each of the two modelling paradigms, approaches differ according to (1) the gray or black box model formulation and (2) the optimization process.

[6]The building model used for BES calibration is identical to the model that generated the measurement data, so no model identification is required.

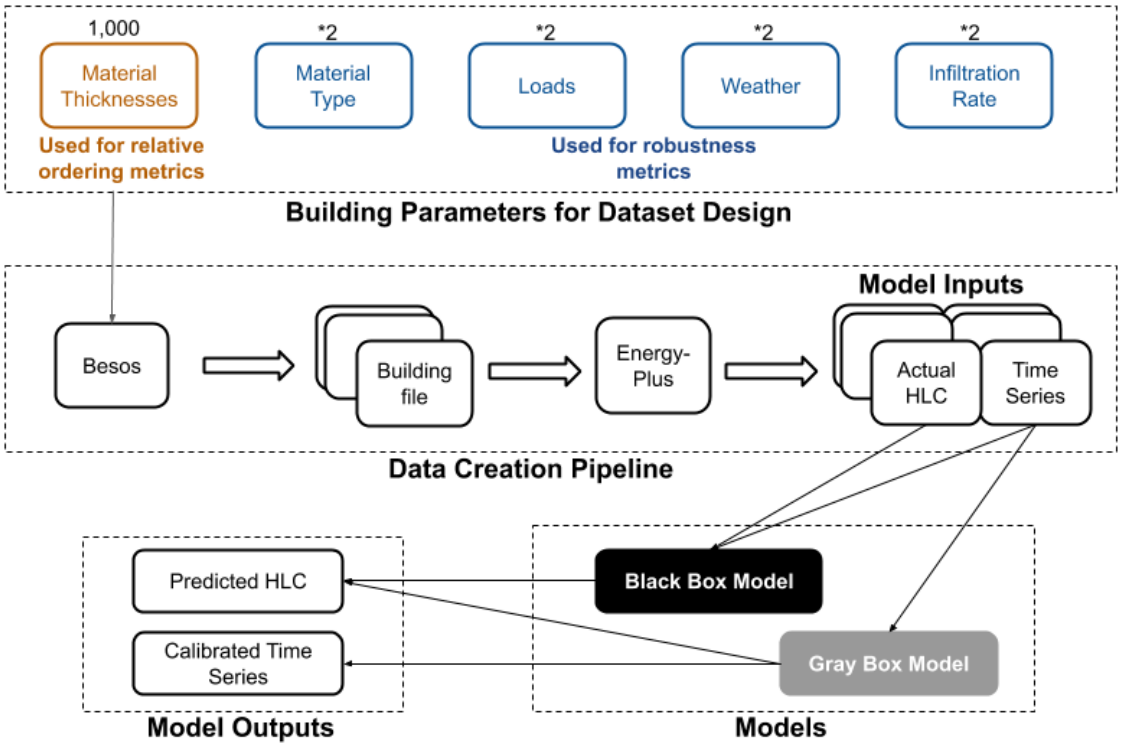


Figure 4.3: Flow diagram describing the methodology presented in this section, including the dataset design parameters, the data creation pipeline and the inputs and outputs of the models. Note that the 1,000 material thicknesses are different for the wooden and concrete buildings (B.3).

Below, these are described for each of the seven studied approaches. The RC model section includes the residual analysis of the selected models to show that the appropriate model order was used in this work. The BES section includes information on the surrogate modelling process. The number of free parameters in each model are listed in Table 4.2.

Energy Signature (ES) Calibration

(1) Model: The underlying model for ES calibration is a basic reformulation of the whole-building energy balance (B.1). When outdoor temperatures are lower than a certain point, known as the 'balance point', the heat exchange in the building is described with the linear[7]

[7]Recent advances in the literature also use non-linear formulations of this approach [77]. Future work should investigate the ground truth performance of these novel and promising methods, but to manage the

equation:

$$\dot{Q}_{h,d}(\overline{T}_{ext}) = HLC_{wb}(\overline{T}_{in} - \overline{T}_{ext}) + \dot{Q}_{solar} - \dot{Q}_{baseline} \tag{4.1}$$

where $\dot{Q}_h$ is the heating system supply, $\overline{T}_{ext}$ is the external temperature, $\overline{T}_{in}$ is the internal temperature, $\dot{Q}_{solar}$[8] is the solar heat gain and $\dot{Q}_{baseline}$ is the baseline heat gain. Typically, $\overline{T}_{ext}$ and $\overline{T}_{in}$ are aggregated daily.

(2) Optimization: HLC_{wb} is found using linear regression to find line-of-best-fit for to the measurement data, which is described by Equation 4.1. This is a standard approach in building energy modelling [39] [35] [36] [73] [24] [14].

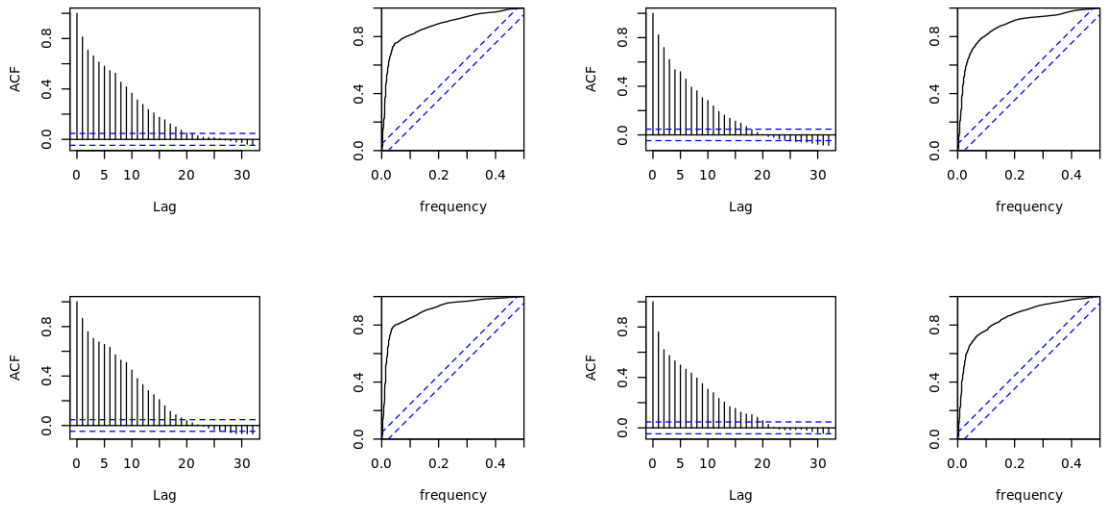
Resistance-capacitor (RC) network calibration

(1) Models: In this popular approach, a building is modelled using an RC network that can be defined at differing orders of complexity, from simple networks with a single lumped capacitance to complex, multi-order systems [11] [65]. RC order 1 (RC1), i.e. Ti and RC order 2 (RC2), i.e. TiTe from [11] were chosen for review in this work. These two model orders were selected according to the criteria described below. They are described by a set of first-order stochastic differential equations (see [11]) whose parameters are calibrated until the model outputs match the observed indoor temperature.

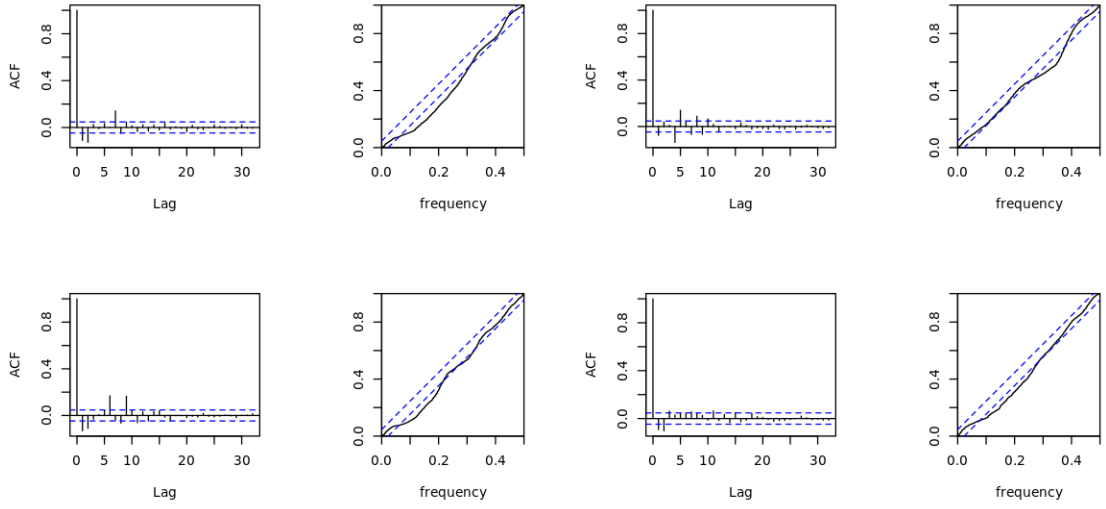
To apply RC models, the appropriate model order must be selected. [11] describe an iterative forward selection procedure using likelihood ratio testing. They suggest choosing a model above which all extensions have likelihood ratio p-value above a specified limit (e.g. 0.05). Due to large computational runtimes, the full iterative procedure is not feasible for large data. Instead, we validate RC2 by assuring that the p-value of the t-tests is below 0.05

scope of this work we only focus on the linear formulation.

[8] In this work $\dot{Q}_{solar}$ is included and the equation is fit with a multiple linear regression. It is often the case that solar is not included in the ES model. The number of free parameters in the ES model scales linearly with the amount of measurement terms that are included.



(a) RC1 residuals do not have the required white noise properties and that the residuals are not independent of the inputs.



(b) RC2 residuals exhibit the required properties.

Figure 4.4: The auto-correlation function and cummulated periodogram of the residuals indicate whether the selected RC network adequately models the physical building behaviour, as suggested by [11].

for the estimated state parameters, and by visually evaluating the autocorrelation function (ACF) and the cumulative periodogram (CP) of the residuals to ascertain whether they

have the appropriate white-noise properties (Figure 4.4[9]). For RC2, any models whose state variable estimates have p-values over 0.05 are filtered out of the results. For RC1, no model validation was performed, but the results are still included to demonstrate the relative performance of the most simple model with the lowest number of parameters. The limitations of this model selection approach are discussed further in Section 4.5.1.

(2) Optimization: Statistical maximum likelihood estimation is applied to estimate the unknown parameters in the model. Specifically, a Kalman filter is used to estimate the likelihood function, and an optimization algorithm is used to find the set of parameters that maximize the likelihood function. Refer to [11] and [65] for more detail.

Surrogate-based BES calibration

(1) Model: Calibration of BES models (here, EnergyPlus) can be computationally expensive, so machine learning based surrogate models are used inplace of the energy simulator [93] [41] [67]. A surrogate model approximates the BES model by learning from a few simulation runs to estimate the effect of changes in parameter values (surrogate model inputs) to changes in simulation outcomes (surrogate model outputs) [105].

(1) Optimization: Two optimization procedures to calibrate the BES model parameter, here the heat-loss coefficient, are used in this work: a genetic algorithm (GA-BES) and Bayesian optimization (B-BES). These are each discussed in turn.

In building design, black box optimization approaches such as genetic algorithms (GAs) are often applied. They can also be used to minimize the summed difference of simulated daily heating demand and measured daily heating demand [25]. Here, the NSGA-II optimization algorithm is used (population size = 200, offspring size = 100, iterations

[9]For brevity only 4 example buildings are included in Figure 4.4. The plots represent randomly selected buildings from the experimental condition for which RC2 produced the widest spread of HLC_{wb} estimates (see the Section 4.4.2 for more details on the performance per experimental condition). This was done to show that even in the worst performing case RC2 has residuals with the required properties. Our analysis showed similar behaviour across all buildings and experimental conditions.

= 3000). The approach is similar to [71], but uses higher frequency data (daily instead of monthly).

Following Bayes' theorem, a posterior for the unknown building parameters, i.e. a probability density function approximation of the calibration parameters, can be inferred using the difference between the measurements and simulated model outputs, and a prior probability for the unknown building parameters [41][29]. [10] Markov-Chain Monte Carlo (MCMC) sampling, here the Metropolis-Hastings (MH) algorithm, is used to approximate the posterior. That MCMC sampling process requires thousands of simulation runs and motivates the use of surrogate models. Commonly a Gaussian Process surrogate model is used (e.g. [41]). For the using of non-GP surrogate models, we follow the approach found in [68]. [11]

Supervised deep learning

(1) Models: Neural networks (NNs) are non-linear transformations of input data which are determined by thousands of trained parameters. Two NN architectures are implemented in this work: recurrent neural networks with gated recurrent units (RNNs) [16] and residual convolutional neural networks (CNNs) [40]. RNNs account for temporal input structure and are therefore a natural choice of architecture for time series data. CNNs are also used because they have exhibited state-of-the-art performance on various tasks, including time series prediction [31] [57].

(2) Optimization: NNs are typically trained using stochastic gradient-based optimization. Here, Adam optimization is used [53]. This is an extension to vanilla gradient-based optimization, where the learning rate is adapted as the model trains.

[10] We specified a uniform prior distribution for each parameter bound by the maximum and minimum heat loss coefficient observed in the data.

[11] The likelihood equals the sum-of-squared errors between measurements and BES time series outputs. This assumes identically distributed errors with zero mean and constant variance σ^2, see [68].

4.3.2 Model Inputs

Model	Input Variables	Granularity	Period	# Buildings	# Free Params	Target Variable
ES	outdoor temp. indoor temp. heating power solar gain	daily	1 year	3200	3	heating power
RC1 RC2	outdoor temp. indoor temp. heating power solar gain	5 minutely	Jan. 1st-7th	3200	6 10	indoor temp
GA-BES B-BES	heating power*	daily	1 month	12,800 (train) 3,200 (test)	1	heating power
RNN CNN	outdoor temp. indoor temp. heating power solar gain HLC_{wb} (*label*)	5 minutely	Jan. 1st-7th	12,800 (train) 3,200 (test)	> 1000 > 1000	HLC_{wb} *(required for training only)*

Table 4.2: Data requirements for each method and the BES-surrogate. *The weather file (here in the EnergyPlus format, .epw) containing the historical weather on building site is required for running the simulations to train the surrogate model, but not for calibration. The collection of the weather file is assumed to be perfect and not further addressed for this study.

Table 4.2 summarizes the inputs for each of modelling approaches described above. The data inputs were selected according to results of previous studies [79] [11] and empirical tests. The RC models were tested several datetime scenarios: in January and July, with 7 and 14 days worth of data. They performed the best on one weeks worth of data in January, so that is the period that is used in this work. ES was tested with 24 hour, 48 hour and 72 hour aggregates. There was no significant difference between the results, so daily aggregates (the most popular in the reviewed literature) are used.

The dataset size requirements are different between the methods, as described in Figure 4.2. Calibration is performed on a building-by-building basis, so the validation dataset can be of any size. A size of 3,200[12] was chosen to manage runtime while still producing

[12] 200 buildings in each of the 16 experimental conditions described in Section 4.3.4.

statistically significant results. Black box methods require both a training and a validation set. The BES surrogate model is trained NN so it also requires a large training and test set (see Section 4.3.1 for more information).

4.3.3 Ground-Truth Performance Metrics

Metric	Measure	Worst	Best	Description
Relative ordering	slope	0*	1	This is the marginal effect, which tells us how much the predicted HLC_{wb} changes when the actual HLC_{wb} changes, when all other building properties are held equal.
Relative ordering	$R^2 - score$	0	1	Indicates how much of the variability in the predicted HLC_{wb} values is attributed to the actual HLC_{wb} value.
Robustness	error distribution	significant difference	identical	Error distributions must be similar regardless of extraneous building properties. Models whose error distributions are shifted for certain conditions systematically over/under predict.

Table 4.3: The metrics that are used to determine (1) whether the models correctly order buildings by HLC, and (2) whether the models are robust to extraneous building properties. (1) is determined by performing regression analysis for buildings that differ by only HLC, but all else is held equal. (2) is determined by evaluating the difference in error distributions for heterogeneous buildings. *The slope can also be less than 0 or greater than 1.

A high degree of accuracy is not required for building stock modelling or to filter retrofit candidates; it is most important for these cases that (1) the relative relationship of the buildings is captured by the models, and (2) that the models are robust, that is, that they do not systematically over or under predict when exposed to particular conditions. To test these criteria, two cases can be considered: homogenous buildings that differ only by HLC_{wb} and heterogeneous buildings that differ according to extraneous properties aside from HLC_{wb} (ie. climate, thermal mass, air-infiltration rate and stochastic occupant behaviour, as described in the section below). The descriptive statistics that capture these criteria are highlighted in Table 4.3.

4.3.4 Synthetic Dataset

Date Creation Pipeline

The dataset is designed to test the aforementioned performance metrics. It is generated by running parametric simulations using BESOS [30] and EnergyPlus [22]; a process similar to that in [10]. The Building and Energy Simulation, Optimization and Surrogate-modelling (BESOS) platform enables quasi-random latin-hypercube-sampling of building design parameters [34]. These parameter combinations are fed as input to the building simulation software EnergyPlus, Version 9.2.0, which outputs a myriad of variables describing the detailed temporal behaviour of a building over the course of a simulation. Some of these represent time-series variables that can be measured with sensors in real buildings[13], and others include detailed information on the building's material properties. HLC_{wb} was calculated from the latter (B.2) A set of relevant time series variables and the computed HLC_{wb} values were stored to form the final, labelled dataset. The distribution of the HLC_{wb} values in the final dataset are displayed in Figure 4.5.

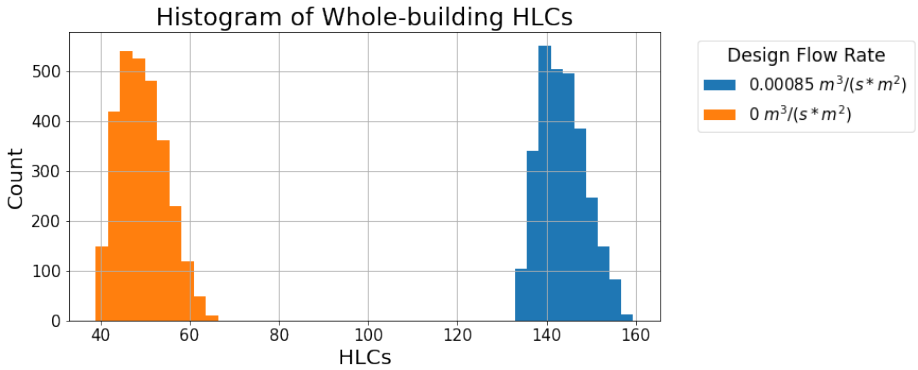


Figure 4.5: Histogram for the whole-building HLC values in the generated dataset.

[13] Such as external temperature, internal temperature, heating system power and solar gains

Dataset Design

To generate the building dataset, two baseline building models - ie.e 5mx5mx3m = 75 m3 box with one zone, four 4mx1.5m = 6m2 windows and no unconditioned spaces - were defined: one wooden building and one concrete building. For each of the wooden and concrete building baselines, the material thicknesses were varied to create 1000 buildings with distinct HLC_{wb} values. The thickness ranges for each of the materials were defined according to engineering standards and randomly sampled for each new building (B.3). Each of these sets of buildings was then simulated with annual weather data from two different climates (Victoria, CA and Chicago, USA), with and without air-infiltration (maximum flow per exterior surface area of 0 and 0.00085 $m^3/s * m^2$), with and without equipment and occupancy loads[14], for a total of $1000 * 2 * 2 * 2 * 2 = 16,000$ simulated buildings; 1,000 buildings differing only by HLC_{wb} for each of the 16 experimental conditions described above (Figure 4.3).

Additional modelling assumptions are listed below:

- The floors are adiabatic. Ground heat loss effects are difficult to simulate and therefore, neglected for now [79].
- No mechanical systems were modelled, EnergyPlus ideal air loads were used instead.
- Constant setpoint schedules were employed across all cases.
- Complex airflow networks and ventilation were ignored.
- Infiltration cases were modelled according to the DOE-2 standard by modifying the Field: Flow per Exterior Surface Area on the Zone Infiltration:DesignFlowRate object.[15]

[14]The stochastic equipment and occupancy loads were generated with the richardsonpy library from https://github.com/RWTH-EBC/richardsonpy

[15]https://bigladdersoftware.com/epx/docs/9-2/input-output-reference/group-airflow.html#zoneinfiltrationdesignflowrate

4.4 Results

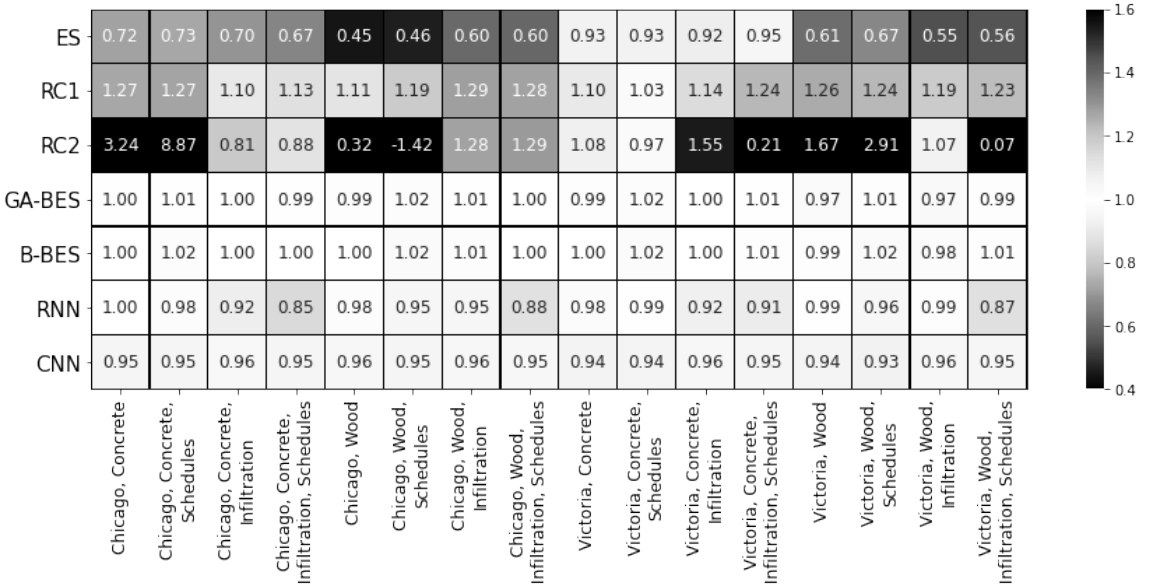
This section presents the performance of the models according to their ability to capture the relative ordering of HLC_{wb} and their robustness to extraneous building properties.

4.4.1 Relative Ordering

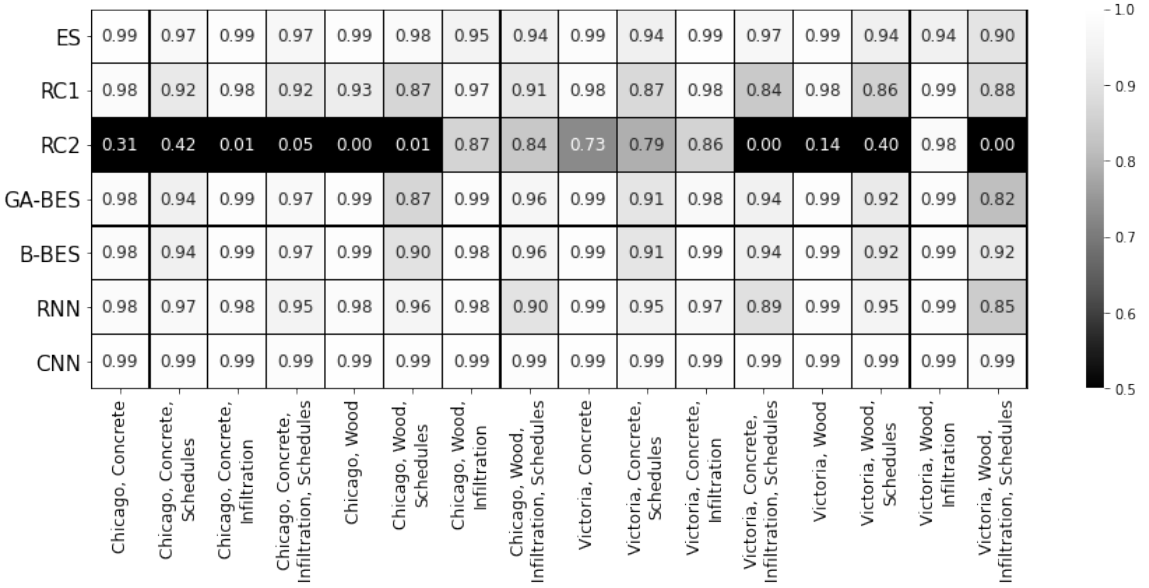
Figure 4.6a shows the instantaneous effect size of the actual HLC_{wb} on the predicted HLC_{wb} values. A slope of 0 indicates that the change in actual HLC_{wb} value has no effect on the model and a slope greater than 1 indicates that the model is biased towards buildings with higher or lower HLC_{wb} values. Naturally, a slope of 1 indicates perfect model performance with respect to marginal effect of HLC_{wb} (see Table 4.3 for a brief description of marginal effect).

The BES calibration approaches have a marginal effect on the predicted values that are closest to 1 (i.e. a slope close to 1), with Bayes calibration slighting outperforming GA calibration. The deep learning approaches perform best after the BES calibration, and the CNN outperforms the RNN. ES finds slopes less than 1, indicating that the actual HLC_{wb} has a low effect on this model. RC2 performs the worst, but even so this method achieves a slope of 1 ± 0.1 for 6/16 cases. RC1 tends to find slopes above 1, indicating that buildings with high or low HLC_{wb} values might have a disproportionate effect on the estimated HLC_{wb}.

The prediction variabilities (i.e. R2-scores) for each method are presented in Figure 4.6b. All of the methods aside from RC2 achieve a score > 0.8 for every experimental condition, which indicates that most of the variability in the predictions is described by changes to the actual HLC_{wb} value. The CNN consistently achieves the highest R2-score, while RC2 performs the worst by far. It is especially surprising that RC1 outperforms RC2, because RC2 had a better model validation score (see Section 4.3.1). Generally, the performance within each model is the worst for the cases with stochastic schedules; of the tested building



(a) Slope: The marginal effect of the actual on the predicted HLC values. 1 is perfect, 0 indicates that the actual HLC values have no effect on the predicted values. Values greater than 1 indicates that the model is biased towards buildings with higher or lower HLC values.



(b) $R^2 - score$: The amount of variability in the predictions that is attributed to the variability in the actual HLC values. A perfect score of 1 means that 100% of the prediction variability is due to actual HLC values, and 0 means the opposite.

Figure 4.6: Metrics that describe the ability of a model to find the correct relative orderings for building HLCs when all other building properties are held equal.

properties, stochastic loads cause the largest variability in the prediction results.

Based on the results above, all of the models except RC2 are able to correctly identify the relative ordering of building HLC_{wb}, while all else is held equal. This indicates that all of the models are suitable for application on a homogenous building stock, while also providing further validation that the models are working as expected.

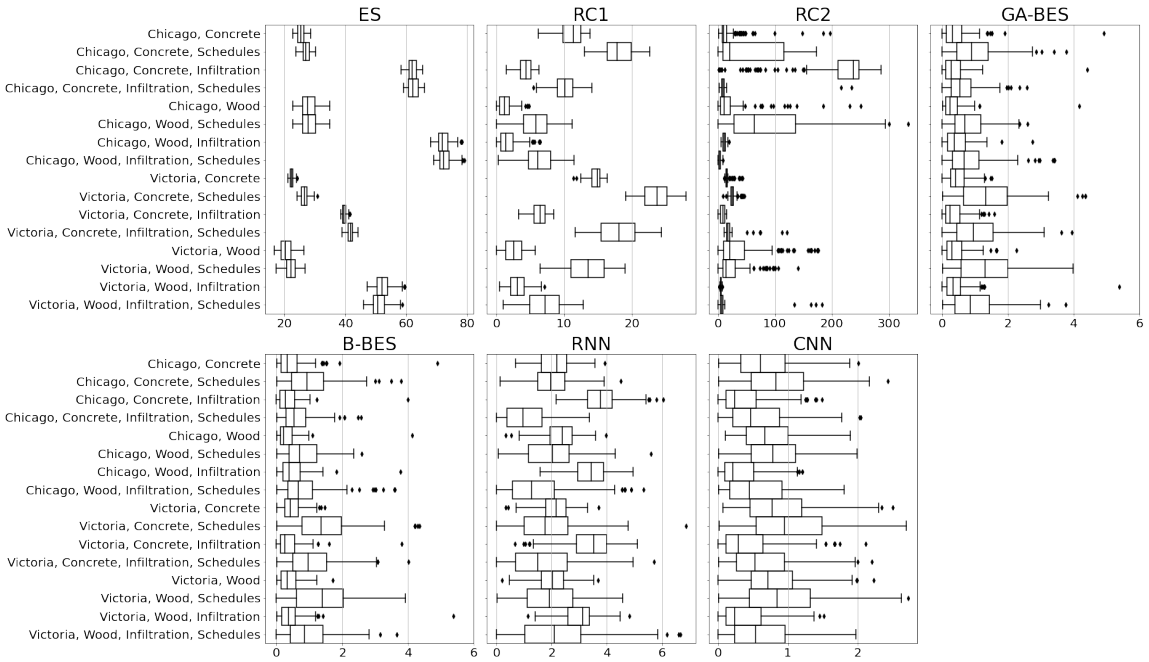
4.4.2 Robustness

Figure 4.7 uses boxplots to show their error distributions, and figure 4.7 provides a numerical summary of the mean absolute errors (MAEs) for each method within each experimental condition. A statistically significant difference ($p<0.05$) between the error distributions for the different building properties was found for all of the evaluated methods, using Wilcoxon signed-rank[16] tests between the distributions for each of the binary experimental conditions (for example, all wooden against all concrete buildings). In other words heterogenous building properties cause statistically significant, systematic bias in all of the models. The size of the difference between the error distributions, however, varies significantly between the methods, and is not always practically significant. In this section we will therefore analyze the error distributions from each model individually. This will also highlight which of the studied building properties have the most significant effect on the modelling results. For brevity, only the most important features of the data are discussed; the interested reader is encouraged to analyze the results further.

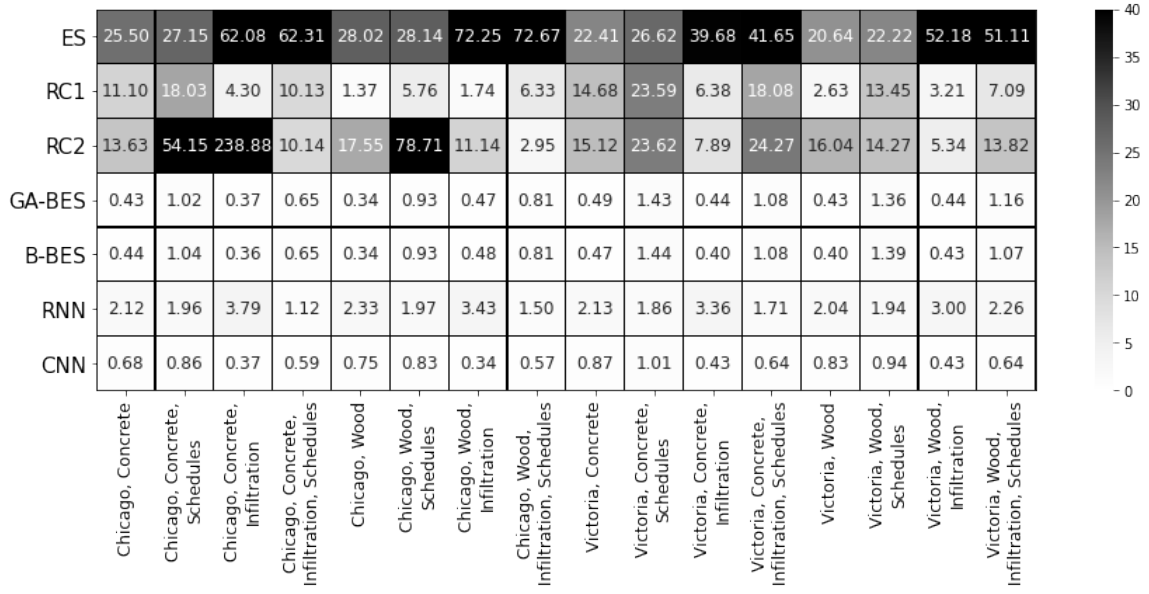
- *ES*: For this method, the buildings with infiltration result in a much higher MAE than the buildings without infiltration. This is a practically significant, systematic bias in the modelling results.

- *RC1*: There is a fairly large difference in the error distributions between the cases for

[16]This is a non-parametric statistical test that is used to compare paired samples that are not normally distributed.



(a) Error Distributions within each experimental condition.



(b) Mean absolute errors (MAEs) of the above distributions.

Figure 4.7: Differences between error distributions capture robustness of the models to climate, stochastic schedules, infiltration and construction material.

this model. The mean and spread of the errors are systematically higher for cases with stochastic occupancy. Moreover, the mean of the distributions for wooden buildings

are lower than those for the concrete buildings. Overall, this model is not robust to the tested extraneous properties.

- *RC2*: This method has clearly poor performance and cannot be considered robust.
- *GA-BES*: Like RC1, the errors for the cases with schedules are larger and more variant for this model. The largest MAE is 1.4 and the lowest is 0.34. In a practical scenario this is likely insignificant, but it is up to the modeller to decide. This approach produces unexpected outliers with poor performance.
- *B-BES*: The results for Bayesian Calibration are very similar to those for GA Calibration.
- *RNN*: Unlike the other methods, the RNN exhibits low errors for buildings with occupancy schedules. Overall, it tends to perform most poorly in the infiltration cases without schedules. Compared with the surrogate-based BES calibration approaches, the RNN has larger differences between the distributions, but these differences are still much smaller than those for the gray box calibration approaches. Again, it is up to the practitioner to decide whether the differences in the error distributions is significant in practice.
- *CNN*: For the CNN the MAE is less than or equal to 1 in every case, and there are few outliers with high errors. The greatest differences in error distributions are caused by infiltration and stochastic schedules, but these differences are likely insignificant in practice

4.5 Discussion

In this work we generated a synthetic data set which offers an experimental environment to test the methods' robustness towards four factors that possibly confound meter-data

based thermal building characterization accuracy. The discussion starts by addressing the assumptions and limitations of our approach. The experimental results will then be analyzed with these constraints in mind. Finally we propose promising directions for future work, where we set the focus on young machine learning based paradigms like BES calibration and deep learning.

4.5.1 Assumptions and Limitations

Assumptions and limitations in this work arise both from the modelling approaches and from the use of synthetic data which will be discussed independently. Each limitation can be overcome in future and presents an opportunity for further research, which we present at the end of this section.

Modelling Assumptions

The constraints of the experimental design in this work bias the results in favour of the BES calibration and the deep learning approaches. For the former, the BES model parameters that are not calibrated are assumed to be perfectly known (for example, the building geometry or the building environment) and the BES calibration model is the same one that was used to generate the synthetic data. In reality, the BES simulations used for training the surrogate model will not perfectly represent the buildings to be calibrated. Future work should focus on evaluating performance depreciation from imperfect BES models. For the latter, the models are trained across all buildings and are overfit to this dataset. More work is required to study how well they generalize to more complicated data.

As mentioned in the methodology section, the model identification in this work is limited compared to the full forward selection procedure suggested in [11]. Regardless, the residual analysis of the RC2 models in this work was promising and suggested that they should outperform RC1. Future studies should consider expanding the scope of this work to include

more robust model selection, however, it is important to note that large data applications are constrained by practical limitations such as computational runtime, which limit the feasibility of finding and deploying higher order models.

Limitations of the Dataset

We used a synthetic data set to conduct controlled experiments on the robustness of building characterization methods. It allows for the estimation of model robustness to the four considered impact factors, but the actual errors will be higher in the case of real buildings. This has multiple reasons including that the heterogeneity of the synthetic building stock is small, eg. only one geometry was considered, that micro-climates were ignored, and that all time series measurements were assumed to be perfect. Future work should consider addressing these points to develop more realistic synthetic datasets.

4.5.2 Summary and Analysis of Results

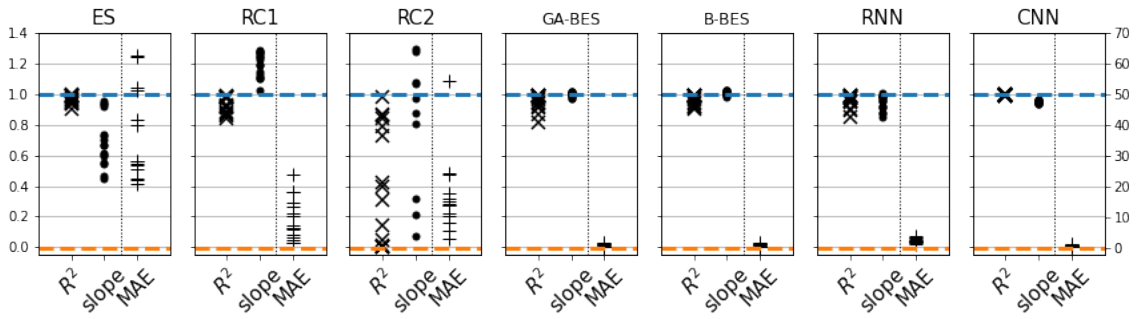


Figure 4.8: Summary of the metrics for relative ordering (R^2 and slope) and robustness (MAE). Some of the results were outside of the axis in the plots but they were excluded for visibility.

Figure 4.8 presents a precise graphical summary of the results. This a quick visual reference of the same points that were discussed in the results sections.

For buildings that only differ by HLC_{wb} (homogeneous building stock, i.e. the buildings within each experimental condition), the relative ordering of the buildings was captured by

all of the models, except the RC2 model[17]. This means that for use cases where the building stock is homogenous, that is the buildings only differ by the envelope, but all else is held equal, any of the models (aside from RC2) can be used. However, none of ES, RC1 or RC2 exhibit robustness to heterogeneous building properties (see Section 4.3.4), which is likely the more common use case in practice.

It is particularly interesting to note that RC1 outperformed RC2 in terms of ground-truth performance, even though RC2 performed better in terms of model validation. This is likely because RC2 is overparameterized compared to RC1; the addition of more model parameters will inevitably result in lower calibration error (see 4.3.1) because, statistically there is less bias in the model, but this does not mean that the building characteristics estimates approach ground-truth. This result strongly suggests that calibration error is not a sufficient metric for thermal property identification for use cases such as retrofit analysis; ground-truth performance evaluation on synthetic data should be a focus in future studies.[18]. Further, the results highlight the sensitivity of well-established methods towards building material choice, air infiltration, stochastic occupant behaviour, and climate.

BES calibration and the black box methods, on the other hand, all perform reasonably well in terms of both model validation and comparison to ground truth, providing a first indication that they may become a key element in data-driven retrofits, stock modelling and demand-response management. The following section therefore suggests future research directions that should be explored to overcome the barriers to application (Section 4.2) of these approaches.

[17]Even though RC2 performs the worst by far; for many cases its R2 -score is close to 0. Still, in some cases (for example the wooden building with infiltration in Victoria) the method performs well. This shows that a single case study might yield the method to be reliable, even if this is not the case in general. Literature tends to run case studies that validate methods on only a single building without varying properties or climatic conditions; the result here provides strong evidence that this is not sufficient.

[18]It is worth emphasizing here that this result does not indicate that calibration approaches should not be used. BES calibration, for instance, was able to identify HLC_{wb} with high accuracy and robustness. Based on the RC results, however, we hypothesize that the performance of the BES calibration will depreciate as more parameters are added. This should be explored in future work.

4.5.3 Future Work

Following from the above discussion, three broad categories for future work are suggested: (1) improving the synthetic dataset, (2) overcoming application barriers and (3) extending this study to benchmark more models against more thermal properties. Suggestions to improve the dataset were already discussed in Section 4.5.1, so this section will focus on the remaining two points.

Based on the current study, we propose that overcoming barriers to application for BES calibration and supervised deep learning should become a primary focus of future work. For BES calibration, future work should integrate model selection and archetype identification into the framework presented in this paper to determine how closely the underlying building simulation model must represent the real building. Regarding the black box models, transfer learning [98] and self supervised learning [70] are state-of-the-art techniques in the machine learning domain that reduce data requirements while maintaining high prediction accuracy.

We encourage researchers to expand on this work to provide more comprehensive benchmarking and to support innovation in this domain. The poor performance of the RC models should be further verified and novel modelling approaches should undergo standardized, ground truth benchmarking. Finally, the results of this study illustrate the challenges of identifying HLC_{wb} from large, heterogeneous sensor data. We therefore highly suggest that other - more simple - approaches to building performance characterization are explored and validated using the approach in this work.

4.6 Conclusions

The goal of this paper was to evaluate gray and black box methods for identification of the HLC_{wb} from large scale, heterogeneous datasets using synthetic labelled data. Comparing the outputs of the models to ground truth lead to several significant findings. First, the only

approach studied in this work that does not suffer from significant application barriers (ES calibration) is not able to produce results that are robust to heterogenous building properties. Second, it is shown that calibration error and residual analysis are not sufficient to validate models for thermal property estimation. This is particularly consequential for RC modelling studies (and other calibration studies that rely on high parameter orders), which are highly prevalent in the literature. Third, BES calibration and supervised deep learning both showed strong performance given the constraints of this study. Neither are ready for deployment in a practical context, but this result indicates that they may become a key component of automated building characterization.

Overall, the results indicate that a strong research effort will be required before methods to predict HLC_{wb} from heterogenous buildings can be established for practical use. To support carbon reduction in the existing building stock, we encourage future work to use this data and to contribute code to the online repository to develop reliable, data-driven methods for building property characterization.

Chapter 5

Visual Explanations from Neural Networks Trained on Simulated Building Sensor Data

5.1 Introduction

Epistemology, the discipline concerned with the nature of knowledge, has piqued the interest of philosophers for centuries. To humankind it is unsatisfactory simply that we "know", we rather seek justification and rationalization for our beliefs and modes of understanding. Thus it is unsurprising that the scientific community is seeking answers to the epistemic questions raised by black-box machine learning models, with supervised deep neural networks at the forefront of this inquiry. Although these networks are highly successful at regression and classification tasks, the nature of a network's predictions - the *why* - remains speculative.

The goal of interpretable machine learning is to provide such speculations with the overarching tenets of fairness and ethical decision making, alongside model transparency, trustworthiness and informativeness [61]. In deep learning, two major approaches to interpretability include the development of proxy models and the creation of saliency maps [37]; the latter is the focus of this work. Saliency maps are used to highlight which features of input data are most important to the model, thus providing transparency and informativeness

with regards to the model predictions [5] [37] [47].

Up until now, saliency maps have most commonly been developed on and applied to image classification, likely because vision tasks are easily interpretable by humans so the maps are clearly meaningful. Modern saliency techniques are rarely used for time series prediction tasks, with only a few examples in the literature [7] [89]. Moreover, saliency maps have seen very limited uptake for interesting application cases (such as those presented by the recent onslaught of building data collection) in which the input data is less easy to interpret by humans. These types of high-impact application cases could, however, prove to be where interpretable machine learning is most advantageous.

This work focuses on one of these application cases. There is a foreseeable future in which the decarbonization of the building stock is supported by integrated sensor networks, big-data collection and analysis and the strategic application of machine learning. Supervised deep learning is state-of-the-art in the buildings domain. It has shown promise for applications such as thermal property estimation and heating system identification for retrofit analysis [10], and socio-demographic classification [101]. Further, machine learning experts recognize the opportunity to improve energy-efficiency in buildings as a high-leverage area for artificial intelligence related to climate change [81]. As such, we believe that interpretable machine learning applied to building data presents a valuable opportunity for enquiry. Particular questions explored in this paper include: what types of features do the models learn, do they discover anything about the physical behaviour of a building, what types of data are most effective for prediction and are any notable ethical or privacy concerns illuminated by the prediction process?

With this vision in mind, this paper presents the novel application of saliency maps (i.e. gradient-based activation maps) to a time series regression task in the buildings domain. Four residual neural networks (ResNets) are trained on a synthetic dataset of 16,000 simulated buildings, and the resulting activation maps are visualized and analyzed. The four models

predict heat loss coefficient (HLC), a physical property that dictates the thermal behaviour of a building. Overall, this paper serves as both a pragmatic foray into machine learning interpretability for a valuable application case and a deep dive into supervised deep learning for infrastructure decarbonization. Along with addressing the questions posited above, this work provides essential insight into the usefulness of simulated building data for supervised deep learning.

5.2 Background

Before the aforementioned questions are addressed we provide a detailed overview of saliency methods. In general, saliency maps, also known as explanation maps, are highly popular and numerous proposed implementations exist. To narrow this scope, this paper focuses specifically on Gradient-based Class Activation Maps (Grad-CAMs) [85] [84]. This section of the paper provides a brief review of existing saliency methods and justifies the use of Grad-CAMs in particular. Gradient-based saliency methods are defined formally and intuitively. Finally, the technical details of Grad-CAMs are presented.

5.2.1 Saliency Maps

Saliency maps, also known as explanation maps, are used to estimate the influence of a datum's features on a particular prediction. For example, the saliency map in Figure 5.3 highlights the pixels that were most important in classifying the image as a meerkat. Formally, a saliency map $E : \mathbb{R}^d \rightarrow \mathbb{R}^d$ maps inputs and input vector $x \in \mathbb{R}^d$ to an output object of the same shape [5]. The output object provides a "mapping" that represents input feature importance.

Gradient-based explanation maps are a particular type of saliency method where input feature importance is calculated by finding the gradient for input x with respect to the

model output. The model, $S : \mathbb{R}^d \rightarrow \mathbb{R}^C$, maps an input vector to an output of C classes.[1] $E_{grad}(x) = \frac{dS}{dx}$ thus represents a gradient-based explanation map for a prediction S(x) with respect to input x. Intuitively, a large gradient for a particular input feature indicates a high rate of change in the prediction, with respect to that feature. In other words, the larger the gradient for a particular input feature, the more influence that feature has on the model output.

A myriad of gradient-based explanation approaches have been studied in the literature; a full overview is outside of the scope of this paper. Hooker et al. and Adebayo et al. provide a robust literature review of existing approaches and, most importantly, provide a quantitative benchmark of existing saliency methods. The methods assessed by these works include base estimators such as Guided Backprop [94], Integrated Gradients [95] and Guided GradCAM [84] as well as ensembling method such as SmoothGrad [91] and VarGrad [4].

According to both Adebayo et al. and Hooker et al., many popular saliency methods may not be suitable for practical application because their feature attribution is not dependent on the trained model. In other words, many saliency methods act similarly to deterministic edge-detectors or randomly assign feature importance. In these cases the saliency method is not an appropriate explanatory tool.

Grad-CAM, which is a highly popular method in the literature, is one of the few saliency maps that passes the criteria defined by Adebayo et al.. While Hooker et al. do not consider Grad-CAM directly, they do study ensemble methods that are based on this approach. According to the criteria defined by these authors, ensemble based approaches perform the best, however, they have a high associated computational cost.

Based on the benchmarking of saliency methods discussed above, Grad-CAM was selected for application in this paper; it passes sanity checks that specify a method's practical applicability without the computational burden of more sophisticated ensemble methods.

[1] In the case of regression, $C = 1$

An advantage that ensemble methods have over basic Grad-CAM is that the tend to better localize feature importance. Figure 5.3, for instance, provides an example of an activation map that could be better localized. It should be kept in mind that, as with the meerkat example, the discovered gradient maps may not be perfectly discriminative.

5.2.2 Grad-CAM

Gradient-weighted Class Activation Mapping, or Grad-CAM, is a technique developed by Selvaraju et al. to provide visuale explanations for predictions from convolutional neural networks (CNNs) [84]. Designed for classification problems, Grad-CAM is based on the assumption that the last convolutional layers in a neural network retain the most spatial and semantic information about an input datum. The gradient of the output class (y^c) is therefore taken with respect to the output activations (A^k) for each of the K feature maps in the final convolutional layer of a trained neural network. For regression, there is only a single output class, that is $C = 1$, so the shorthand y will be used. Global average pooling is used on the resulting gradient to obtain the neuron importance, α_k.

$$\alpha_k = AvgPool\left(\frac{dy}{dA^k}\right) \tag{5.1}$$

This process is repeated for each of the K feature maps in the convolution last layer. The outputs α_k are then combined linearly and passed through a ReLU function as seen in equation 5.2.

$$E_{Grad-CAM} = ReLU\left(\sum_k \alpha_k A^k\right) \tag{5.2}$$

where $E_{Grad-CAM}$ is the explanatory saliency map that represents the final attribution of feature importance. Overlaying $E_{Grad-CAM}$ onto the original input that was used for prediction illustrates which input features were most influential for a prediction (see Figure 5.3).

5.3 Methods

As mentioned above, Grad-CAM was originally defined for classification problems; we are using it for regression simply by specifying C=1. For the remainder of the paper we will therefore refer to Gradient-based Activation Maps (Grad-AM, instead of Grad-CAM).

For this work, $E_{Grad-AM}$ was found for all the buildings in the validation dataset for four deep neural networks. The networks were trained on a synthetic dataset of 16,000 buildings. The buildings were programmatically generated using the Building Simulation, Optimization and Surrogate Modelling (BESOS) platform [30], and simulated using EnergyPlus[2] version 9.2.0 using a process similar to that in [10]. Each of the four neural networks was trained to predict a building's HLC using a distinct set of multivariate time series inputs that were chosen to match sensor data that might be collected in a real world context. The same network structure (shown in Figure 5.2) was used for each case. From these networks, Grad-AMs were extracted, analyzed and compared in order to interpret the prediction results for each of the four models. The Grad-AMs were visualized both as overlays on the original input (as is common with image data) and as time series plots (which is an approach unique to this paper). The data creation, model structure and visualization approaches will now be discussed in more detail.

5.3.1 The Dataset

Building energy simulation (BES) software such as EnergyPlus allows for the generation of synthetic building datasets that can be used to run controlled and tractable experiments. This sandboxed environment provides a simplified antecedent to the real world, where computational models can be tested and explored. As such, a synthetic dataset was generated for this study, using the process described below.

[2]https://energyplus.net/

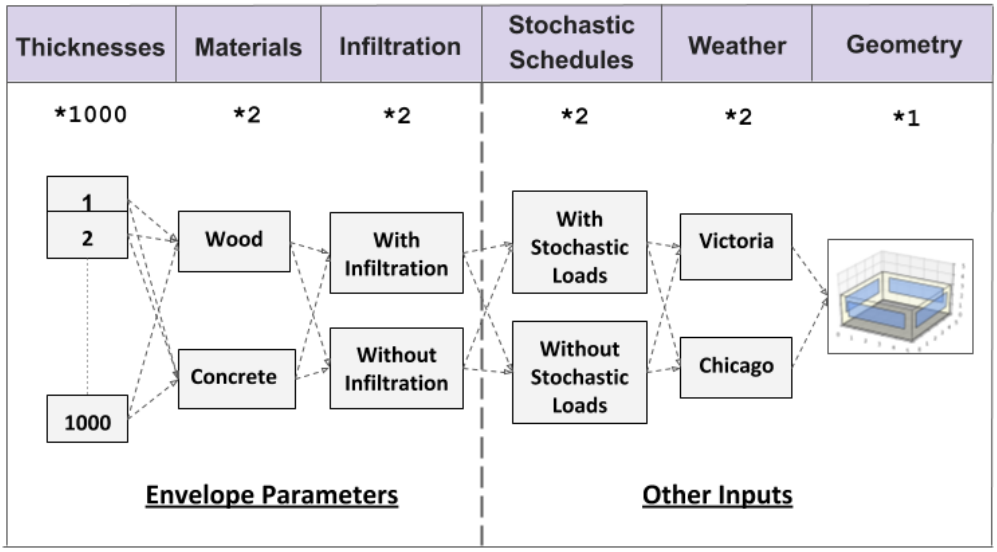


Figure 5.1: The building properties that were manipulated to create the synthetic dataset.

A single EnergyPlus simulation takes a building model as input, and outputs a variety of information including construction details and time-resolved building behaviour. By running many such simulations, a time series dataset for machine learning model training was created. BESOS was used to support rapid the creation of many building prototypes by programmatically manipulating the underlying building model according to the set of predefined attributes (Figure 5.1).[3] Temporal EnergPlus outputs such as outdoor and indoor temperatures, solar gains, and heating power consumption were used as model inputs (X). HLCs were calculated analytically from the simulation outputs and used as training labels (y).[4] Each time series input was scaled to be between 0 and 1 before model training.

5.3.2 Model Structure & Training

In total, four distinct models were trained using different lengths and subsets of the aforementioned time series inputs. Two of the models accept daily inputs (288, 5 minute time

[3]Note that the generated dataset does not represent the full complexity of a real world building stock, but rather provides us with tractable constraints and cases for comparative analysis.

[4]HLC includes both the thermal resistivity of the envelope and the infiltration rate.

steps) and two accept weekly inputs (2000, 5 minute time steps). The included input data was from either the first week or the first day in January.

Each of the four models included solar gains, outdoor temperature and indoor temperature, but only two of the four included the heating power. Otherwise, the training and validation data for each of the four models was identical. Throughout this paper the models will be referred to by a short-hand name that describes their inputs, specifically (1) Daily-NoHeat, (2) Daily-Heat, (3) Weekly-NoHeat and (4) Weekly-Heat.

A convolutional ResNet (Figure 5.2) was used for the four models trained on the time series input types described above. ResNets allow for the training of very deep convolutional networks by including shortcut connections that propagate information from lower layers to higher layers in the network [40]. As demonstrated in by Fawaz et al., convolutional ResNets have achieved high accuracy on time series classification tasks [31]. The output HLC is numerical so its prediction is a regression rather than a classification task but similarly high performance should be expected.

The building dataset was divided into training data and validation data using an 80/20 split which was stratified according to the cases presented in Figure 5.1. The model was trained over 100 epochs with decreasing learning rates (starting at $1x10^{-3}$ and ending at $1x10^{-5}$). Adam optimization [54] was used for all cases, and mean-squared error (MSE) was the defined error metric. The training and validation errors are presented in the results section.

5.3.3 Visualizing Grad-AM for Time Series

As mentioned previously, saliency maps are most commonly used for image classification tasks. In this paper they are repurposed for time series regression. The gradients presented in the next section represent the derivative of the predicted HLC value with respect to the last convolutional layer in the ResNet, as indicated by Figure 5.2 and explained in Section

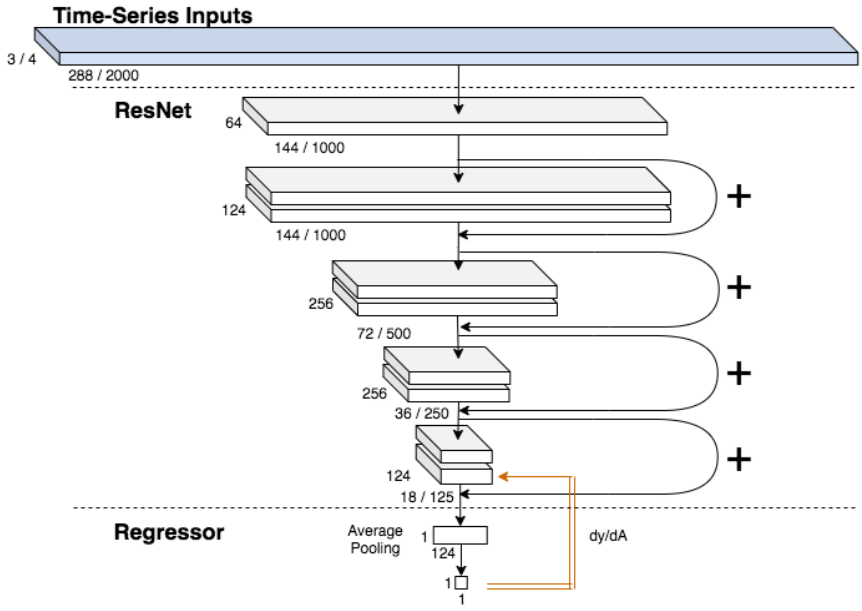


Figure 5.2: The convolutional, ResNet architecture pictured above was used for all four training cases, two of which accept daily inputs (288 time steps) and two of which accept weekly inputs (2000 time steps). The Grad-AMs were retrieved by taking the gradient of the prediction with respect to the last convolutional layer in the network.

5.2.2. These gradients are expanded by a factor of 16 (from 18 to 288 values in the daily case and from 124 to 2000 vales in the weekly case) to match the original input size so that they can be more easily attributed to particular input features.[5] Saliency maps for image data are relatively easy to understand: it is clear when we look at the explanation map overlaying a meerkat that the head and upper body are the most important indicators for prediction. Evaluating time series data is not so easy because the data itself is less interpretable for humans.

Figure 5.3 illustrates an analogy between image-based explanation maps and time series-based explanation maps. The displayed time series includes four input variables over one week. The x-axis represents 2000 timesteps at 5-minute intervals. The yellow portions of the time series overlay represent the areas of highest feature importance. In this example the most relevant features fall periodically in the middle of the day.

[5]This same process is used to reshape gradients calculated on image data so that they overlay the original input.

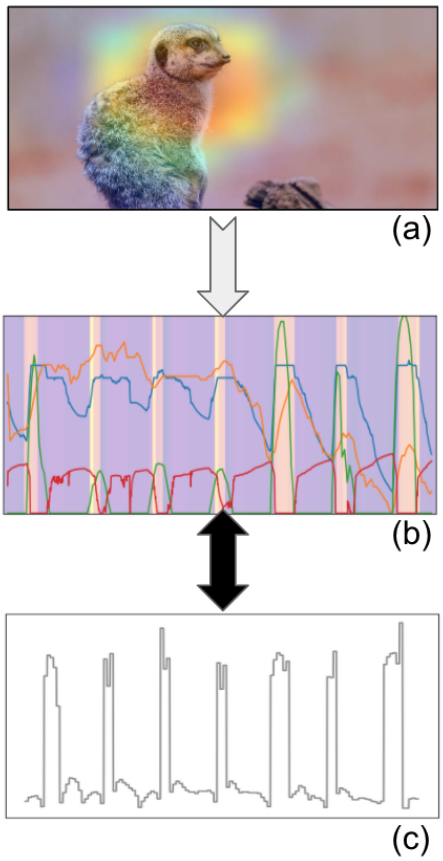


Figure 5.3: Saliency maps are commonly used on image data to attribute picture importance to a final prediction. Analogous heatmaps can be created for time series data to attribute importance to a particular time step. For temporal input, the discovered Grad-AM is technically a 1-D vector so it can also be represented as a time series plot.

Though plotted in Figure 5.3 (b) as a coloured overlay, the time series Grad-AM is really a 1-D vector, where each index is a time step and the value at an index represents the relative influence of that particular time step on the final prediction. The activation maps can therefore also be plotted as time series, as seen in Figure 5.3 (c). This is a valuable way of visualizing the gradients; using this approach, activation maps from many inputs can be plotted on the same axes for comparison.

5.4 Results

This section of the paper presents the Grad-AMs retrieved for buildings in the validation set for each of the four neural networks. Important features of the plots are highlighted in this section, with a full interpretation of the results in the discussion section.

First, the training and validation errors for each of the four neural networks are presented. Next, the activation maps for the networks are visualized for a small subset of cases. Evaluating fewer cases helps to highlight distinctive features of each of the models and build intuition about the Grad-AMs. The time series interpretation of the activation maps are then plotted and compared for all of the buildings in the validation set so that patterns of behaviour across the data and models can be discovered. Finally, correlations between the input variables and the discovered Grad-AMs will be presented and used to quantitatively analyze the feature importance across all cases.

5.4.1 Model Performance

The prediction results for each of the four models on the training and the validation data are presented in Table 5.1. In terms of validation error, the worst performing model was Daily-NoHeat, followed by Weekly-NoHeat, so heating power input improved model performance in general, as we would expect.

Weekly-Heat had a lower validation MSE than training MSE. This result is somewhat unusual; overfitting most commonly occurs when the validation error is higher than the traininer error, while underfitting occurs when both errors are high. A lower validation error than training error likely indicates something else about the data, perhaps that the training set contains more outlier examples than the validation set.

Note that the relative model errors on the validation set might impact the quality of the retrieved Grad-AMs. Models with higher errors may exhibit Grad-AMs that are less well

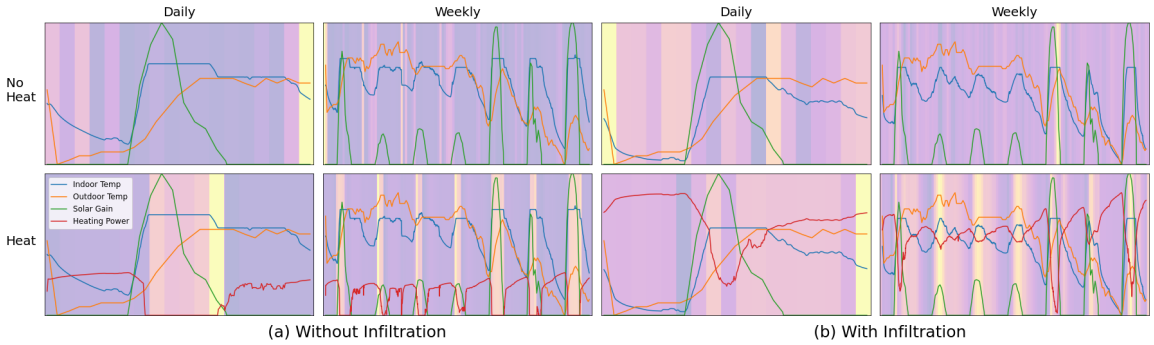


Figure 5.4: Grad-AMs for a wooden building in Chicago. Remember that heating power is always included in the building simulation. It is only excluded as a model input.

localized when compared to those from models with very low errors. Similarly, gradients from models with high validation error are more likely to misattribute feature importance because the final prediction is further away from the ground truth.

Table 5.1: Model prediction results.

Model Name	Training MSE	Validation MSE
Daily-NoHeat	0.9843	5.4405
Daily-Heat	0.5410	0.8378
Weekly-NoHeat	0.5468	2.7195
Weekly-Heat	1.1359	0.4889

5.4.2 Single-Building: Heatmap Representation

Figure 5.4 displays gradient heatmaps for a randomly selected building. It is a wooden building in Chicago that includes stochastic occupancy and equipment load schedules: for subplot (a) the building has no infiltration, while for subplot (b) infiltration is included. Everything else about the building is identical, including its material composition. Gradient heatmaps for each of the four model types are displayed for both cases. These are now qualitatively examined with respect to each of the four trained models.

(a) Without Infiltration

- *Daily-NoHeat*: For this building, this model attributes the highest feature importance to the end of the day, followed by the early morning. On first glance it appears that the model is discovering the most information during the periods when the heating power is dropping.
- *Daily-Heat*: In direct contrast, this model attributes the highest feature importance to the middle of the day. It is very interesting that it is at this time that the heating power is the lowest - for this particular building it is actually 0 - so the model uses its knowledge of heating input to find periods where it is minimal.
- *Weekly-NoHeat*: This model attributes the highest feature importance to the first half of the time series. The weekly model and daily models without heating power attribute the highest feature importance to the night times.
- *Weekly-Heat*: Similar to the daily model, the weekly model with heating power periodically attributes the highest feature importance to periods in the middle of the day.

(b) With Infiltration

- *Daily-NoHeat*: This model finds a relatively similar Grad-AM for the cases with and without infiltration. For both, the highest feature importance occurs in the middle of the night when the indoor temperature is dropping. Unlike the no infiltration case, the infiltration case also finds an important period in the middle of the day, again, when the indoor temperature is dropping. In general, for the infiltration case there is less distinction between features of high and low importance.
- *Daily-Heat*: For this model, the discovered Grad-AM more closely resembles Daily-NoHeat than Daily-Heat in the no infiltration case. That is, instead of finding the period in the middle of the day when heating power input is low it

finds the period at the end of the day when indoor temperature starts dropping.

- *Weekly-NoHeat*: This model does not exhibit as strong a periodicity for the infiltration case, as compared to the non infiltration case. Rather, it seems to find one moment of particularly high importance.
- *Weekly-Heat*: Unlike Daily-Heat with infiltration, which attributed feature importance to the night times, Weekly-Heat attributes feature importance to the middle of the day. Thus, Weekly-Heat in the infiltration cases exhibits similar behaviour to Daily-Heat in the non infiltration case.

5.4.3 Multi-Building: Time Series Representation

As shown in Figure 5.3, the Grad-AM heatmaps can also be plotted as univariate time series variables. Figure 5.5 plots the activation maps in this way for all of the buildings in the validation set. Evaluating these plots helps to highlight the patterns of behaviour of the neural networks across the all of the buildings.

The plots in Figure 5.5 are divided to match the 16 cases shown in Figure 5.1 that were used to create the synthetic dataset but, for brevity, the cases with and without stochastic schedules were combined into one.

- *Daily-NoHeat*: This model exhibits activation maps that have variation within each of the cases displayed, especially compared to the models with heating power input. Even so, some patterns arise. Across all cases aside from the infiltration case in Chicago, Daily-NoHeat tends to find the highest feature importance in the night times or early mornings. It is also notable that the no infiltration and infiltration cases exhibit a pattern of behaviour that is somewhat comparable.
- *Daily-Heat*: The discovered activations for this model are indubitably different from Daily-NoHeat. For the cases without infiltration the model consistently finds the

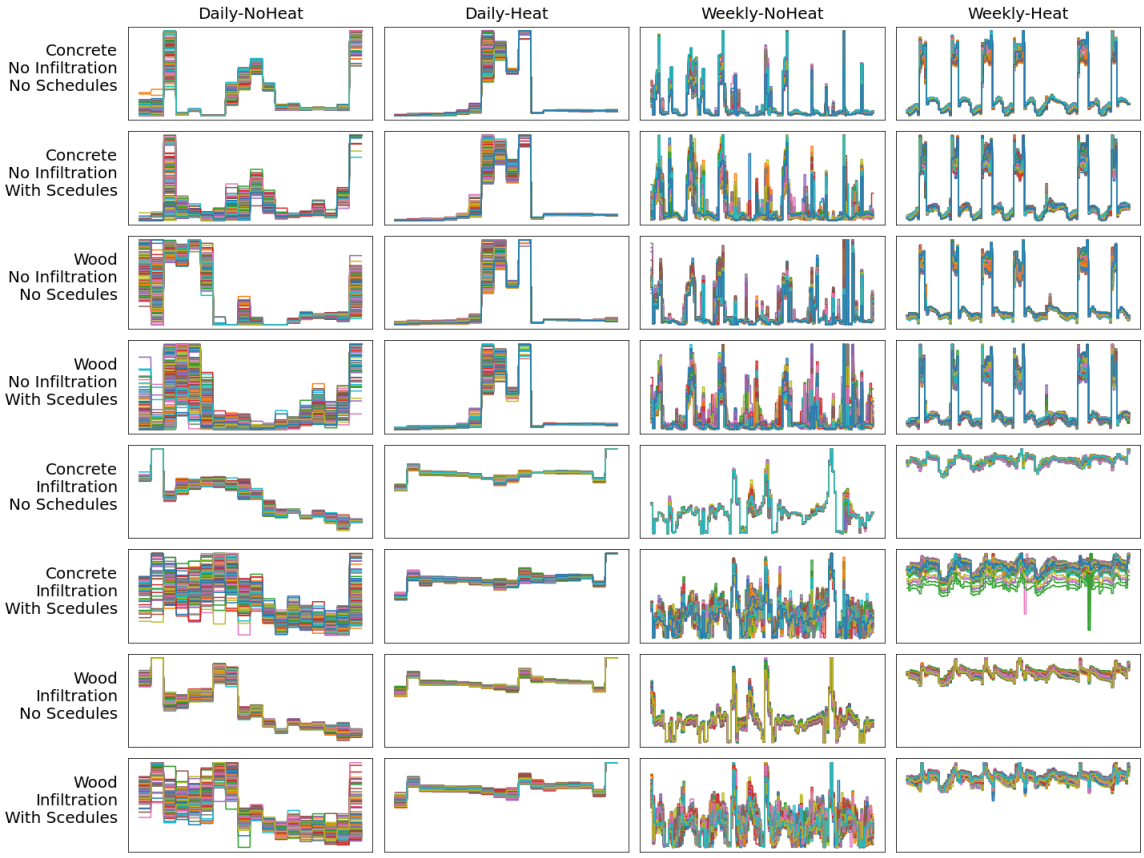


Figure 5.5: Univariate time series representation of the Grad-AM for every building in the validation set and for all four models.

highest importance in the middle of the day when the heating power input is at its lowest, with a spike at the end of this period when the heating power turns on and the indoor temperature begins to drop. The infiltration cases exhibit a distinctly different (but consistent) pattern in which the last timestep exhibits the highest influence on the prediction.

- *Weekly-NoHeat*: It is not immediately easy to discern patterns of behaviour for this model simply by examining the time series plots. We might conclude that the cases without infiltration exhibit some periodicity, but must also point out that it is not very well defined. The infiltration cases in Chicago show a distinct spike at a particular time (as seen in the single building case in Figure 5.4) but, otherwise, no distinct pattern of behaviour is discernible.

- *Weekly-Heat*: The periodicity in this model is much stronger than that for Weekly-NoHeat. It seems clear by examining the gradients that the weekly model with heating input finds feature importance at similar times of day as the daily model with heating input. This periodicity is apparent for the buildings with and without infiltration, but much more clear for the former.

The Grad-AMs for Daily-Heat and Weekly-Heat show distinct and consistent patterns between the infiltration and no infiltration cases. Such distinctions are not as clear for Daily-NoHeat and Weekly-NoHeat. There are also noticeable patterns in the Grad-AMs between Victoria and Chicago for each of the models.

The variability within each case for Daily-NoHeat and Weekly-NoHeat is high. This is mostly caused by the buildings with schedules; our analysis showed that the addition of schedules had a much larger effect on the models without heating power. An illustrative example is included in Figure C.1 in the Appendix.

5.4.4 Correlations Between Grad-AMs and Input Variables

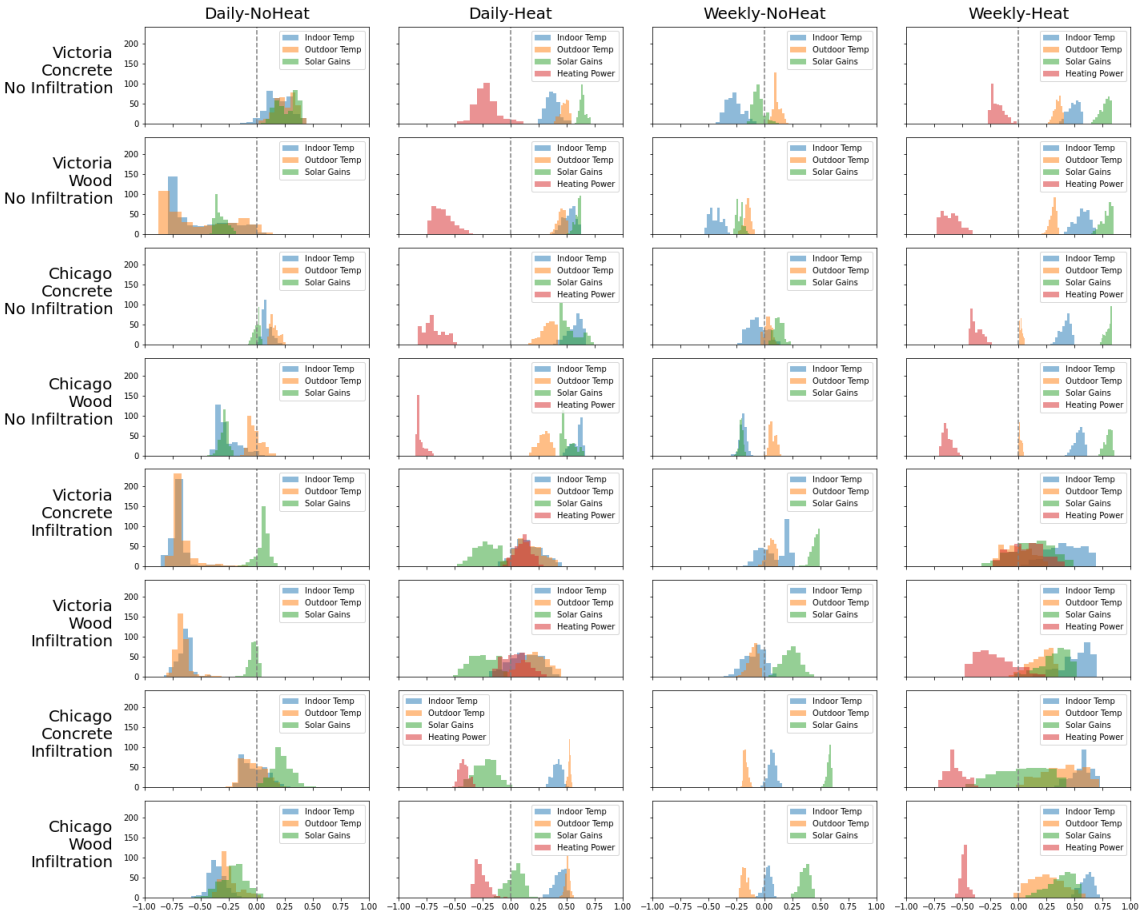


Figure 5.6: Histograms of the Pearson correlation between the time series input and the discovered Grad-AMs for all of the buildings in the validation set for each of the four trained models.

In the previous two sections, the Grad-AMs for all buildings and all four neural networks were evaluated qualitatively. In order to quantify this analysis, correlations between input variables and discovered activations are now considered. Specifically, the Pearson correlation between the various time series inputs and the gradient-based activation values was found for each of the buildings and a histogram of these correlations were plotted.[6] Note that these correlations only provide a partial picture of the nature of the activation maps. Future work

[6]During the initial analysis, the correlation was also taken with the slopes of the time series inputs, but almost no correlation was found.

should emphasize patterns as well, perhaps through clustering.

- *Daily-NoHeat*: For the Victoria infiltration cases and for the wooden building without infiltration in Victoria, there is a strong negative correlation between the gradient-base activations and the indoor and outdoor temperatures for this model. This means that the model is most influenced when the indoor and outdoor temperatures are both low. From the time series plot in Figure 5.4 we can see that this occurs in the early mornings. The Chicago cases, on the other hand, do not exhibit strong correlations. This makes sense considering the large variation in the Grad-AMs for these cases (seen in Figure 5.5).

- *Daily-Heat*: For the cases without infiltration and for the infiltration cases in Chicago, the Daily-Heat Grad-AMs exhibit a strong negative correlation with the heating power input (which is in the middle of the day, so it also when solar power is high). This is consistent with the patterns seen in Figures 5.4 and 5.5. It indicates that the portions in the time series with the lowest heating input are the most influential on model prediction. For the Victoria cases with infiltration, on the other hand, the gradient correlations are not very strong.

- *Weekly-NoHeat*: In general, Weekly-NoHeat exhibits the weakest correlations of all the models. Moreover, the correlations do not match those of Daily-NoHeat, which indicates that the weekly model without heat input does not always distinguish the same patterns as its associated daily model. In the Chicago cases there is a weak correlation with solar, which might indicate a slight daily periodicity in the data.

- *Weekly-Heat*: The correlations in model this are similar to Daily-Heat, confirming the apparent periodicity that can be seen in Figure 5.5. In the infiltration cases, the correlation with heating power input is stronger than the correlation with solar input,

which indicates that the model is specifically seeking periods of low heat input, as opposed to the periods in the middle of the day.

5.5 Discussion

The section above presented a quantitative and qualitative analysis of the Grad-AMs that were retrieved for the whole validation set of synthetically generated buildings. In this section, the results will be interpreted and contextualized. We are particularly concerned with the features that the models learn, whether or not they discover physically meaningful behaviour characteristics, the types of data that are most effective for prediction, whether or not privacy concerns are raised and how we might use saliency maps to inform the design a more robust synthetic dataset. A hypothesis regarding model behaviour will be formulated. Future work should address this hypothesis in more detail.

5.5.1 Does the Model Learn Physically Meaningful Features?

Before continuing, it is important to consider the how difficult this prediction problem was for the model. The dataset used was generated by manipulating the material properties of 16 baseline buildings (Figure 5.1). Within each of these 16 cases, the relationship between the building HLC and the sum total of the heating input is highly linear. For the models with heating power as an input variable, this might therefore be considered a simple prediction problem. In theory, the models simply need to classify the buildings into one of the 16 cases and then find the linear relationship between the sum total of the heat input and the building HLC.

Keeping this in mind it is highly interesting to note that, for the cases without infiltration, the models with heating power actually attribute the highest feature importance to periods of time where the heating power is the lowest. We hypothesize that the models are therefore

learning about the building behaviour when there is no heat input.

This is an interesting and relevant outcome that warrants more exploration. Consider a building's behaviour when the heating power input is low. At these periods of time the thermal behaviour of the building is governed by solely physical properties such as the HLC. The model learns the most during these periods, indicating that it is perhaps learning about the physical, thermal dynamic behaviour of the building.[7]

The correlations with between the gradients and heating power input were weakest for the buildings with infiltration in Victoria. This is likely because the infiltration cases the buildings do not exhibit as distinct a drop in heating power input as in the no infiltration cases; in the former case cold air blows through the buildings at all times of day so heating is more consistently required.[8]

From Figure 5.5 we can see that, for the buildings with infiltration, Daily-Heat and Weekly-Heat consistently consider the period at the end of the day and in the early mornings to make their prediction. Evaluating the input time series we see that at these times the indoor temperature is dropping. Considering that we are trying to predict the building HLC, or the rate at which heat is lost from the building, it seems very logical that the models learn the most from these periods.

The models that did not include heating power did not perform as well as the models that did. There are two plausible reasons for this. First, it could simply be due to the linear relationship between building HLC and sum total heating power discussed above. In this case, if heating power is not provided as input, it would be difficult for the model to find this relationship. Second, it might be because the models cannot as easily find places in the input time series that are most informative, that is, periods where the thermal dynamics are

[7] It is also possible, however, that the model is using this time period to classify the building signature into one of the 16 cases. This can be tested by introducing more diversity into the dataset, as described in the Section 5.6.

[8] Visual inspection of the plots further confirms this, but, for the sake of brevity, the time series for all the building cases were not included.

governed by HLC, opposed to heating power.

While the first explanation is plausible, the behaviour of the models with heating power input seems to provide evidence for the second. That is, the models that do not accept heating power as an input cannot find the periods of time where the thermal behaviour of the building is dictated by physical thermal properties such as the HLC. If this were true, the implication is that the deep learning models do in fact "learn" building physics. The future work section discusses how this hypothesis could be tested and verified. The remainder of this section provides brief discussions on some of other practical implications uncovered through interpretability analysis.

5.5.2 Effective Data Collection

The validation errors of the four trained neural networks (Table 5.1) seem to indicate that including heating power in the input data improves model performance. Analysis of the Grad-AMs, however, indicates that what is actually the most important is the periods of time in which the building has no heat input. This knowledge might assist in targeted data collection programs.

Additionally it is important to note that, based on the correlations in Figure 5.6, indoor temperature is an important indicator for prediction. This is relevant because many data collection initiatives (for example, smart meters that collect temporal energy usage) do not include indoor temperature. These data may not contain enough information for the deep learning models to learn the physical characteristics of buildings that govern thermal behaviour.

5.5.3 Privacy and Ethical Concerns

Any data collection program must consider ethics and privacy. The collection of building data for targeted retrofit programs might introduce equity issues that should be considered.

For instance, if models that include heating input or occupancy information provide much higher performance than those that do not, only building owners with access to this type of data (eg. through ownership of a smart thermostat) will receive benefits in the design of data collection programs. This should be acknowledged and addressed early.

Regarding privacy, tests should be run to ensure that Grad-AMs from a trained model do not attribute highest feature importance to periods in the day for which a building is unoccupied. If this were the case this could pose a security threat.

5.6 Limitations & Future Work

The results and discussion presented provide insight into the behaviour of the deep learning models, but there is still a large degree of subjectivity to the interpretation. More work must be done to confirm the hypothesis that the machine learning models learn the most about the building HLC because of the physical, thermal dynamic behaviour at periods of low heating. Future work should focus on (1) exploring this hypothesis further and (2) expanding this study by improving the dataset.

In order to better understand if the machine learning models learn physically meaningful building features, a new dataset should be created in which no heating power is included in the building simulation. The diversity of the dataset should be increased to include differing geometries, load schedules, infiltration rates, climates and mechanical system types. Using interpretability methods with this dataset could help to clarify some of the questions raised by this study.

The discovered features could provide deeper insight into learned physical behaviours. For example, we might expect that the model attributes the highest importance to periods of time where the indoor temperature is dropping and the outdoor temperature is stable, because at these periods the rate of decay of the indoor temperature directly depends on

the building HLC. Other thermal properties such as the building capacitance should also be examined using this dataset. Finally, infiltration rates must be considered further. The HLC used for prediction in this paper included heat loss from infiltration. Comparison of Grad-AMs found for prediction on HLC with and without infiltration might indicate what the model learns in either case. For example, the model that predicts HLC with infiltration might find correlations with wind speed, while the model that predicts HLC without infiltration might not.

As noted, the dataset used for training and validation can be significantly improved; many variables were not considered in this study. Diversifying the dataset will help overcome overfitting, reduce the linearities in the data and improve our confidence that the model is learning meaningful features.

Despite the simplicity of the dataset, it is a surprising and interesting result to find that the model learns the most at periods where the heating power is the lowest. Presumably, when the thermal dynamics of the building are dictated by the HLC. The models that do not include heating power as an input may perform more poorly because they have trouble finding these periods. If this were true, the implication is that the deep learning model does in fact "learn" building physics. This would be significant, as it likely means that neural networks trained on synthetic data could be re-purposed for use on real world data, for example through transfer [98] or self-supervised [70] learning.

5.7 Conclusion

Innovation has long been driven by epistemological enquiry. The improvement of computational models should be no exception. The application of modern machine learning interpretability methods to real-world cases is rare, but the results of this paper show that this should not be the case. By applying gradient-based saliency maps to four neural net-

works trained on building data, this paper penetrates the black-box nature of the networks to provide novel and applicable insights into their behaviour. Based on these results, we suggest that the continued application of interpretability approaches can help to accelerate the strategic use of machine learning for decarbonizing the existing building stock.

Chapter 6

Conclusions

Challenges such as (1) a lack of data, (2) limited model generalizability and reliability and (3) un-reproducible studies have resulted in restricted industry adoption of machine learning research [45]. The purpose of this thesis was to rigorously evaluate multiple methods for identifying quantitative building characteristics from large, heterogeneous datasets, considering these challenges.

In the first chapter, it was determined that by using gray box models an accurate ranking of *RC* and *RK* is achievable, but that absolute values are harder to determine. A high degree of accuracy is not required to filter retrofit candidates, so it was concluded that the three methods presented are likely sufficient for this purpose. A major point of future work identified in this chapter was to validate model performance against a known ground truth. This inspired the work in Chapter 4.

Chapter 3 was motivated by the observation that deep learning models in particular are affordable, scalable and reusable, and their successful application could prove invaluable in the building performance assessment industry. The findings in this study indicated the potential for the use of deep learning in targeted retrofit analysis. These methods are unused in the literature and it is not clear how well they perform when compared to gray box approaches.

The work comprising Chapter 4 was a natural synthesis of the previous two chapters, which identified (1) a requirement for validation against ground truth and (2) the need for performance benchmarking of novel methods. This chapter benchmarked multiple methods to estimate the heat loss coefficient on quantitative building characteristics, on a novel, extensible synthetic building meter data set. The findings of this work were significant; based on qualitative and quantitative criteria it was determined that none of the evaluated methods are currently suitable for the identified application cases, but that deep learning and surrogate-based calibration showed the most promising ground truth performance. The gray box methods did not perform well in terms of robustness towards heterogenous building properties. This result indicates serious shortcomings to the state-of-the-art in the literature.

A major barrier for supervised deep learning is a lack of datasets containing relevant labels. Chapter 5 begins to address this limitation by evaluating the learning behaviour of models that are trained on synthetic data to see if they might be transferable to real data. This work provides first indication that the models learn physically meaningful features from synthetic data, but there is still a large degree of subjectivity to the interpretation of the results. More work must be done to confirm this result. Long-term, this work could help to overcome the data shortage in this domain and encourage the use of machine learning for energy reduction in buildings.

An important consideration when performing research in this domain (and one reason for the aforementioned data scarcity) is data privacy and ethics. This should always be a fundamental consideration for researchers, industries, governments or other stakeholders who are looking to develop and deploy data-driven methods. Residential building data in particular offers insight into occupant behaviour and lifestyle, so special care should be taken to anonymize and protect user information. Researchers and other stakeholders should work closely together to ensure best practices, in order to provide the most safe and secure future possible.

In conclusion, this thesis provides insight into the use of gray and black box models for thermal property estimation in buildings. It summarizes the state of research in terms of empirical model performance, and determines that significant effort is needed to support the adoption of methods for data-driven applications such as large-scale, targeted retrofit analysis and building stock modelling.

Chapter 7

Future Work

Significant research and development effort will still be required to identify scalable methods for estimating physical thermal properties for retrofit analysis from large data. The methodology presented in Chapter 4 should be used to evaluate more methods, including the decay curve and energy balance methods from Chapter 2 (balance points i.e. energy signatures were already included within the scope of Chapter 4). The same methodology should be used to aid in the development and validation of novel approaches for thermal property characterization.

A major focus of this work was the whole-building HLC, but additional properties are often needed by building energy modellers. They may be continuous or discrete, e.g. the primary heating system of a building [104]. Future work should extend this scope to include more types of building properties. The data creation pipeline used for this work can easily accommodate this.

Chapters 3 and 4 identify strong predictive performance of black box methods, and Chapter 5 indicates that synthetic data might help ease real-world requirements for labelled data. This should be a strong focus of future work. Possible avenues of study include transfer learning [98], self-supervised learning [70], and pretraining with labelled synthetic data sets.

Chapter 4 showed that, in addition to deep learning, surrogate-based calibration shows strong promise. This should also be explored in future work. In particular, surrogate-baesd calibration relies on an underlying physical model. Representative archetype models can be derived if a large number of buildings is to be calibrated. In fact, segmenting a building stock into groups of similar buildings (archetype classification) and deriving a suitable building energy model (architecture characterization) are decisive steps in common calibration processes [93][56][52]. The generalizeability of these apporaches must be explored further.

Finally, the synthetic dataset that was used in this work should be extended to include more building properties such as geometry and number of zones. A more comprehensive dataset will support all of the aforementioned future works.

Overall, significant effort will be required to identify useable models for thermal property estimation from large datasets in practice.

Bibliography

[1] BetterHomesBC-RebateChart.

[2] CleanBC | Government of British Columbia.

[3] Pan-Canadian Framework on Clean Growth and Climate Change : Canada's plan to address climate change and grow the economy.

[4] Julius Adebayo, Justin Gilmer, Ian Goodfellow, and Been Kim. Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values. *arXiv:1810.03307 [cs, stat]*, October 2018. arXiv: 1810.03307.

[5] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. *arXiv:1810.03292 [cs, stat]*, October 2018. arXiv: 1810.03292.

[6] K Arendt, M Jradi, H R Shaker, and C T Veje. COMPARATIVE ANALYSIS OF WHITE-, GRAY- AND BLACK-BOX MODELS FOR THERMAL SIMULATION OF INDOOR ENVIRONMENT: TEACHING BUILDING CASE STUDY. page 8, 2018.

[7] Roy Assaf and Anika Schumann. Explainable Deep Neural Networks for Multivariate Time Series Predictions. In *Proceedings of the Twenty-Eighth International Joint*

Conference on Artificial Intelligence, pages 6488–6490, Macao, China, August 2019. International Joint Conferences on Artificial Intelligence Organization.

[8] Fidel Aznar, Victor Echarri, Carlos Rizo, and Ramón Rizo. Modelling the thermal behaviour of a building facade using deep learning. *PLOS ONE*, 13(12):e0207616, December 2018.

[9] Gaby Baasch, Adam Wicikowski, Gaëlle Faure, and Ralph Evins. Comparing gray box methods to derive building properties from smart thermostat data. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '19, pages 223–232. Association for Computing Machinery.

[10] Gaby M Baasch and Ralph Evins. Targeting Buildings for Energy Retrofit Using Recurrent Neural Networks with Multivariate Time Series. page 6, 2019.

[11] Peder Bacher and Henrik Madsen. Identifying suitable models for the heat dynamics of buildings. *Energy and Buildings*, 43(7):1511–1522, July 2011.

[12] Phillip Biddulph, Virginia Gori, Clifford A. Elwell, Cameron Scott, Caroline Rye, Robert Lowe, and Tadj Oreszczyn. Inferring the thermal resistance and effective thermal mass of a wall using frequent temperature and heat flux measurements. *Energy and Buildings*, 78:10–16, August 2014.

[13] Samuel Dalton Borgeson. *Targeted efficiency: Using customer meter data to improve efficiency program outcomes*. PhD thesis, UC Berkeley, 2013.

[14] Morten Brøgger, Peder Bacher, and Kim B. Wittchen. A hybrid modelling method for improving estimates of the average energy-saving potential of a building stock. *Energy and Buildings*, 199:287–296, September 2019.

[15] Jonathan D. Chambers and Tadj Oreszczyn. Deconstruct: A scalable method of as-built heat power loss coefficient inference for UK dwellings using smart meter data. 183:443–453.

[16] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *arXiv:1406.1078 [cs, stat]*, June 2014. arXiv: 1406.1078.

[17] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv:1412.3555 [cs]*, December 2014. arXiv: 1412.3555.

[18] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated Feedback Recurrent Neural Networks. *arXiv:1502.02367 [cs, stat]*, February 2015. arXiv: 1502.02367.

[19] Dorota Chwieduk. *Solar energy in buildings: thermal balance for efficient heating and cooling*. Academic Press, San Diego, Calif., 2014. OCLC: 881887858.

[20] Daniel Coakley, Paul Raftery, and Marcus Keane. A review of methods to match building energy simulation models to measured data. *Renewable and Sustainable Energy Reviews*, 37:123–141, September 2014.

[21] Tim Cooijmans, Nicolas Ballas, César Laurent, Çağlar Gülçehre, and Aaron Courville. Recurrent Batch Normalization. *arXiv:1603.09025 [cs]*, March 2016. arXiv: 1603.09025.

[22] Drury Crawley, Linda Lawrie, Frederick Winkelmann, W.F. Buhl, Y.Joe Huang, Curtis Pedersen, Richard Strand, Richard Liesen, Daniel Fisher, Michael Witte, and

Jason Glazer. EnergyPlus: Creating a New-Generation Building Energy Simulation Program. *Energy and Buildings*, 33:319–331, April 2001.

[23] 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, April 2001.

[24] S. Danov, J. Carbonell, J. Cipriano, and J. Martí-Herrero. Approaches to evaluate building energy performance from daily consumption data considering dynamic and solar gain effects. *Energy and Buildings*, 57:110–118, February 2013.

[25] Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and Tanaka Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In *International conference on parallel problem solving from nature*, pages 849–858. Springer, 2000.

[26] An-Heleen Deconinck and Staf Roels. Comparison of characterisation methods determining the thermal resistance of building components from onsite measurements. *Energy and Buildings*, 130:309–320, October 2016.

[27] An-Heleen Deconinck and Staf Roels. Is stochastic grey-box modelling suited for physical properties estimation of building components from on-site measurements? *Journal of Building Physics*, 40(5):444–471, March 2017.

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[29] Marian Farah, Paul Birrell, Stefano Conti, and Daniela De Angelis. Bayesian Emulation and Calibration of a Dynamic Epidemic Model for A/H1N1 Influenza. *Journal of*

the American Statistical Association, 109(508):1398–1411, October 2014. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2014.934453.

[30] Gaëlle Faure, Theo Christiaanse, Ralph Evins, and Gaby M. Baasch. BESOS: a Collaborative Building and Energy Simulation Platform. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, BuildSys '19, pages 350–351, New York, NY, USA, November 2019. Association for Computing Machinery.

[31] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, July 2019. arXiv: 1809.04356.

[32] Federal Energy Management Program. M&v guidelines: Measurement and verification for performance-based contracts version 4.0.

[33] Samuel F. Fux, Araz Ashouri, Michael J. Benz, and Lino Guzzella. EKF based self-adaptive thermal model for a passive house. *Energy and Buildings*, 68:811–817, January 2014.

[34] Sushant S Garud, Iftekhar A Karimi, and Markus Kraft. Design of computer experiments: A review. *Computers & Chemical Engineering*, 106:71–95, 2017.

[35] Cristian Ghiaus. Experimental estimation of building energy performance by robust regression. *Energy and Buildings*, 38(6):582–587, June 2006.

[36] Panagiota Gianniou, Christoph Reinhart, David Hsu, Alfred Heller, and Carsten Rode. Estimation of temperature setpoints and heat transfer coefficients among residential buildings in Denmark based on smart meter data. *Building and Environment*, 139:125–133, July 2018.

[37] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. *arXiv:1806.00069 [cs, stat]*, February 2019. arXiv: 1806.00069.

[38] Virginia Gori, Phillip Biddulph, and Clifford A. Elwell. A Bayesian Dynamic Method to Estimate the Thermophysical Properties of Building Elements in All Seasons, Orientations and with Reduced Error. *Energies*, 11(4):802, April 2018.

[39] Stig Hammarsten. A critical appraisal of energy-signature models. *Applied Energy*, 26(2):97–110, January 1987.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. arXiv: 1512.03385.

[41] Y. Heo, R. Choudhary, and G. A. Augenbroe. Calibration of building energy models for retrofit analysis under uncertainty. *Energy and Buildings*, 47:550–560, April 2012.

[42] Jennifer Hiscock. Smart grid in canada 2014. Technical report, report# 2015-018 RP-ANU 411-SGPLAN, Natural Resources Canada, March 2015, 32 . . . , 2014.

[43] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[44] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wanni Zhang. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212:109831, April 2020.

[45] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wanni Zhang. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212:109831, April 2020.

[46] Tianzhen Hong, Zhe Wang, Xuan Luo, and Wanni Zhang. State-of-the-art on research and applications of machine learning in the building life cycle. *Energy and Buildings*, 212:109831, 2020.

[47] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks. *arXiv:1806.10758 [cs, stat]*, November 2019. arXiv: 1806.10758.

[48] Jeremy Howard et al. fastai. https://github.com/fastai/fastai, 2018.

[49] Brent Huchuk, William O'Brien, and Scott Sanner. A longitudinal study of thermostat behaviors based on climate, seasonal, and energy price considerations using connected thermostat data. *Building and Environment*, 139:199–210, July 2018.

[50] Srinivasan Iyengar, Stephen Lee, David Irwin, Prashant Shenoy, and Benjamin Weil. WattHome: A Data-driven Approach for Energy Efficiency Analytics at City-scale. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 396–405, New York, NY, USA, 2018. ACM. event-place: London, United Kingdom.

[51] Amirhosein Jafari and Vanessa Valentin. An optimization framework for building energy retrofits decision-making. *Building and Environment*, 115:118–129, April 2017.

[52] Fatemeh Johari, Giuseppe Peronato, Paria Sadeghian, Xiaoyun Zhao, and Joakim Widén. Urban building energy modeling: State of the art and future prospects. *Renewable and Sustainable Energy Reviews*, 128:109902, 2020.

[53] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization.

[54] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014. arXiv: 1412.6980.

[55] Kevin J. Kircher and K. Max Zhang. On the lumped capacitance approximation accuracy in RC network building models. *Energy and Buildings*, 108:454–462, December 2015.

[56] Martin Heine Kristensen, Rasmus Elbæk Hedegaard, and Steffen Petersen. Hierarchical calibration of archetypes for urban building energy modeling. *Energy and Buildings*, 175:219–234, September 2018.

[57] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. Number: 7553 Publisher: Nature Publishing Group.

[58] Marc Lee. A Green Industrial Revolution: Climate Justice, Green Jobs and Sustainable Production in Canada. page 64.

[59] Marc Lee and Seth Klein. *Winding down BC's fossil fuel industries planning for climate justice in a zero-carbon economy*. 2020. OCLC: 1184685971.

[60] Xinyi Lin, Zhe Tian, Yakai Lu, Hejia Zhang, and Jide Niu. Short-term forecast model of cooling load using load component disaggregation. *Applied Thermal Engineering*, 157:113630, July 2019.

[61] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]*, March 2017. arXiv: 1606.03490.

[62] Mikael Lundin, Staffan Andersson, and Ronny Östin. Development and validation of a method aimed at estimating building performance parameters. *Energy and Buildings*, 36(9):905–914, September 2004.

[63] Zhenjun Ma, Paul Cooper, Daniel Daly, and Laia Ledo. Existing building retrofits: Methodology and state-of-the-art. *Energy and Buildings*, 55:889–902, December 2012.

[64] M. Maasoumy, M. Razmara, M. Shahbakhti, and A. Sangiovanni Vincentelli. Handling model uncertainty in model predictive control for energy efficient buildings. *Energy and Buildings*, 77:377–392, July 2014.

[65] H. Madsen and J. Holst. Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22(1):67–79, March 1995.

[66] D. Majcen, L. C. M. Itard, and H. Visscher. Theoretical vs. actual energy consumption of labelled dwellings in the Netherlands: Discrepancies and policy implications. *Energy Policy*, 54:125–136, March 2013.

[67] Massimiliano Manfren, Niccolò Aste, and Reza Moshksar. Calibration and uncertainty analysis for computer models–a meta-model based approach for integrated building energy simulation. *Applied energy*, 103:627–641, 2013.

[68] Paul R Miles and Ralph C Smith. Parameter estimation using the python package pymcmcstat. 2019.

[69] Evan Mills. Building commissioning: a golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. *Energy Efficiency*, 4(2):145–173, May 2011.

[70] Ishan Misra and Laurens van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. *arXiv:1912.01991 [cs]*, December 2019. arXiv: 1912.01991.

[71] Shreshth Nagpal, Caitlin Mueller, Arfa Aijazi, and Christoph Reinhart. A methodology for auto-calibrating urban building energy models using surrogate modeling techniques | Request PDF. *Journal of Building Performance Simulation*, April 2018.

[72] Zoltán Nagy, Dino Rossi, Christian Hersberger, Silvia Domingo Irigoyen, Clayton Miller, and Arno Schlueter. Balancing envelope and heating system parameters

for zero emissions retrofit using building sensor data. *Applied Energy*, 131:56–66, October 2014.

[73] Gustav Nordström, Helena Johnsson, and Sofia Lidelöw. Using the Energy Signature Method to Estimate the Effective U-Value of Buildings. In Anne Hakansson, Mattias Höjer, Robert J. Howlett, and Lakhmi C Jain, editors, *Sustainability in Energy and Buildings*, Smart Innovation, Systems and Technologies, pages 35–44, Berlin, Heidelberg, 2013. Springer.

[74] Alex Nutkiewicz and Rishee K Jain. Exploring the integration of simulation and deep learning models for urban building energy modelling and retrofit analysis. page 8.

[75] Nilavra Pathak, James Foulds, Nirmalya Roy, Nilanjan Banerjee, and Ryan Robucci. A Bayesian Data Analytics Approach to Buildings' Thermal Parameter Estimation. In *Proceedings of the Tenth ACM International Conference on Future Energy Systems*, e-Energy '19, pages 89–99, New York, NY, USA, 2019. ACM. event-place: Phoenix, AZ, USA.

[76] Samuel Prívara, Jiří Cigler, Zdeněk Váňa, Frauke Oldewurtel, Carina Sagerschnig, and Eva Žáčeková. Building modeling as a crucial part for building predictive control. *Energy and Buildings*, 56:8–22, January 2013.

[77] Christoffer Rasmussen, Peder Bacher, Davide Calì, Henrik Aalborg Nielsen, and Henrik Madsen. Method for Scalable and Automatised Thermal Building Performance Documentation and Screening. *Energies*, 13(15):3866, January 2020. Number: 15 Publisher: Multidisciplinary Digital Publishing Institute.

[78] Christoph F Reinhart and Carlos Cerezo Davila. Urban building energy modeling–a review of a nascent field. *Building and Environment*, 97:196–202, 2016.

[79] G. Reynders, J. Diriken, and D. Saelens. Quality of grey-box models and identified parameters as function of the accuracy of input and observation signals. *Energy and Buildings*, 82:263–274, October 2014.

[80] Staf Roels. EBC annex 71 building energy performance assessment based on in-situ measurements.

[81] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli, Konrad P. Kording, Carla Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling Climate Change with Machine Learning. *arXiv:1906.05433 [cs, stat]*, November 2019. arXiv: 1906.05433.

[82] S. S. Sablani, A. Kacimov, J. Perret, A. S. Mujumdar, and A. Campo. Non-iterative estimation of heat transfer coefficients using artificial neural network models. *International Journal of Heat and Mass Transfer*, 48(3):665–679, January 2005.

[83] Padraig Scully. Smart meter market report. Technical report, IOT Analytics, 2019.

[84] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, October 2019. arXiv: 1610.02391.

[85] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Why did you say that? *arXiv:1611.07450 [cs, stat]*, January 2017. arXiv: 1611.07450.

[86] Marieline Senave, Staf Roels, Glenn Reynders, Stijn Verbeke, and Dirk Saelens. Assessment of data analysis methods to identify the heat loss coefficient from on-board monitoring data. *Energy and Buildings*, 209:109706, February 2020.

[87] Marieline Senave, Staf Roels, Stijn Verbeke, Evi Lambie, and Dirk Saelens. Sensitivity of Characterizing the Heat Loss Coefficient through On-Board Monitoring: A Case Study Analysis. *Energies*, 12(17):3322, January 2019. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

[88] Mohammad Haris Shamsi, Usman Ali, and James O'Donnell. A generalization approach for reduced order modelling of commercial buildings. *Journal of Building Performance Simulation*, 12(6):729–744, November 2019. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19401493.2019.1641554.

[89] Shoaib Ahmed Siddiqui, Dominique Mercier, Mohsin Munir, Andreas Dengel, and Sheraz Ahmed. TSViz: Demystification of Deep Learning Models for Time-Series Analysis. *IEEE Access*, 7:67027–67040, 2019.

[90] Ramvir Singh, R. S. Bhoopal, and Sajjan Kumar. Prediction of effective thermal conductivity of moist porous materials using artificial neural network approach. *Building and Environment*, 46(12):2603–2608, December 2011.

[91] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. *arXiv:1706.03825 [cs, stat]*, June 2017. arXiv: 1706.03825.

[92] Leslie N. Smith. Cyclical Learning Rates for Training Neural Networks. *arXiv:1506.01186 [cs]*, June 2015. arXiv: 1506.01186.

[93] Julia Sokol, Carlos Cerezo Davila, and Christoph F. Reinhart. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. *Energy and Buildings*, 134:11–24, January 2017.

[94] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for Simplicity: The All Convolutional Net. *ICLR*, 2015.

[95] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328, Sydney, NSW, Australia, August 2017. JMLR.org.

[96] Seyed Amin Tabatabaei. A Data Analysis Approach for Diagnosing Malfunctioning in Domestic Space Heating. January 2016.

[97] Seyed Amin Tabatabaei, Wim Van der Ham, Michel C. A. Klein, and Jan Treur. A Data Analysis Technique to Estimate the Thermal Characteristics of a House. *Energies*, 10(9):1358, September 2017.

[98] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A Survey on Deep Transfer Learning. *arXiv:1808.01974 [cs, stat]*, August 2018. arXiv: 1808.01974.

[99] Wim Van der Ham, Michel Klein, Seyed Amin Tabatabaei, Dilhan Thilakarathne, and Jan Treur. Methods for a Smart Thermostat to Estimate the Characteristics of a House Based on Sensor Data. *Energy Procedia*, 95:467–474, September 2016.

[100] City of Vancouver. Climate Emergency Action Plan.

[101] Yi Wang, Qixin Chen, Dahua Gan, Jingwei Yang, Daniel S. Kirschen, and Chongqing Kang. Deep Learning-Based Socio-Demographic Information Identification From

Smart Meter Data. *IEEE Transactions on Smart Grid*, 10(3):2593–2602, May 2019. Conference Name: IEEE Transactions on Smart Grid.

[102] Yi Wang, Qixin Chen, Tao Hong, and Chongqing Kang. Review of smart meter data analytics: Applications, methodologies, and challenges. *IEEE Transactions on Smart Grid*, 10(3):3125–3148, 2018.

[103] Phil Webber, Andy Gouldson, and Niall Kerr. The impacts of household retrofit and domestic energy efficiency schemes: A large scale, ex post evaluation. *Energy Policy*, 84:35–43, September 2015.

[104] Paul Westermann, Chirag Deb, Arno Schlueter, and Ralph Evins. Unsupervised learning of energy signatures to identify the heating system and building type using smart meter data. *Applied Energy*, 264:114715, 2020.

[105] Paul Westermann and Ralph Evins. Surrogate modelling for sustainable building design–a review. *Energy and Buildings*, 198:170–186, 2019.

[106] Yilmaz et al. BEIS, smart metering-based innovation and building performance: A BEIS / TEDDINET exploratory workshop, 2016.

[107] Diana Ürge Vorsatz, Luisa F. Cabeza, Susana Serrano, Camila Barreneche, and Ksenia Petrichenko. Heating and cooling energy trends and drivers in buildings. *Renewable and Sustainable Energy Reviews*, 41:85–98, January 2015.

Appendix A

Chapter 3

The synthetic dataset used for regression over R was generated using the Building Energy Simulation, Optimization and Surrogate Modelling (BESOS) platform[1] and EnergyPlus, as described in Figure 2. BESOS is a cloud-based research platform used for building energy simulation and optimization. Amongst other things, the platform provides functionality to produce many distinct sample buildings by parameterizing model inputs. Usually the generated samples are used for optimization (ex. using Genetic Algorithms) or for training surrogate models (ex. using Artificial Neural Networks). For the purpose of this project, the sampling functionality provided by BESOS was used to randomly vary the thickness and the density of each of the building materials, thus varying the whole building R-value and the simulated energy usage. 10 initial building designs were used to generate a total of 966 homes. Future work will continue to use the BESOS platform to generate a more robust dataset by including more building geometries, parameterizing more model inputs other than material properties and varying inputs such as weather and occupant schedules.

After many building designs were generated using BESOS, the energy use of each design was simulated with EnergyPlus, a standard software for building energy modelling [23]. This produced a multivariate time series dataset at 10 minute granularity, where the input

[1] https://besos.uvic.ca/

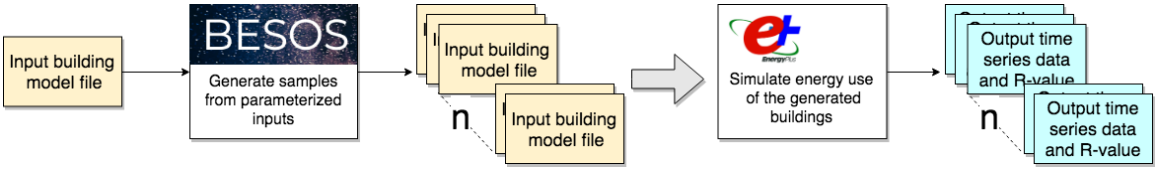


Figure A.1: *Step 1*: Use the BESOS platform to generate many example buildings from a single EnergyPlus model. *Step 2*: Use EnergyPlus to run an annual simulation for each building generated in step 1.

variables consist of indoor temperature, outdoor temperature, and heating system power, for each thermal zone in the building. The time series are summed together for each thermal zone to produce a total of 3 features for each building. The output variables for prediction are the whole-building values for R, as derived from the EnergyPlus input data.

Appendix B

Chapter 4

B.1 Whole-building heat loss coefficient

The whole-building HLC quantifies the rate at which heat is lost through the building envelope via convective, conductive and radiative forces. This knowledge is instrumental for estimating the benefits of building retrofits [72] or assessing the quality of a building post-construction [66]. HLC can be determined by evaluation the thermal energy balance of a building.

Equation B.1 describes the dynamic heat flows in a building as a function of timestep t, where $\dot{Q}_{int}$ is the heat flow from internal gains, $\dot{Q}_{hsys}$ is the heating system supply, $\dot{Q}_{sol}$ is the heat flow from solar gains, $\dot{Q}_{env}$ is the heat flow through the envelope, $\dot{Q}_{inf}$ is the heat flow due to infiltration,[1] C is the effective heat capacity, or capacitance, in J/K, and $dTin/dt$ is the rate of change of the indoor temperature. All of the heat flows are measured in W.

$$C\frac{dT_{in}}{dt}(t) = \dot{Q}_{int}(t) + \dot{Q}_{hsys}(t) + \dot{Q}_{sol}(t) + \dot{Q}_{env}(t) + \dot{Q}_{inf}(t) \tag{B.1}$$

Rearranging equation B.1 shows that infiltration is implicitly included in the whole-building HLC values. This is relevant for calculating HLC from EnergyPlus outputs in the

[1] For simplicity, ventilation is not considered here.

data creation pipeline (B.2).

Equations B.2 and B.3 express $\dot{Q}_{env}$ and $\dot{Q}_{inf}$ in terms of the difference between external and internal temperature.

$$\dot{Q}_{env}(t) = \frac{1}{R}(T_{ext}(t) - T_{in}(t)) \tag{B.2}$$

where R is the thermal resistance of the building envelope [K/W], T_{ext} is the external temperature and T_{in} is the internal temperature.

$$\dot{Q}_{inf}(t) = m * c_{p,air}(T_{ext}(t) - T_{in}(t)) \tag{B.3}$$

where m is the air mass flow rate [kg/s] and $c_{p,air}$ is the air specific heat capacity [J/kg K]. Equation B.1 can thus be rewritten as:

$$C\frac{dT_{in}}{dt}(t) = \dot{Q}_{int}(t) + \dot{Q}_{hsys}(t) + \dot{Q}_{sol}(t) + HLC_{wb}(T_{ext} - T_{in}) \tag{B.4}$$

$$HLC_{wb} = HLC_{inf} + HLC_{env} \tag{B.5}$$

where $HLC_{inf} = \dot{m} * c_{p,air}$ is the heat lost due to infiltration and $HLC_{env} = \frac{1}{R}$, where R is the thermal resistivity of the building envelope in K/W. HLC_{wb} is the whole-building heat loss coefficient. We can see now that HLC depends on both the infiltration rate and the thermal resistivity of the building envelope.

B.2 Calculating HLC from EnergyPlus outputs

In [86] Senave et al. present a methodology for solving these values from the Trnsys simulation software. In this work we adapt their approach to use EnergyPlus outputs.

HLC_{inf} (B.1) is the product of the air mass flow rate, $\dot{m}$, and the air specific heat capacity,

$c_{p,air}$. The air mass flow rate was calculated directly by EnergyPlus and recorded as a time series output.[2] The mean yearly value of this output variable was multiplied by the specific heat capacity for air to calculate HLC_{inf}.

Note that for all the cases in which infiltration was 0, HLC_{inf} was also 0. The calculation for HLC_{env} is considerably more complicated. It can be calculated using an analogy to RC circuit model, where the thermal resistances of the building envelope are analogous to resistors in a circuit. The building envelope can be represented by three resistors in series: (1) the interior surface resistance, R_{int}, (2) the resistance of the material layers, R_{mat}, and (3) the exterior surface resistance, R_{ext}:

$$HLC_{env} = (R_{int} + R_{mat} + R_{ext})^{-1} \tag{B.6}$$

R_{int}, R_{mat} and R_{ext} represent the respective resistances of all the building surfaces in parallel. For instance, R_{int} represents the parallel resistances for each individual indoor surface. These values can therefore be found by taking the sum of the reciprocals of the resistances of each surface, as seen in equation B.7. The resistances of each surface are reported by EnergyPlus, but the models outputs do not account for area. Therefore, the reciprocals of the resistances are multiplied by their associated surface areas as follows:

$$\frac{1}{R_i} = \sum_{s \in S} A_s * \frac{1}{R_{i:s}} \tag{B.7}$$

where $i \in \{\text{int, mat, ext}\}$, S is the set of all surfaces, $R_{i:s}$ is the resistance of the surface and A_s is the area of the surface.

The values for $R_{i:s}$ are calculate differently for the three resistance types. $R_{mat:s}$ is output directly by EnergyPlus. The calculation for R_{int} and R_{ext} requires the evaluation of time-resolved heat transfer coefficients (HTCs), measured in W/m^{2}K. HTCs are proportionality

[2]The EnergyPlus output variable is called: Zone Infiltration Current Density Volume Flow Rate

constants that dictate the given amount of heat exchange by convective and radiative forces at a building surface. Each HTC can be viewed as the inverse of a resistance. Each HTC at a given surface acts in parallel, so $h_{total} = h_1 + h_2 + ... + h_n$. In EnergyPlus, the equations for heat exchange due to convection and radiation depend directly on HTCs so

$$R_{y:s} = 1/mean(h_{y:s:conv} + h_{y:s:rad}) \tag{B.8}$$

where $y \in \{\text{int}, \text{ext}\}$, $h_{y:s:conv}$ is the surface convective HTC and $h_{y:s:rad}$ is the surface radiative HTC.

The remaining calculation considerations for each of the three R values are summarized below:

1. $R_{int:s}$: The internal surface radiation HTCs calculated by EnergyPlus are modelled by the software internally and are not easily accessible to the user (see the EnergyPlus documentation[3] for more information). Therefore, for the purpose of this study, only convection is included in the calculation for $R_{int:s}$. The exclusion of the radiative HTCs may result in a slightly larger absolute errors in HLC estimation, but it should not affect the comparisons between methods or the parametric analysis within the methods.

2. $R_{mat:s}$: The material R value is the sum of the resistance of the material layers that compose the surface. It is calculated directly by EnergyPlus and is reported as the surface U-value, or $1/R_{mat:s}$, in [W/m^{2}K].

3. $R_{ext:s}$: As described by the EnergyPlus documentation,[4] at the external surfaces convection and radiation to the ground, air and sky are modelled by EnergyPlus and

[3] https://bigladdersoftware.com/epx/docs/9-2/engineering-reference/inside-heat-balance.html

[4] https://bigladdersoftware.com/epx/docs/9-2/engineering-reference/outside-surface-heat-balance.html#outside-surface-heat-balance

the associated HTCs are directly available to the user as time-resolved output variables. These output variables are used to find $R_{ext:s}$.

B.3 Material Property Ranges

Surface	Material Layers	Thickness Ranges (m)
Wall	Stucco	[0.015, 0.030]
	Plywood or Concrete	[0.006, 0.03] or [0.2, 0.3]
	Insulation	[0.035, 0.3048]
	Gypsum	[0.00633, 0.0159]
Window	Glass	[0.001, 0.01]
	Air Gap	[0.006, 0.02]
	Glass	[0.001, 0.01]
Floor	Plywood or Concrete	0.0127 or 0.1016
Roof	Roof Membrane	[0.0012, 0.0095]
	Insulation	[0.1, 0.3]
	Metal Decking	[0.0007, 0.0015]

Table B.1: Material composition of the buildings and the thickness ranges used for parametric generation of buildings meter data for our synthetic data set.

Appendix C

Chapter 5

The variation in Grad-AMs within a particular case for the models without heating input (Figure 5.5) is mostly attributed to the buildings that include stochastic schedules. Figure C.1 illustrates this by plotting heatmaps for Daily-Heat for a subset of the buildings. The buildings in each row have the same HLC, they only differ according to their occupancy and equipment schedules.

Figure C.1: Grad-AMs for the daily models for wooden buildings in Victoria with infiltration, separated by the schedule and no schedule cases. The heat maps are plotted in ascending ordered according to predicted HLC.