JOURNAL OF BUILDING PERFORMANCE SIMULATION
2022, VOL. 15, NO. 4, 445–464
https://doi.org/10.1080/19401493.2020.1864474

Evaluation of data-driven thermal models for multi-hour predictions using residential smart thermostat data

Brent Huchuk [a], Scott Sanner [a] and William O'Brien [b]

[a]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada; [b]Department of Civil and Environmental Engineering, Carleton University, Ottawa, Canada

ABSTRACT

Predictive residential HVAC controls can reduce a building's energy consumption; however, they require customized thermal models for each home. In this setting, detailed physical models are not practical. Fortunately, the recent availability of fine-grained thermostat data from residential buildings combined with modern machine learning creates an unprecedented opportunity to build customized data-driven thermal models. We trained and evaluated a range of promising candidate data-driven thermal models for multi-hour predictions using a sliding training window over logged temperature and equipment runtime data from 1000 smart thermostats. The models included machine learning methods, time series models, grey box models, and a simple baseline. Since many models can incorporate exogenous data, we also investigate which combination of features and history provides the best predictions of indoor air temperature. We conclude that lasso and ridge regression with solar, fan, heating and cooling runtime, and 20-minutes of history provided the lowest errors across our sample.

ARTICLE HISTORY
Received 10 July 2020
Accepted 4 December 2020

KEYWORDS
Residential buildings; thermal models; data-driven models

1. Introduction

Residential buildings account for a significant proportion of North America's total energy use; approximately 20% in both Canada and the U.S. (Natural Resources Canada 2019; U.S. Energy Information Administration 2019a). More than half of that energy is consumed by space conditioning (Natural Resources Canada 2019; U.S. Energy Information Administration 2019b), and is predominantly managed by rudimentary reactive thermal controls (U.S. Energy Information Administration 2015) – a technology first introduced in the early 1900's (Peffer et al. 2011). Replacing reactive with predictive controls can reduce peak and total energy use (Candanedo, Dehkordi, and Tamasauskas 2015; Afram and Janabi-Sharifi 2017; Ławryńczuk and Ocłoń 2019), but depends on the ability to build accurate thermal models of the home.

For residential heating, ventilation, and air conditioning (HVAC) applications, a thermal model replicates the response (e.g. air temperature) of the building to near-term environmental conditions and system inputs. Unfortunately, developing a model remains the 'most time-demanding and costly part' of implementing predictive control in buildings (Prívara et al. 2013). Generating a detailed, physics-based model of each home is not feasible. Residential buildings are unique, and it is impractical to suggest gathering each building's physical properties – the U.S. alone has over 118 million homes (U.S. Energy Information Administration 2015). However, with the data available from Internet-connected smart thermostats, there is now an opportunity for widespread development and testing of data-driven thermal models.

To date, there has been no thorough evaluation of data-driven approaches for residential HVAC thermal modelling. In this paper, we provide such an evaluation using a comprehensive dataset of 1000 homes. Our objectives were to determine (1) which features are crucial for predictions with data-driven models and (2) which models give the lowest temperature prediction errors across this population of thermostats. Our selection of features and models were based on their simplicity and ability to be used by edge-computing available on the thermostats. We considered various model classes including grey box, random forest, lasso and ridge regression, and auto-regressive models with exogenous variables. We trained each model online with the data from the 14 days prior to prediction (shown in Figure 4). The trained model was used for multiple prediction sequences each day. We determined that fan runtime, solar data, and a time series history of 20 minutes lead to the best temperature

CONTACT Brent Huchuk brent.huchuk@mail.utoronto.ca

prediction accuracy. Furthermore, we observe that the simple lasso and ridge regression models that have been carefully tuned to prevent overfitting work well, are efficient to train, and outperform a variety of other state-of-the-art data-driven thermal models. Ultimately, we remark that the best approaches can be efficiently deployed on-device for real-time data-driven thermal modelling that offers significant potential for energy savings through better prediction.

The remainder of the paper is structured as follows. In Section 2 we present an overview of past thermal modelling methods and applications for residential buildings. Section 3 discusses the features included, candidate models, data source, and how the evaluation of our sliding window predictions were conducted. Sections 4 and 5 presents the results and associated discussion on how the models and feature combinations performed on thermal prediction. Finally, Section 6 addresses the best overall model selection and future directions regarding the application of our prescribed solution.

2. Review of thermal modelling

Thermal models represent the effects of the heat transfer mechanisms (i.e. radiation, convection, and conduction) between the building, the ambient conditions, internal gains (e.g. occupants and their contents), and the HVAC equipment. A model can broadly be classified into one of three categories: white box, grey box, or black box.

2.1. White box models

White box models are the most detailed solution as, built from first principles, they capture all the various forms of energy flow. Implementing a white box model is typically done utilizing a software program such as TRNSYS (Klein 2017) or EnergyPlus (U.S. Department of Energy 2020). These programs solve the intricate series of underlying heat transfer equations either analytically with transfer methods or numerically using the finite difference method (Clarke 2007). Constructing white box models requires detailed knowledge of the building characteristics (e.g. geometry, orientation, construction), and specific details of the HVAC equipment (Prívara et al. 2013).

White box models have been used sparingly to represent residential buildings for control applications. Often a white box model is utilized instead of relying on a physical building for study (Surles and Henze 2012; Cetin, Manuel, and Novoselac 2016) or as part of a larger testbed (Alibabaei, Fung, and Raahemifar 2016; Alibabaei et al. 2017). Alibabaei et al. (2016; 2017) incorporated a white box model of a demonstration home during the development of a co-simulation platform. The TRNSYS model originally was developed using the physical dimensions of the home and the thermal resistive properties of the windows and walls (Safa, Fung, and Kumar 2015). The constructed model was able to be calibrated using their extensive sensor network installed in the demonstration facility. This process of model commissioning is generally not an option for most practitioners or use cases. In an alternative approach to white box model development, researchers have relied on supercomputers to construct and tune the models using multiple data sources such as smart meters and satellite imagery (Sanyal, New, and Edwards 2013; New et al. 2018). While this alleviates some problems, such as collecting physical property data, scaling this solution is challenging. In particular, practitioners would need to address getting access to a supercomputer, different data sources being available for each home, and the need to recalibrate or retrain models. Finally, these models would be calibrated only to energy usage patterns, because of their use of the metre data, and not the temperature response of the home. To calibrate to temperature response, data would need to be made available from something like a smart thermostat which exposes new privacy concerns compared to the other open data used.

2.2. Grey box models

Grey box models are a common tool employed by building researchers to address some of the challenges with white box models (Coley and Penman 1996; Afram and Janabi-Sharifi 2015b; Burger and Moura 2016; Zeifman, Lazrak, and Roth 2019; Baasch et al. 2019). The model type incorporates expert knowledge when defining the model but also uses measured data from a given space to train the model. When trained, the parameters of the grey box are intended to retain physical meanings. For example, in a trained model of a single thermal zone (Figure 1)

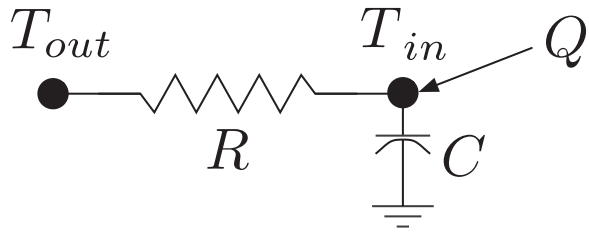

Figure 1. Schematic of a grey box model for a single thermal zone. The indoor (T_{in}) and outdoor (T_{out}) air temperatures represent the nodes of the circuit. The thermal resistance (R) and capacitance (C) of the zone are based on the construction and envelope the space. Thermal loads (Q) are assumed to be deposited directly to the indoor air node.

the values of thermal resistance (*R*) and thermal capacitance (*C*) capture how energy flows in and out of the zone, and how energy is stored or released from the thermal mass. We would describe the grey box model in Figure 1 as a *1R1C* model because of the single resistance and single capacitance term. Instead of assigning the *R* and *C* values based on detailed knowledge of the building's construction (as done with white box models), the parameter are estimated from measured data in the thermal zone (Coley and Penman 1996; Missaoui et al. 2014; Dong and Lam 2014; Burger and Moura 2016; Zeifman, Lazrak, and Roth 2019).

In practice it can be challenging to train physically accurate grey box models. The learned model parameters are a relative value and do not directly translate to physical values (Zeifman, Lazrak, and Roth 2019). Since the number of feasible solutions can be vast, grey box models should be verified against reasonable value ranges to ensure reasonable physical meaning (Dimitriou et al. 2015; Brastein et al. 2018). In order to improve the performance of the models, it is common to increase the model complexity beyond a *1R1C* configuration. While these additional temperature nodes, and connected resistive or capacitance values are physically justifiable, they are a challenge to validate with data. For instance, often a temperature node for the wall or the predominant thermal mass is included (Madsen and Holst 1995; Zeifman, Lazrak, and Roth 2019; Wang and Chen 2019), which in most residential buildings are not measured.

The configurations of grey box models used in practice has been highly varied. Multiple researchers have utilized a single resistance and capacitance term (Reddy, Norford, and Kempton 1991; Burger and Moura 2016; Baasch et al. 2019) while others have implemented two or three of each. In models with a large number of parameters (Dong and Lam 2014; Missaoui et al. 2014) the models were developed from simulated buildings or off of well-instrumented test homes; allowing for at least some specific knowledge of the building. Baasch et al. (2019) utilized data from commercially available smart thermostats to evaluate multiple different grey box methods for deriving building characteristics. By using only the data from smart thermostats, the researchers were limited to indoor and outdoor temperatures and HVAC equipment duty cycling (the normalized value of runtime over a fixed timestep). Similar to Baasch et al.'s (2019) data constraints, Burger and Moura (2016) utilized the same inputs except using a binary on/off flag for equipment.

The features used in grey box models have varied just like the model configurations. Some models have included measurements for wind speed (Zeifman, Lazrak, and Roth 2019) and restricted modelling to only during the night to avoid uncertainty with non-HVAC related loads (Baasch et al. 2019; Zeifman, Lazrak, and Roth 2019). Wang et al. (2019) included wind speed in their model along with normal direct irradiance and internal gains. Other models incorporated non-HVAC related loads by including variables such as the global irradiation on a single façade (Wang and Chen 2019). The façade specific data was possibly because of the test house they had used (Chen, Athienitis, and Galal 2010; Chen, Galal, and Athienitis 2010); but even the best performing configuration (*3R2C*) was outperformed by other data-driven methods.

2.3. Black box models

White or grey box models, when properly built and trained, can be utilized for multiple applications. For example, grey box models though trained using temperature measurements are often used to compare the thermal characteristics of spaces and not ultimately used for thermal prediction (Zeifman, Lazrak, and Roth 2019; Baasch et al. 2019). This general applicability, while advantageous in reducing the number of models being trained, can be sacrificed in favour of specialized black box models that perform singular tasks well. Black box (or sometimes called data-driven) models have traditionally been derived from statistical time series formulations such as auto-regressive model with exogenous variables (ARX) (Molina et al. 2013; Burger and Moura 2017; Wang and Chen 2019), auto-regressive moving average model with exogenous variables (ARMAX) (Bacher and Andersen 2014), or seasonal auto-regressive integrated moving average model with exogenous variables (SARIMAX) (Hossain, Zhang, and Ardakanian 2019). Compared to grey box models, the data-driven models can have better performance for temperature prediction (Afram and Janabi-Sharifi 2015a; Wang and Chen 2019) but their solutions can lack explainability. This lack of explainability is especially true when they are built using modern machine learning methods such as neural networks (Moon and Kim 2010; Ashtiani, Mirzaei, and Haghighat 2014; Afram and Janabi-Sharifi 2015a; Afram et al. 2017, 2018b; Wang and Chen 2019). The neural networks, while powerful remain problematic to implement at scale and potentially challenging to optimize with during control. They also represent a more complicated tool for problems that otherwise can be captured by previously linear models.

Jin et al. (2017); Burger and Moura (2017) both implemented black box models that were very similar in structure to grey box ARX models – the finite difference equation for a *1R1C* model is in fact analogous to a first-order ARX model. However, the general forms were

relaxed to allow each term to be trained individually and not tied to the physical relationships forced by grey box models. Burger and Moura (2017) noted that to better approximate the system, additional exogenous variables could be set to a higher order lag than the autoregressive terms. The order of lag on the exogenous variable was not made with physics-based justifications. Afram et al. (2018a) compared methods for modelling the thermal zone of a single test house using air temperature measurements and flow rates from the equipment. Neural nets were beaten by both grey box and ARX models for RMSE and mean absolute error (MAE); though other white box methods did appear to outperform them all.

3. Methodology

3.1. Problem definition

We required a thermal model to predict the indoor air temperature change ($\Delta T_{in,t}$) over a timestep t. The change in temperature is a result of the energy gained or lost to the ambient conditions (T_{out}) and thermal loads (Q) from the previous timesteps $t-M$ to the current timestep t. The general relation is expressed in Equation (1a). For our predictions, we reformulated the air temperature relation to predict the next observed temperature instead of the change in temperature (Equation (1b)). The remainder of this section addresses the expansion of the thermal load terms (Q) and selection of recent history length (M), the models used to express our objective, and how those models were trained and evaluated.

$$\Delta T_{in,t} = f\left(T_{out,t}, Q_t, \ldots, T_{out,t-M}, Q_{t-M}\right) \quad (1a)$$

$$T_{in,t+1} = f\left(T_{in,t}, T_{out,t}, Q_t, \ldots T_{out,t-M}, Q_{t-M}\right) \quad (1b)$$

3.2. Features

From the definition of the thermal model (Equation (1b)), the factors included in the thermal load term (Q_t) needed to be defined. The heating and cooling equipment's runtime were deemed essential (q_{heat} and q_{cool}). For additional thermal loads, we limited ourselves only to easily accessible information (i.e. measurements made by the thermostat or easily accessible to the thermostat), and values which did not require user input (e.g. HVAC equipment capacities). We considered the introduction of many potential terms but present only the most promising: solar data (Section 3.2.1) and fan runtime (Section 3.2.2). Finally, because of the established delayed effects in a building's temperature response (Athienitis, Stylianou, and Shou 1990; Clarke 2007; Doiron, O'Brien, and Athienitis 2011) we investigated the addition of various lengths of recent history (Section 3.2.3). During model training, all the feature data is assumed to have been logged and stored. During evaluation (i.e. prediction), all exogenous data is assumed to be known perfectly with only the chaining of predictions and the errors in the model itself being uncertain. In an application of the model, predictions for the outdoor temperatures and solar would be required as would predictions of the heating, cooling, or fan runtimes. These additional predictions would all contribute more uncertainty to the models than is quantified in our analysis.

3.2.1. Solar data

Solar gains in residential buildings are a major contributor to both the heating and cooling loads (Huang et al. 1999). Anecdotally, we found that changes in indoor air temperature lag the solar data while other environmental conditions remained consistent (see Figure 2 in the early afternoon). Despite the observable and known effects of solar gains, solar data has been inconsistently included in past thermal models. Some applications have omitted solar or avoided it entirely by limiting training to only nighttime periods (Cole et al. 2014; Zeifman, Lazrak, and Roth 2019; Baasch et al. 2019); a limitation we did not wish to impose on our predictions. Other researchers have included solar data but broken down the data by orientations (Bacher and Andersen 2014), used a site-specific weather station (Wang et al. 2019; Wang, Tang, and Song 2020), or used standard ground-surface measurements (Wang and Chen 2019). We deemed the use of a building specific solar data (e.g. accounting for orientation of surfaces or local weather stations) or calibrating solar data for different surfaces, as not being feasible. The inclusion of these additional tunable components would further increase the number of parameters needing to be trained and would become inaccurate based on occupant's shading decisions. The use of a single standard, community-level, weather-station based measurements seemed the most practical.

Solar data was not a provided value in our dataset so we needed an additional source of data. Each smart thermostat is already matched to a weather station so it is plausible that solar data could be accessible to the individual thermostats. In a predictive system, solar data would need to be provided both historically for training and as part of the weather forecast for prediction. These implementation details are left for future investigation. We downloaded global horizontal irradiance from the National Solar Radiation Database API[1]. The API's data (for direct normal, global horizontal, and diffuse horizontal irradiance) was in 30-minute intervals and was resampled to five-minute intervals, to match the frequency of our thermostat data, using linear interpolation between

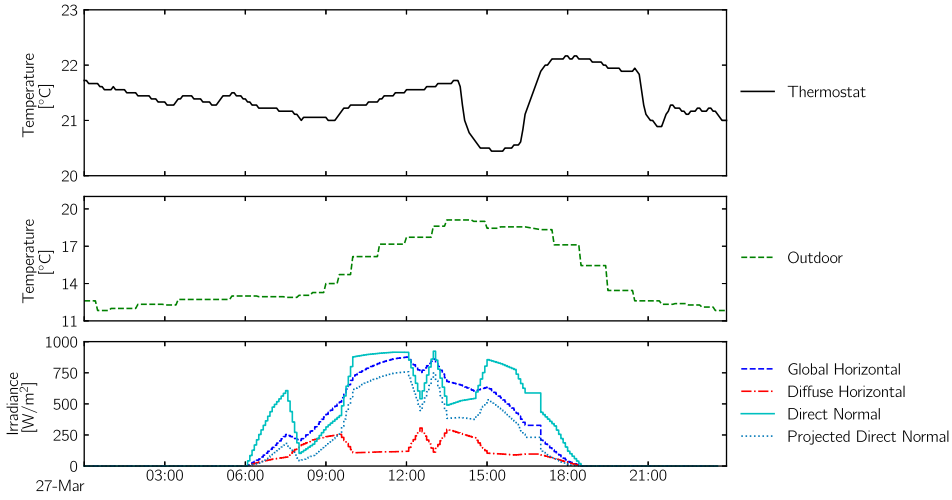

Figure 2. A 24-hour period of thermostat interval data showing thermostat temperature (top), outdoor temperature (middle), solar irradiance measurements (bottom). A sudden drop in direct normal and global horizontal (or increase in diffuse horizontal) appears to proceed a temperature drop around 13:00.

measurements. The data provided was in units of watts per square-meter.

3.2.2. Fan runtimes

In centralized air-based systems, the fan, located with the furnace, is generally used to move air over the heating or cooling coil and distribute conditioned air throughout the home's duct work. A fan can also be used to circulate the air within the home without active heating or cooling from the coils. On many smart thermostats today, users are able to specify a minimum amount of time each hour that the fan should run to help prevent stratification, improve air quality, and also help to balance the temperature throughout the home. As a result, the amount of fan runtime is often greater than the amount of heating or cooling runtime in the home. The fan runtime is measured independently of the heating and cooling runtime by the thermostats and reported as its own value. The effects of fan operation vary greatly depending on the home but can have a considerable impact on indoor air temperatures when neither heating nor cooling is running. Anecdotally, Figure 3 clearly exhibits periods during which the indoor air temperature changes with fan runtime and no active heating or cooling. We speculated that the inclusion of fan runtime should improve the thermal models of some users in the general population but not all.

3.2.3. Recent history terms

The final feature consideration was the length of history given (M) for all the input features. Recent history terms help data-driven models approximate the dynamics and true underlying state of the system. The delayed effects need to be considered both for the indoor air temperature and the exogenous variables (Burger and Moura 2017). In fact, multiple horizons could be considered given the multiple response times in buildings (Athienitis, Stylianou, and Shou 1990). We allowed the models to determine correct weightings for these features during training instead of excluding features using pre-training feature selection. We tested models with different lengths of history; building up from using only time t up to including history from time $t-4$. Given the five-minute timesteps of our interval data, this represents 20 minutes of historical data. The length of history was deemed adequate for general effects of air temperature given trade-offs in performance later observed (Section 4.2). We note that this time range may be brief for systems with large thermal mass or when the home is undergoing precooling or preheating of the thermal mass.

3.3. Candidate models

From our review of thermal modelling (Section 2) we concluded that grey and black box models were our preferred solutions for training a thermal model using smart thermostat data. For grey box models we chose to utilize only the *1R1C* configuration (Section 3.3.1). For black box models we limited our study to four alternatives: the general time-series model ARX (Section 3.3.2), and three standard regression models from machine learning with lasso and ridge regression (Section 3.3.3), and random forest regression (Section 3.3.4). Finally, we designed a single simple baseline to provide context to other models' prediction accuracy (Section 3.3.5).

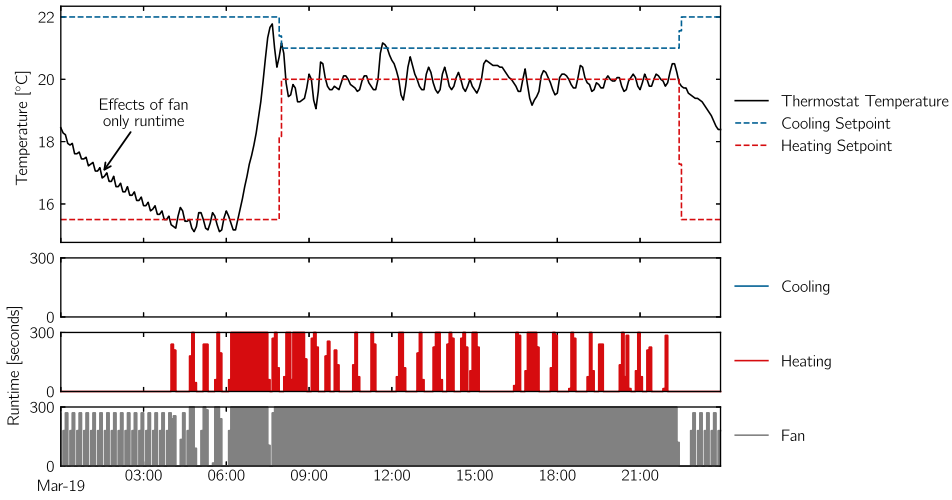

Figure 3. A 24-hour period of thermostat interval data showing thermostat temperature and setpoints (top), heating and cooling equipment runtimes (middle), fan runtime (bottom). Fan cycles early in the early morning (00:00 to 08:00) are seen corresponding to temperature changes on the thermostat.

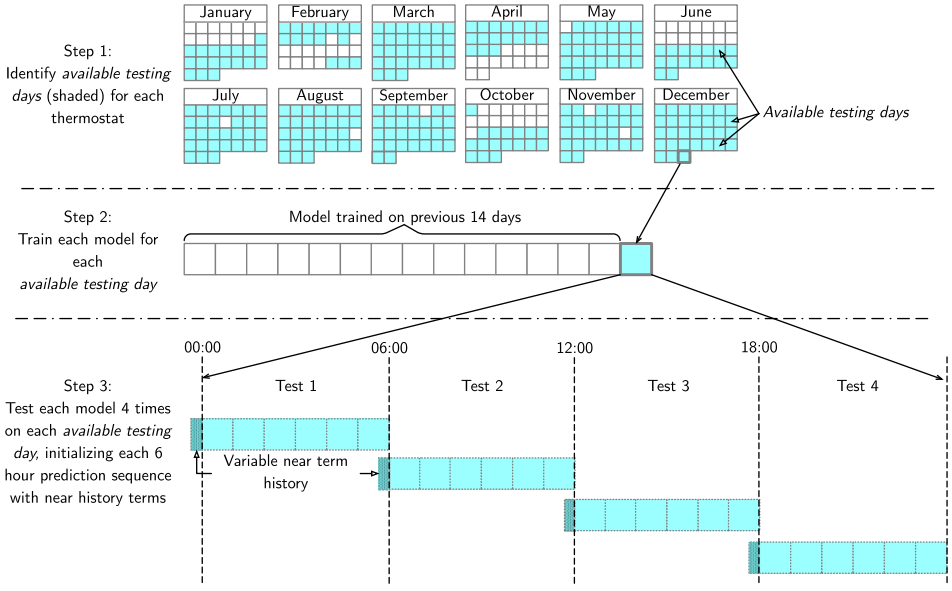

Figure 4. Model evaluation steps for each thermostat.

3.3.1. Grey box

We tested the same single resistance and capacitance configuration (*1R1C*), as was shown in Figure 1, and which has been used effectively by others (Reddy, Norford, and Kempton 1991; Burger and Moura 2016; Baasch et al. 2019). For our purposes, the original problem formulation (Equation (1b)) was expressed using the finite difference representation of the *1R1C* model (Equation (2)). The thermal load term (Q_t) of the *1R1C* model (Equation (2a)) was expanded out to incorporate the features for heating and cooling runtime ($q_{heat,t}$ and $q_{cool,t}$) in addition to a term for solar irradiance ($q_{solar,t}$), shown in Equation (2b). These values were measured in seconds for runtimes and as a power density for solar irradiance. Without required information on homes such as equipment sizes, orientation or façade details, we reformulated Equation (2b) to Equation (2c) to allow the model to learn appropriate scaling factors with α and the β terms. Neither the fan runtime nor were any recent history terms included with the grey box model as both are observational effects and did not come from the formalism of the finite difference equation. The parameters of the model were found using a non-linear least square method in the *SciPy* package (Virtanen et al. 2020).

$$T_{in,t+1} = T_{in,t} + \frac{\Delta t}{C}\left[\frac{(T_{out,t} - T_{in,t})}{R} + Q_t\right] \tag{2a}$$

$$= T_{in,t} + \frac{\Delta t}{C}\left[\frac{(T_{out,t} - T_{in,t})}{R}\right.$$

$$+ q_{heat,t} + q_{cool,t} + q_{solar,t}\Big] \tag{2b}$$

$$= T_{in,t} + \Delta t\alpha\Big[\left(T_{out,t} - T_{in,t}\right) + \beta_1 q_{heat,t} + \beta_2 q_{cool,t} + \beta_3 q_{solar,t}\Big] \tag{2c}$$

3.3.2. Auto-regressive with exogenous variables

Auto-regressive models with exogenous variables (ARX) have frequently been used by building researchers (Burger and Moura 2017; Afram et al. 2018b; Wang and Chen 2019). ARX models have a similar linear form (and are identical for a first-order version) as the finite difference formulation (Equation (2)) but it relaxes the physical relations of the terms. By relaxing the physical meaning, we were able to introduce the recent history terms (in the form of time lags $m \in M$) and the additional effects of the fan ($\phi_{fan,t}$). Using a similar formulation as in Equation (2b), the general form of the ARX equation is shown in Equation (3). The auto-regressive components and the exogenous variables both were considered to have lags up to order M. The individual exogenous variables may each require their own value of the lag terms (i.e. cooling and heating may function at different rates), however we left this to be handled by the parameter estimation. The noise term (ϵ_t) was assumed to be Gaussian. We used the *PyFlux* (Taylor 2016) package to fit the models using maximum likelihood estimation.

$$T_{in,t+1} = \sum_{m=0}^{M}\big[\beta_{1,m}T_{in,t-m} + \beta_{2,m}T_{out,t-m} + \beta_{3,m}q_{heat,t-m} + \beta_{4,m}q_{cool,t-m} + \beta_{5,m}q_{solar,t-m} + \beta_{6,m}\phi_{fan,t-m}\big] + \epsilon_t \tag{3}$$

3.3.3. Lasso and ridge regression

The ARX model improves on the grey box model by allowing for additional features and previous history terms to better approximate the system dynamics. Yet, the ARX model's typical training methodology does not prevent over-fitting to the training set. To train a better predictor, we used an alternative training methodology on the same general form as in Equation (3). Lasso and ridge regressions are an ordinary least squares regression with an ℓ_1 or ℓ_2 regularization respectively. During the minimization of the sum of squared error between the observed next timestep value and predicted values, the coefficients ($\beta_{j,m}$'s) are penalized based on the ℓ_1-norm (Equation (4)) or ℓ_2-norm (Equation (5)) over all training data points D. The two regularization methods have some distinct benefits. With the lasso regression, the ℓ_1 regularization will try and provide sparsity and will remove unimportant features. With the ridge regression, the ℓ_2 regularization will instead only try to minimize the weights of unimportant features but never eliminate them. We implemented both regression methods in *scikit-learn* (Pedregosa et al. 2011) with a three-fold cross validation on the training set to select the best regularization strength (λ) for each model trained. Values of $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ were selected during training based on which value of λ resulted in the lowest mean squared error.

$$\arg\min_{\beta_{1..6,0..M}} \sum_{t=0}^{D-1}\left(T_{in,t+1} - \sum_{m=0}^{M}[\beta_{1,m}T_{in,t-m} + \beta_{2,m}T_{out,t-m} + \beta_{3,m}q_{heat,t-m} + \beta_{4,m}q_{cool,t-m} + \beta_{5,m}q_{solar,t-m} + \beta_{6,m}\phi_{fan,t-m}]\right)^2 + \lambda \underbrace{\sum_{m=0}^{M}\sum_{j=1}^{6}|\beta_{j,m}|}_{\ell_1\text{regularizer}} \tag{4}$$

$$\arg\min_{\beta_{1..6,0..M}} \sum_{t=0}^{D-1}\left(T_{in,t+1} - \sum_{m=0}^{M}[\beta_{1,m}T_{in,t-m} + \beta_{2,m}T_{out,t-m} + \beta_{3,m}q_{heat,t-m} + \beta_{4,m}q_{cool,t-m} + \beta_{5,m}q_{solar,t-m} + \beta_{6,m}\phi_{fan,t-m}]\right)^2 + \lambda \underbrace{\sum_{m=0}^{M}\sum_{j=1}^{6}\beta_{j,m}^2}_{\ell_2\text{regularizer}} \tag{5}$$

3.3.4. Random forest regression

Random forest regression was the final machine learning technique we considered. Tree-based methods are able to provide a non-linear regression as opposed to both the ARX and ridge regression – a similar advantage to neural networks while also easier to train. As an ensemble technique, random forests predict using multiple decision trees (Breiman 2001). Random forests are less susceptible to over-fitting than a single decision tree (Breiman 2001) as it determines a final value by taking the mean of all the tree's predictions. The random forest models were trained using the same input features as the ridge regression and ARX model. We applied a three-fold cross validation on the training of the random forest using *scikit-learn* (Pedregosa et al. 2011) to determine

the best hyperparameters for each model. The hyperparameters tuned were the maximum depth limit of the trees (choices: {1, 5}) and the number of trees (choices: {10, 100}).

3.3.5. Baseline

We used a single baseline to help assess if our methods were in fact learning and help gauge their performance. Previous research has often neglected to include even the most rudimentary baselines to understand how much better their proposed solutions may be doing. Our baseline method was to hold the last measured air temperature measurement ($T_{air,t}$) constant over the prediction window. For example, if the last air temperature measurement was 23°C then the baseline would predict 23°C over the entire prediction window.

3.4. Model evaluation

Our model evaluation framework was designed to replicate a scenario for training and testing in which both were occurring on the thermostats themselves, a.k.a. 'on the edge'. In this situation, the thermostats would have relatively little computation power and not be able store full histories of data. Figure 4, illustrates the three main steps in the evaluation process. *Available testing days* for a thermostat were identified based on how complete the data was over the past 14 days and on the day of testing (Figure 4, Step 1). We set an 80% minimum requirement for the previous 14 days and the *testing day* needed to be missing fewer than 8 timesteps (or 40 minutes). The prediction's sensitivity to the training data requirements were not explored, though Wang et al. (Wang et al. 2019) found using either 10 or 20 days did not appreciably change predicting performance. After implementing these data requirements, the median number of *available testing days* per thermostat in our sample of 1000 homes was 59.

The actual model training and testing was conducted once *available testing days* had been identified. For each thermostat, and on each *available testing day* all of the model and feature configurations were trained using all available data from the 14 previous days (Figure 4, Step 2). Each model's potential features are shown in Figure 5. The indoor and outdoor temperatures (T) along with runtimes (R) were used in all combinations except for the baseline. Solar data (S) and fan runtimes (F) were introduced independently and used together in all the data-driven (black box) models. The recent history was provided for all the features when incorporated into the ridge regression, random forest, and ARX models.

With the data and computation limitations on the thermostat, we aimed to minimize the number of models being trained before prediction and reused a single model over an entire *testing day* (Figure 4, Step 3). For each *available testing day*, the same prediction process was repeated with each model and feature combination. For each day, four tests were conducted using six-hour prediction sequences (i.e. 72 timesteps) made starting at 00:00, 06:00, 12:00, and 18:00. During each of these prediction sequences, the indoor air temperature predictions were fed forward at each timestep. For example, a prediction at 01:00 would be made using up to the previous four indoor temperature predictions (based on the current number of recent history terms being tested) for the ARX, random forest, and lasso and ridge regression models. In contrast, the grey box model would have used only the temperature prediction that was made for 00:55 because recent history was not a viable feature. All the other exogenous variables (i.e. runtimes, solar, outdoor air temperature) used the true historical

	Indoor Temperature	Outdoor Temperature	Heating/ Cooling Runtime	Solar	Fan Runtime	Recent History
	(T)		(R)	(S)	(F)	t −4:t−1
Baseline	✓					
Grey Box	✓	✓	✓	✓		
Lasso Regression	✓	✓	✓	✓	✓	✓
Ridge Regression	✓	✓	✓	✓	✓	✓
Random Forest	✓	✓	✓	✓	✓	✓
ARX	✓	✓	✓	✓	✓	✓

Figure 5. Available features used with each model during testing.

measurements during prediction. Six hours was deemed to be an acceptable time length for most predictive applications however longer horizons could be required for specific applications (e.g. day ahead price arbitrage) or for thermally massive buildings.

Two specific metrics were logged for each prediction sequence (p). For each sequence, the root mean squared error (RMSE) and mean absolute error (MAE) were calculated at each timestep of the prediction horizon (N) between the measured air temperature ($T_{in,t+1}$) and predicted ($\hat{T}_{in,t+1}$). Both the RMSE and MAE were aggregated per thermostat (r) as the average of the metric over all predictions (P) for that thermostat. The explicit definitions of both these values are shown in Equations (6) and (7) respectively. Later in our analysis of errors by time of day (Section 4.5) we restricted the sets of P to only those predictions sequences starting at the specified hour.

$$\mathrm{RMSE}_r = \frac{1}{|P|}\sum_{p}^{P}\left[\left(\frac{\sum_{t=0}^{N-1}\left(T_{in,t+1}-\hat{T}_{in,t+1}\right)^2}{N-1}\right)^{\frac{1}{2}}\right] \tag{6}$$

$$\mathrm{MAE}_r = \frac{1}{|P|}\sum_{p}^{P}\left[\frac{\sum_{t=0}^{N-1}\left|T_{in,t+1}-\hat{T}_{in,t+1}\right|}{N-1}\right] \tag{7}$$

3.5. Data overview

We used the Donate Your Data (DYD) dataset (ecobee Inc 2020) as the source of all our smart thermostat data. The dataset is provided by a single manufacturer of smart thermostats (ecobee Inc.). The dataset contains over 100,000 thermostats found globally, but predominately in North America. For each thermostat, there is user-provided metadata (e.g. square-footage, house style, etc.) and interval data collected by the thermostat. The interval data is five-minute measurements of temperature, equipment runtime, schedule setpoints, overrides, and motion data. A more detailed description of the data has been conducted in other previous work (Huchuk, O'Brien, and Sanner 2018; Huchuk, Sanner, and O'Brien 2019).

We associated a latitude and longitude to each thermostat in order to map to a climate zone and retrieve solar data. To protect the privacy of the thermostat users, only city and province/state information is provided in the DYD dataset. We used the *uszipcode* python package (Hu 2019) and the available location information to estimate the latitude and longitude of the homes. Solar data was provided by the National Renewal Energy Laboratory's National Solar Radiation Database based on the determined latitude and longitude (National Renewable Energy Laboratory 2018).

We screened the available thermostats before sampling 1000 devices for use in our study. The study sample was drawn from a population of thermostats which:

(1) the thermostat user had only one thermostat,
(2) they were listed as located in the U.S.,
(3) they had $<$ 10% missing data in 2016,
(4) they were successfully matched to a latitude and longitude, and
(5) they had available solar data in 2016.

4. Results

The results of the model evaluations are presented in the same order as they were conducted. The first evaluations (Section 4.1) utilized the various feature combinations at a single timestep (t). Based on these findings we introduced increasing amounts of recent history information and evaluated performance changes (Section 4.2). Finally using the cumulative results we are able to address best performance overall for our sample of thermostats (Section 4.3) and the performance characteristics for the top performing feature set (Sections 4.5 and 4.6). We illustrate the performance of the various models on single thermostats in Section 4.4.

4.1. Adding additional features

From all the predictions (P) made per thermostat (i.e. on all *available testing days* and for the four six-hour prediction sequences per day), the average RMSE_r and MAE_r were calculated. Figure 6 shows the distributions in average performance per thermostat for (a) MAE_r and (b) RMSE_r. The baseline method is shown beside other models which included runtime (R) for illustrative purposes; though it did not use this feature. The grey box model is not included in combinations where fan (F) was introduced as it was not trained using that feature.

We note the following specific items regarding the performance of the models and feature combinations. The random forest performs the same as the baseline both in absolute (MAE_r) error and variation (RMSE_r) in the errors. This trend implies the random forest has often learned simply to hold the same last indoor temperature observation.

Ridge regression, lasso regression, and ARX consistently have the best performance, which is unsurprising given they are alternative forms of the same equation. The addition of just the solar data (S) feature had a notable

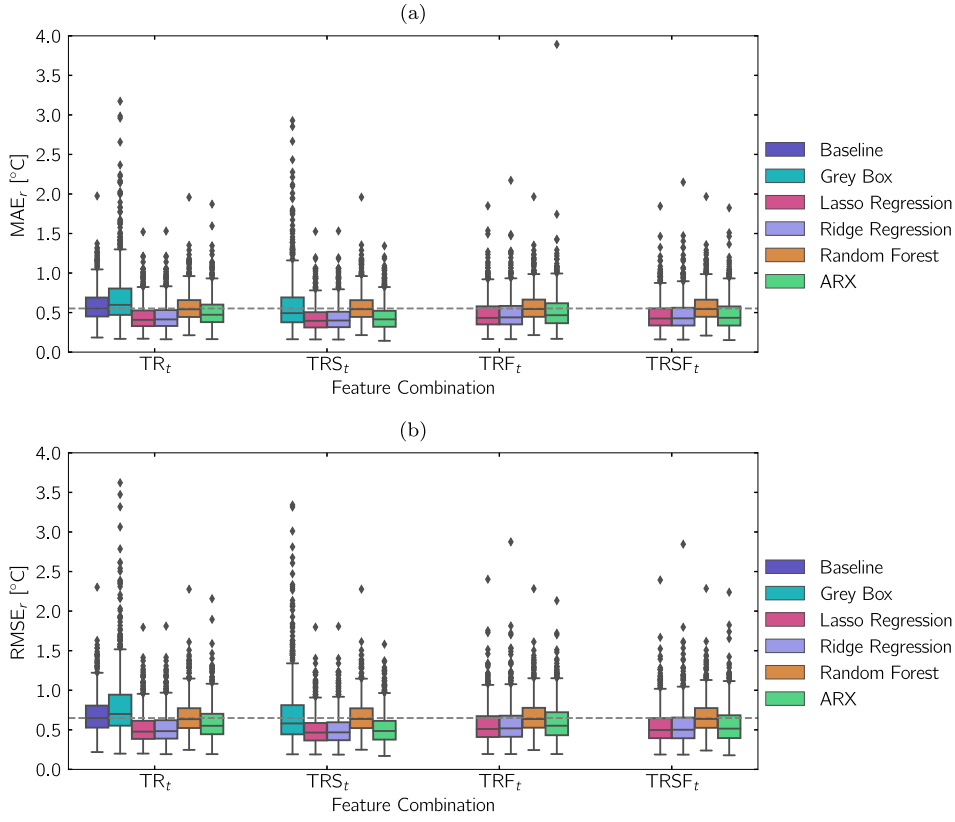

Figure 6. The distributions in (a) MAE$_r$ and (b) RMSE$_r$ for all predictions as the models used the different feature combinations. Features were limited to indoor and outdoor temperatures (T), heating or cooling equipment runtime (R), solar data (S) and fan runtime (F) at time *t*. The median value of the baseline method is indicated across the figure by the dashed line.

performance improvement while the addition of the fan runtime (F) did not. It was our initial impression that the inclusion of fan would not benefit all users, which would reduce its impact on the entire distribution. Based on our anecdotal evidence (Figure 3), we continued to include fan runtime in future investigations (i.e. Section 4.2).

4.2. Adding recent history terms

Based on the performance of the models in Section 4.1, we introduced recent history terms on the feature set of TRSF$_t$ with the black box models. Figure 7 shows the distributions in (a) MAE$_r$ and (b) RMSE$_r$ for feature combinations with increasing amounts of history. We observe that the addition of history lowers the error rates of all models but does begin to increase the magnitude of the outliers of each model. The ridge regression remains the best performing model, with the median MAE$_r$ being approximately 0.1°C better than the other models and feature combinations. Ridge and lasso regression with the inclusion of history up to $t-4$ reduces the median MAE$_r$ by over 20% from when only accounting for features at time *t*. We note a tabular summarization of all features and model combinations is included in the Appendix.

4.3. Best performing model and feature combinations per thermostat

The top performing model and feature sets were found per thermostat using the same MAE$_r$ and RMSE$_r$ data as in Sections 4.1 and 4.2. Figure 8(a) provides counts of the top performing model per thermostat. The ridge regression model is the best performing model based on RMSE$_r$ and MAE$_r$ for the largest number of thermostats. As a set, ridge and lasso regressions are the top performing model for the majority of devices. Seeing that ridge and lasso regression along ARX rank the top three makes sense given their similarities. Since both the lasso and ridge regressions are regularized in training, they should outperform the non-regularized form in ARX during prediction. Similar to the existence of a single top performing model, a single feature set of all indoor and outdoor temperatures, heating or cooling runtimes, solar, and fan runtimes (TRSF$_{t-4:t}$) is the dominant performer (Figure 8(b)). That feature set has the lowest MAE$_r$ and RMSE$_r$ for more than half of the thermostats. It is interesting to note that the second most popular feature set is not just the same parameter set with less history (i.e. TRSF$_{t-3:t}$) but rather TRS$_t$. That could suggest that for some very responsive homes additional history only hurts predictions.

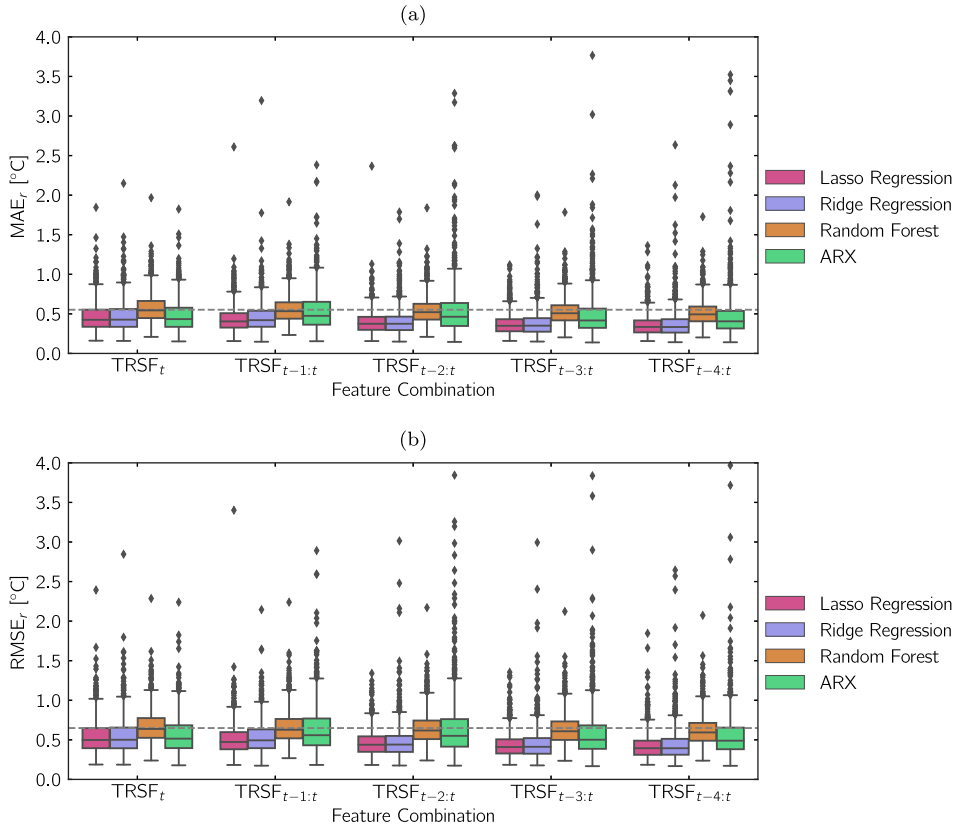

Figure 7. The distributions in (a) MAE$_r$ and (b) RMSE$_r$ for all predictions as the models used increasing lengths of historical data previous to time *t*. Features used were all of the indoor and outdoor temperatures (T), heating or cooling equipment runtime (R), solar data (S), and fan runtime (F). The median value of the baseline method is indicated across the figure by the dashed line.

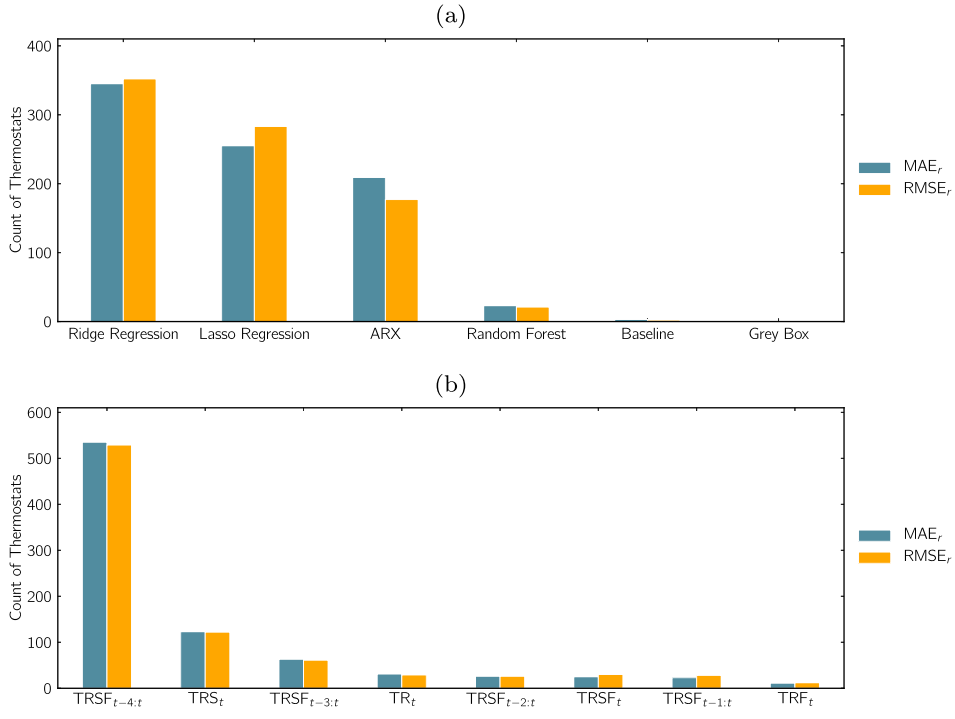

Figure 8. A breakdown of the top performing (a) model and (b) parameter set for each thermostat based on all thermal predictions. For all models and parameters, both RMSE$_r$ and MAE$_r$ are presented.

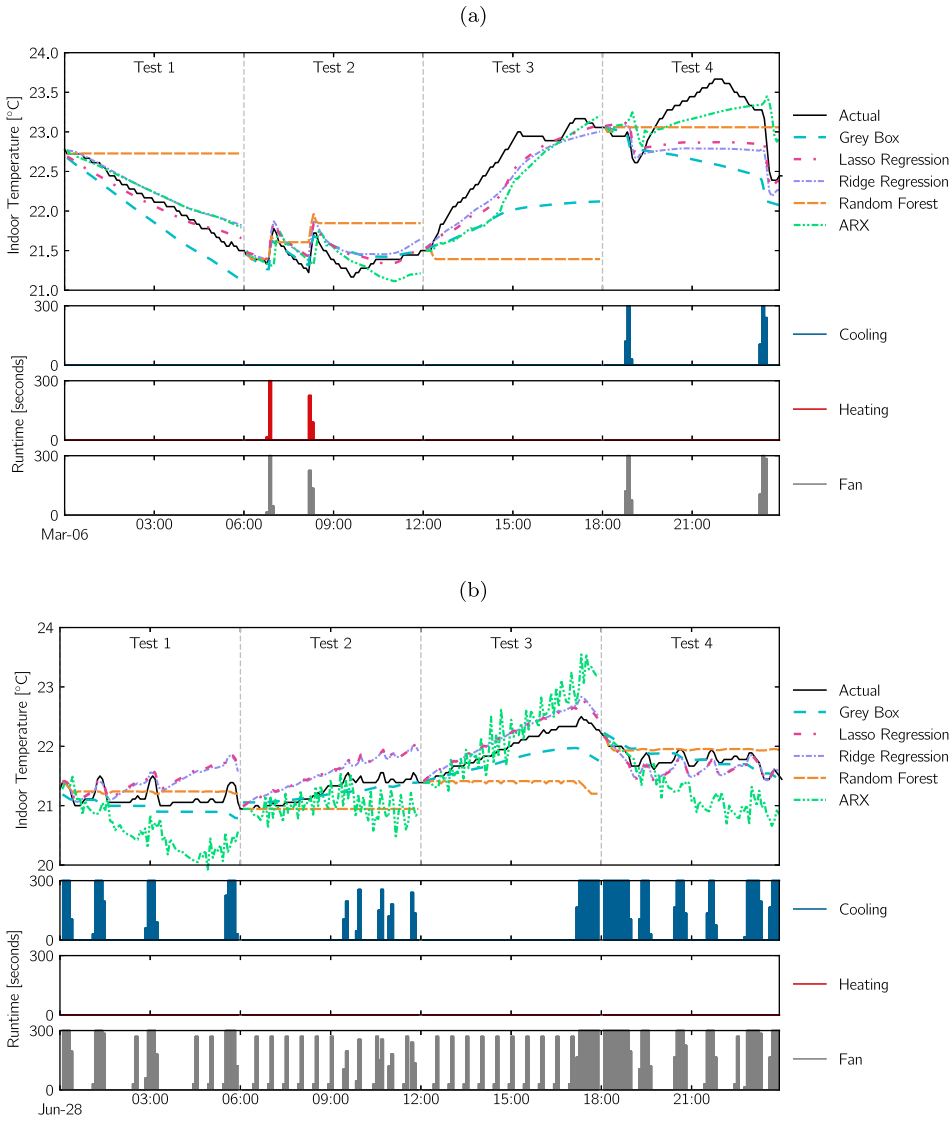

Figure 9. Example predictions for two days with two different thermostats (a and b) for all models except the baseline. The grey box model used TRS_t while all other models used $TRSF_{t-4:t}$. The four six hour tests had the prediction sequences started at 00:00, 06:00, 12:00, and 18:00.

4.4. Demonstration of predictions and error dispersion

The predictions made by each model, given their varied performances (Figures 6 and 7) and varied methods, were expected to differ. Figure 9 shows the prediction progressions for two thermostats (Figures 9(a) and 9(b)) on an *available testing day* for all models except the baseline. Both lasso and ridge regression, the random forest, and the ARX model used $TRSF_{t-4:t}$ feature set. In both figures, some consistent patterns begin to emerge. For example, the lasso and ridge regressions appear to perform similarly and generally track the actual temperature well. Meanwhile, the random forest while reacting to equipment runtime, does not pick up the general decays in temperature. This relatively flat temperature response agrees with our previously observed similarity in performance to the static baseline (Figures 6 and 7). In Figure 9(a) during Test 4, the actual temperature increased substantially and no model was able to predict the change. This situations emphasizes that dynamics in the home are still not fully captured. In Figure 9(b), the ARX model appears to become overly sensitive to the fan runtime, which the other regularized models appear to have avoided.

In addition to the varied predictions made by the different models, each model was expected to have a different error profile. While RMSE and MAE are informative across the population in terms of variation in error and the magnitude of error, a more specific analysis on the variation of absolute error was conducted for a single randomly selected thermostat. Figure 10 shows the distribution in maximum and minimum errors for all the

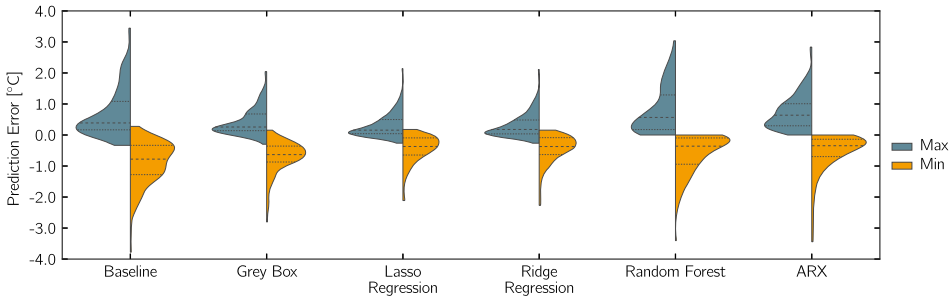

Figure 10. Maximum and minimum prediction error for a single thermostat for all models. The grey box model used TRS_t while all other models used $TRSF_{t-4:t}$. For each violin, the 25th, 50th, and 75th percentile are indicated.

models. The same feature combinations as in Figure 9 were used. For this single thermostat, the distributions in the data reveal that in terms of the absolute worst errors, ARX, random forest, and the baseline have fairly similar performance. Meanwhile, the grey box model had an average error higher than either the lasso or ridge regression but has similar error here in the worst case scenarios. The causes of the largest errors could be attributed to both the challenges in modelling and to stochastic dynamics that are not captured in the models. These dynamics include the number of occupants or the occupant activities in the homes. Users are also free to introduce components to the home which drastically changes how it operates. For instance, they may utilize secondary heating and cooling devices in specific areas of the home or decide to open or close all the blinds.

4.5. Prediction errors based on time of day

Homes are a dynamic space used for a wide range of activities at different times of the day, as such we wanted to know if any general trends appeared in the predictive performance over a day. Results for each thermostats predictions (P) were recalculated based on the starting hour of

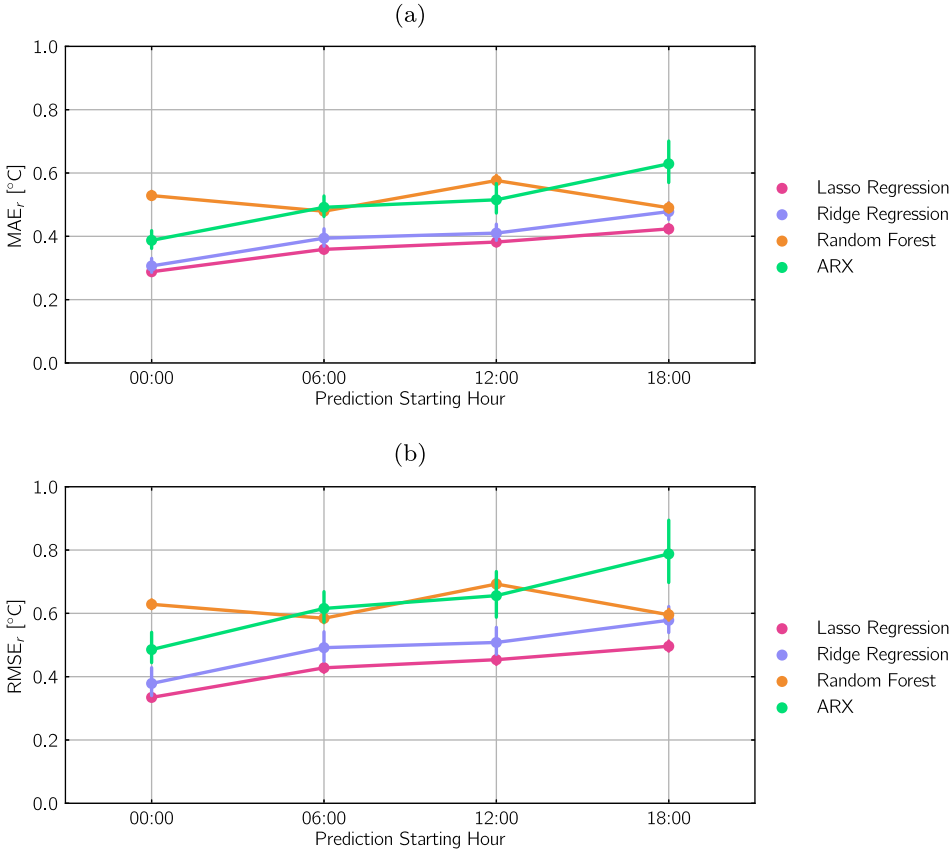

Figure 11. The average of (a) MAE_r and (b) $RMSE_r$ for prediction for all models based on the starting hour of the sequence. The length of the line at each data point represents the standard deviation of each average. All models used the $TRSF_{t-4:t}$ feature set.

the prediction (i.e. 00:00, 06:00, 12:00, 18:00). The effect of time of day on predictions is shown in Figure 11. The best general feature set, based on Figures 6 to 8, of $TRSF_{t-4:t}$ was used along with the three data-driven models. For the ARX, lasso and ridge regressions, the absolute error (Figure 11(a)), and variance in predictions (Figure 11(b)) increase as the day progresses. This increase agrees with the presumed effects of errors caused by the increasing amount and intensity of occupant activities throughout the day. Random forest appears to be the most consistent over the day based on both metrics and in fact it outperforms the ARX during evening (18:00) predictions. This suggests the ARX model overestimates effects from the disturbances as the random forest model was previously observed in Section 4.4 to provide static temperature predictions.

4.6. Housing characteristics

We were interested to know if the quality of model predictions were affected by two specific latent characteristics (i.e. underlying conditions) of the prediction problem. Both the effects of the climate region (Section 4.6.1) and the physical size of the home (Section 4.6.2) were considered. The three data-driven models (i.e. ridge regression, random forest, and ARX) were used with the $TRSF_{t-4:t}$ feature set.

4.6.1. Climate region

Climate region was used as a proxy for regional specifics such as building codes. Figure 12 shows the average (a) MAE_r and (b) $RMSE_r$ for thermostats based on their associated Building America climate zone[2]. For ridge regression, the performance across climate regions in absolute error is fairly consistent. The Marine climate appears to have the lowest absolute error values (MAE_r) while Mixed-Dry/Hot-Dry has the highest error (Figure 12(a)). We postulate the Marine climate's performance is related to the relatively temperate and consistent outdoor temperatures in the colder months (see Figure 13). Meanwhile, the Mixed-Dry/Hot-Dry climate, with its desert-like conditions, has more drastic daily outdoor temperature changes. The ARX model in the Marine climate has the largest standard deviations in

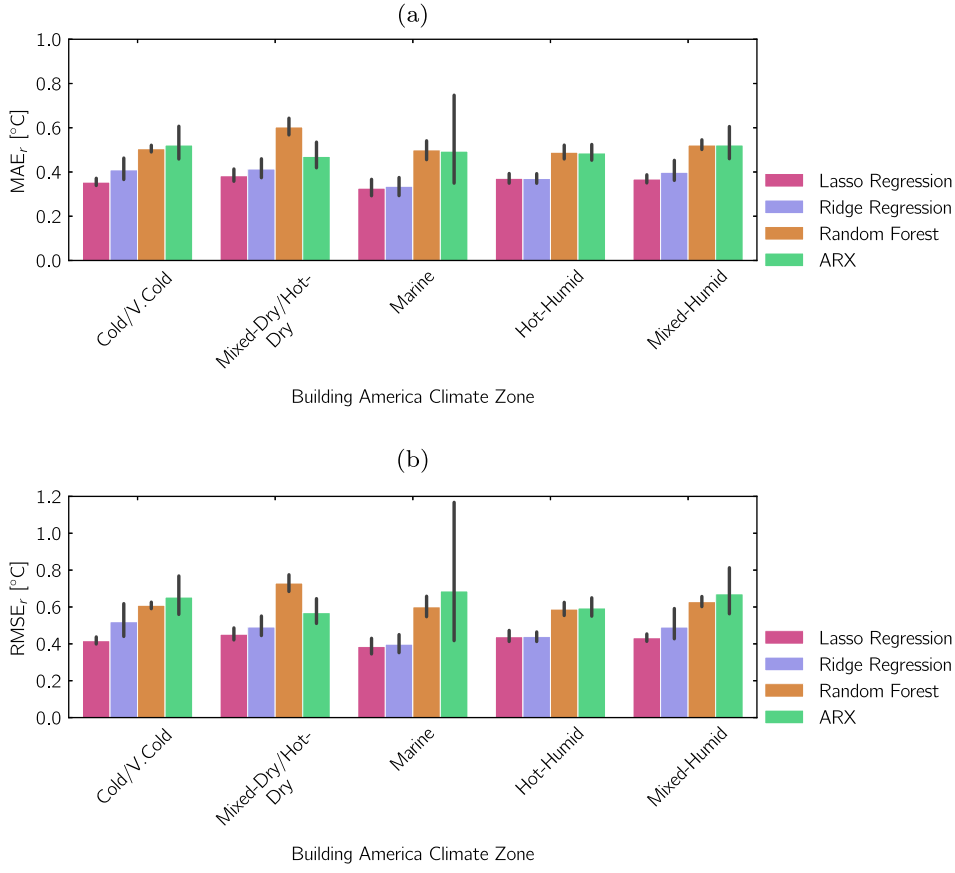

Figure 12. The average (a) MAE_r and (b) $RMSE_r$ for all predictions based on Building America climate zones. Lines on each bar indicate standard deviation of each average.

error. ARX would be susceptible to overfitting in the Marine climate, and a large dispersion in prediction, because of the lack regularization of the parameters and reliance on the exogenous variables other than outdoor air temperature to explain changes in the indoor temperature.

4.6.2. Floor area

Figure 14 presents the average (a) MAE_r and (b) $RMSE_r$ based on the provided housing size for each home. Increasing floor area appears to cause a slight decrease in the magnitude (Figure 14(a)) and variation (Figure 14(b)) in prediction errors. This trend continues up to 4000ft^2 (371.6 m^2) homes. The smaller homes may be condominiums or row housing which could have additional and unobserved energy transfer mechanisms with neighbouring units. These unobserved mechanisms would make the home more difficult to model. The standard deviations increase for homes above 4000ft^2; however, this coincides with considerably smaller sample sizes for these larger homes.

5. Discussion

While our investigation is a useful guide future practitioners developing control models for residential building, we highlight five key points of discussion.

5.1. The relative performance of the grey box model

To clarify reasons for the poor performance of the grey box model, we first note that we had suspected when selecting our models that our grey box model would not outperform all of our data-driven methods. While a physics-based grey box model provides physically meaningful values, the different units of the exogenous variables (Equation (2b)) require unique scaling. With the imposed data restrictions, such as no knowledge of the wall area or orientation, these scaling values are required to be learned and begin to make the formulation more similar to a linear black box model. The grey box model was additionally at a disadvantage as it was not provided either the information of the fan runtime or the recent history terms for any input parameters. As previously stated, the effects of fan runtime on the indoor temperature were found to vary by building. More broadly, the lack of historical terms impacted the performance of the grey box models as the models possessed less complete state information of the entire house. Our initial hypothesis regarding the *1R1C* grey box model being outperformed was supported by the ARX and both ridge and lasso regressions in head-to-head comparisons. The performance of grey box models in general may be improved using other configurations that better match specific homes, by utilizing a richer data set, or by applying more specialized training methods.

5.2. The poor performance of the random forest model

Given the ability of random forest models to handle non-linearity and ability to generally avoid overfitting, we had expected it to perform better in our tests. We suspect that with such rich data provided in the various features, including the near history, the linear models were more than adequate to capture the dynamics in the indoor air temperature. Meanwhile the random forest models had too large a decision space to not overfit to the training data. Finally, the random forest model has an inability to extrapolate from what had been seen in the training set. Given the sliding window used in training, it is possible that the model was hindered by a covariate shift.

5.3. The importance of regularization during regression

The predictive performance of our data-driven models illustrates the importance of proper training methods being applied. The use of regularization was the only difference between ridge regression and the ARX model. With a small number of parameters (i.e. when we only used only variables at time t), the lack of regularization resulted in only minor performance differences between ridge regression and ARX. As we extended the history even longer however, the lack of regularization became more consequential in the relative error rates between to the two black box models. Ultimately the ridge regression was able to have a lower median $RMSE_r$ and MAE_r by almost 0.1°C, or over 20%, in comparison to the ARX model with the same $TRSF_{t-4:t}$ feature set.

5.4. Continued improvement of data-driven models

Since data-driven models are so flexible, there are a number of potential ways to improve them. The first method to improve data-driven model would be to introduce new features. The $TRSF_{t-4:t}$ feature set represents, at present, the majority of observable information on a home's HVAC system. Additional sensors, potentially part of a smart home, could help reduce the amount of system disturbance that is currently treated as random noise. For example, information on the occupant counts in the home, occupant activity, or window/door contacts may help improve predictive accuracy.

The second method to improve the data-driven model's performance would be to extend the historical

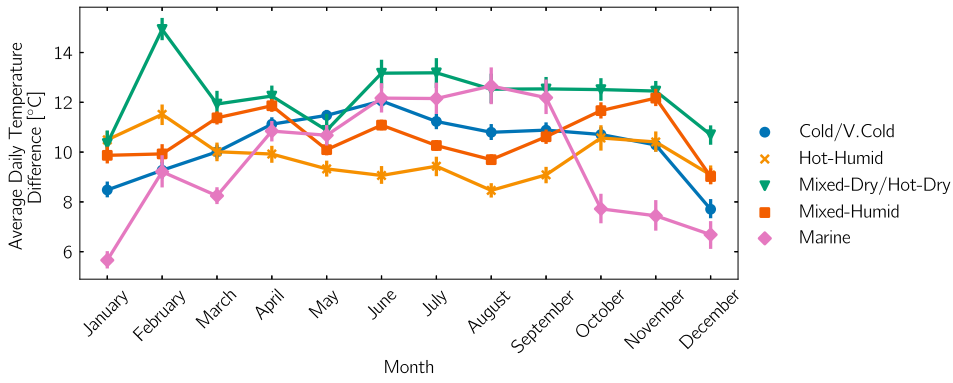

Figure 13. Average daily temperature difference (max-min outdoor temperature) for each month by climate zone. The bars on each point represent the standard deviation of each average.

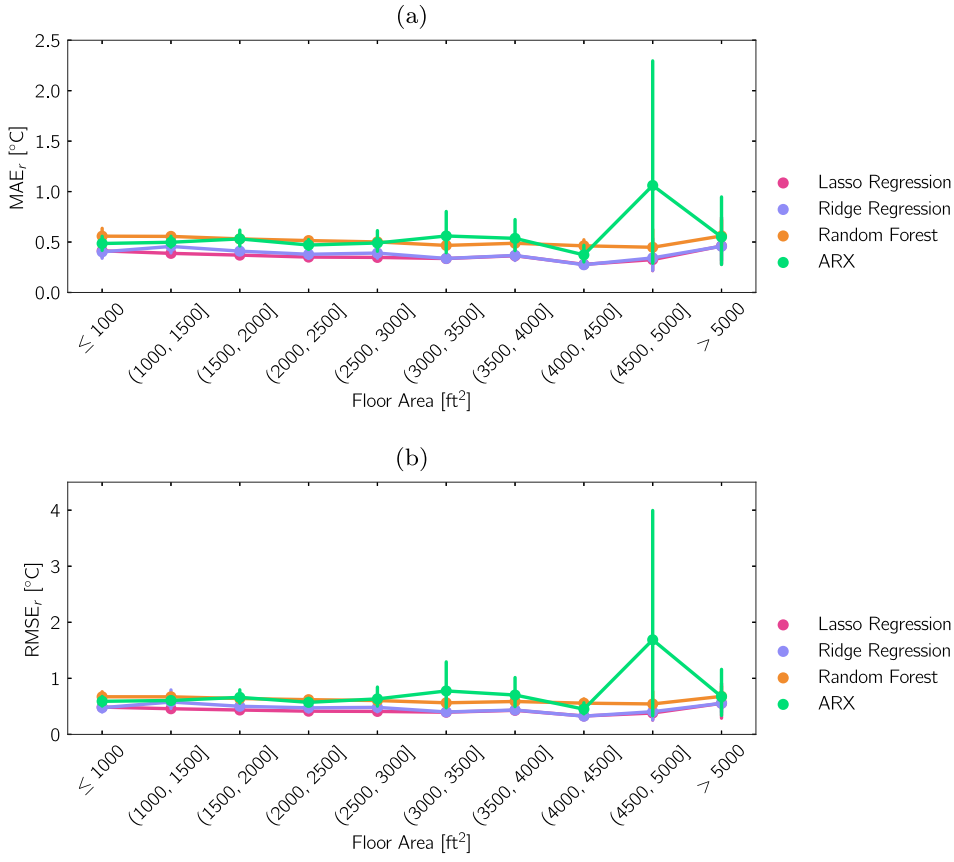

Figure 14. The average (a) MAE$_r$ and (b) RMSE$_r$ based on reported square footage of the home. Bars on each point indicate a standard deviation of the average.

information. We observed that adding recent history provided greater performance improvements than adding additional features. Yet, how much history to provide was shown to be a trade-off. Our results indicated that the benefits of adding more recent history terms of the same form (i.e. a single timestep at a time) had diminishing improvements to both RMSE$_r$ and MAE$_r$ and caused increased outlier magnitudes. This trade-off would be very dependent on the application. For instance, with slow-moving hydronic-based systems or systems using lower temperature air such as heat pumps, additional history beyond $t-4$ may be required. One potential method to address this, without also continuing to drastically increase the number of parameters, would be including features which aggregate multiple timesteps together. For example, the amount of heating or cooling runtime from the previous hour. Alternatively, if automated, the training process could identify the correct lag of features using the autocovariance as shown by Dodier and Henze (2004). However, these lags may not be consistent

and constantly need to be re-evaluated. These alternative features once again illustrate the benefits of the flexibility of data-driven methods.

A final alternative to reduce prediction errors would be to adjust the method of training the models. Since our top-performing model (ridge regression) and feature set ($TRSF_{t-4:t}$) were not universally agreed upon, model selection could be conducted for each home during model training. Model selection allows for the best model and feature set to be used at any time. One approach would be to base the selection of the model and feature set on the lowest cross-validated training errors at the time of training. Applying model selection however would increase the amount of model training happening on devices. The added complication would need to be justified against the size of error or uncertainty from using a 'sub-optimal' model in any application.

5.5. Utilizing the data-driven models in practice

This paper sought to demonstrate that future thermal conditions in homes can be predicted using data and models that can be implemented in residential thermostats. Specifically, we selected and evaluated candidate models with future on-thermostat deployment in mind, where all of the models we evaluated can both train and make predictions with limited computation and a relatively small memory footprint that should be well within the operating parameters of modern smart thermostat devices. Utilization of a model for prediction is reliant on access to outdoor temperature and solar forecasts; the former of which is already present on many smart thermostats and the latter of which could be provided by a forecasting service or estimated with knowledge of the geographic location of the home. The model also requires predicted future control inputs for heating, cooling, and fan runtime, which can be estimated from the device control policy itself or left as decision variables to be optimized by a predictive controller (i.e. in a model predictive control paradigm). Potential implementations of this work may need to consider broader use-cases in which multi-stage or variable capacity equipment is installed. The dynamics of these systems may be adequately captured by the inclusion of more terms for different stages or some indication of the level of operation by the equipment.

6. Conclusion and future work

To properly control residential HVAC to reduce operational costs and energy usage, predictive thermal models need to be developed for each home based on their available data. We utilized a sample of 1000 thermostats from a large and diverse dataset of representative smart thermostat data to compare and evaluate the performance of candidate thermal models. We determined a ridge regression utilizing recent history of 20 minutes for air temperatures, heating and cooling runtime, fan runtime, and solar data was the single best performer for our sample of thermostats. The method outperformed strong data-driven alternatives such as auto-regressive models with exogenous variables, random forest, and a *1R1C* configuration of a grey box model. In addition to its strong predictive performance, the ridge regression is an ideal solution for deploying to devices given its simplicity of training and predicting.

In the future we seek to address a number of remaining questions regarding how to maximize the utility of our available data. For example, we did not investigate the minimum number of days required to train an acceptable model or whether a prescribed number of training days should be replaced with a minimum amount of equipment runtime being observed before training. Determining the appropriate amount of training data is dependent on the application at-hand and corresponding accuracy requirement. To reduce computational loads and data requirements, the length of time a model could be reused should be assessed. Finally, consideration should be given to accessing more relevant features and more detailed data. As we mentioned earlier, the features we included capture the majority of the currently available observational data in a home. In the future, new basic information could drastically improve how a thermal model may be incorporated into control. For instance, knowledge of whether windows and doors are open could improve model accuracy and appropriate decision making.

Notes

1. https://nsrdb.nrel.gov
2. https://basc.pnnl.gov

Acknowledgments

This work is funded by a research grant provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and ecobee Inc. Two of the authors are active participants of International Energy Agency Energy in Buildings and Communities (IEA-EBC) Programme Annex 79.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Natural Sciences and Engineering Research Council of Canada [CRDPJ 508857-17].

ORCID

Brent Huchuk http://orcid.org/0000-0002-7720-7830
Scott Sanner http://orcid.org/0000-0001-7984-8394
William O'Brien http://orcid.org/0000-0002-0236-5383

References

Afram, Abdul, Alan S. Fung, Farrokh Janabi-Sharifi, and Kaamran Raahemifar. 2018a. "Development and Performance Comparison of Low-Order Black-Box Models for a Residential HVAC System." *Journal of Building Engineering* 15: 137–155.

Afram, Abdul, Alan S. Fung, Farrokh Janabi-Sharifi, and Kaamran Raahemifar. 2018b. "Development of An Accurate Grey-box Model of Ubiquitous Residential HVAC System for Precise Performance Prediction During Summer and Winter Seasons and Improved Control System Design." *Energy and Buildings* 171: 168–182. http://linkinghub.elsevier.com/retrieve/pii/S0378778816310817.

Afram, Abdul, and Farrokh Janabi-Sharifi. 2015a. "Black-box Modeling of Residential HVAC System and Comparison of Gray-box and Black-box Modeling Methods." *Energy and Buildings* 94: 121–149. http://dx.doi.org/10.1016/j.enbuild.2015.02.045.

Afram, Abdul, and Farrokh Janabi-Sharifi. 2015b. "Gray-box Modeling and Validation of Residential HVAC System for Control System Design." *Applied Energy* 137: 134–150. http://dx.doi.org/10.1016/j.apenergy.2014.10.026.

Afram, Abdul, and Farrokh Janabi-Sharifi. 2017. "Supervisory Model Predictive Controller (MPC) for Residential HVAC Systems: Implementation and Experimentation on Archetype Sustainable House in Toronto." *Energy and Buildings* 154: 268–282. http://linkinghub.elsevier.com/retrieve/pii/S0378778817301743.

Afram, Abdul, Farrokh Janabi-Sharifi, Alan S. Fung, and Kaamran Raahemifar. 2017. "Artificial Neural Network (ANN) Based Model Predictive Control (MPC) and Optimization of HVAC Systems: A State of the Art Review and Case Study of a Residential HVAC System." *Energy and Buildings* 141: 96–113. http://linkinghub.elsevier.com/retrieve/pii/S0378778816310799, http://dx.doi.org/10.1016/j.enbuild.2017.02.012.

Alibabaei, Nima, Alan S. Fung, and Kaamran Raahemifar. 2016. "Development of Matlab-TRNSYS Co-Simulator for Applying Predictive Strategy Planning Models on Residential House HVAC System." *Energy and Buildings* 128: 81–98. http://www.sciencedirect.com/science/article/pii/S0378778816304716.

Alibabaei, Nima, Alan S. Fung, Kaamran Raahemifar, and Arash Moghimi. 2017. "Effects of Intelligent Strategy Planning Models on Residential HVAC System Energy Demand and Cost During the Heating and Cooling Seasons." *Applied Energy* 185: 29–43. https://linkinghub.elsevier.com/retrieve/pii/S0306261916315082.

Ashtiani, Arya, Parham A. Mirzaei, and Fariborz Haghighat. 2014. "Indoor Thermal Condition in Urban Heat Island: Comparison of the Artificial Neural Network and Regression Methods Prediction." *Energy and Buildings* 76: 597–604. http://dx.doi.org/10.1016/j.enbuild.2014.03.018.

Athienitis, A. K., M. Stylianou, and J. Shou. 1990. "Methodology for Building Thermal Dynamics Studies and Control Applications." *ASHRAE Transactions*, 839–848.

Baasch, Gaby, Adam Wicikowski, Gaëlle Faure, and Ralph Evins. 2019. "Comparing Gray Box Methods to Derive Building Properties from Smart Thermostat Data." In *BuildSys 2019 – Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, New York, USA, 223–232.

Bacher, Peder, and Philip Hvidthøft Delff Andersen. 2014. *IEA Common Exercise 4: ARX, ARMAX and Grey-box Models for Thermal Performance Characterization of the Test Box*. Technical Report. Technical University of Denmark. https://orbit.dtu.dk/en/publications/iea-common-exercise-4-ar x-armax-and-grey-box-models-for-thermal-p.

Brastein, O. M., D. W. U. Perera, C. Pfeifer, and N. O. Skeie. 2018. "Parameter Estimation for Grey-box Models of Building Thermal Behaviour." *Energy and Buildings* 169: 58–68. http://linkinghub.elsevier.com/retrieve/pii/S0378778817331791.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. http://link.springer.com/10.1023/A:1010933404324.

Burger, Eric M., and Scott J. Moura. 2016. "Recursive Parameter Estimation of Thermostatically Controlled Loads Via Unscented Kalman Filter." *Sustainable Energy, Grids and Networks* 8: 12–25.

Burger, Eric M., and Scott J. Moura. 2017. "ARX Model of a Residential Heating System with Backpropagation Parameter Estimation Algorithm." *In Proceedings of the ASME 2017 Dynamic Systems and Control Conference. Volume 3: Vibration in Mechanical Systems; Modeling and Validation; Dynamic Systems and Control Education; Vibrations and Control of Systems; Modeling and Estimation for Vehicle Safety and Integrity; Modeling and Control of IC Engines and Aftertreatment Systems; Unmanned Aerial Vehicles (UAVs) and Their Applications; Dynamics and Control of Renewable Energy Systems; Energy Harvesting; Control of Smart Buildings and Microgrids; Energy Systems, Tysons*, USA.

Candanedo, José A., Vahid R. Dehkordi, and Justin Tamasauskas. 2015. "Impact of Heat Pumps and Predictive Control on Residential Energy Use, Load and Grid Interaction: A Canadian Perspective." In *Proceedings 14th International Conference of IBPSA – Building Simulation 2015, BS 2015*, Hyderabad, India, 1954–1961.

Cetin, Kristen S., Lance Manuel, and Atila Novoselac. 2016. "Effect of Technology-Enabled Time-of-Use Energy Pricing on Thermal Comfort and Energy Use in Mechanically-Conditioned Residential Buildings in Cooling Dominated Climates." *Building and Environment* 96: 118–130. http://linkinghub.elsevier.com/retrieve/pii/S036013231530175X.

Chen, Yuxiang, A. K. Athienitis, and Khaled Galal. 2010. "Modeling, Design and Thermal Performance of a BIPV/T System Thermally Coupled with a Ventilated Concrete Slab in a Low Energy Solar House: Part 1, BIPV/T System and House Energy Concept." *Solar Energy* 84 (11): 1892–1907. http://dx.doi.org/10.1016/j.solener.2010.06.013.

Chen, Yuxiang, Khaled Galal, and A. K. Athienitis. 2010. "Modeling, Design and Thermal Performance of a BIPV/T System Thermally Coupled with a Ventilated Concrete Slab in a Low Energy Solar House: Part 2, Ventilated Concrete Slab." *Solar Energy* 84 (11): 1908–1919. http://dx.doi.org/10.1016/j.solener.2010.06.012.

Clarke, Joseph A. 2007. *Energy Simulation in Building Design*. Oxford: Butterworth-Heinemann.

Cole, Wesley J., Kody M. Powell, Elaine T. Hale, and Thomas F. Edgar. 2014. "Reduced-order Residential Home Modeling for Model Predictive Control." *Energy and Buildings* 74: 69–77. http://dx.doi.org/10.1016/j.enbuild.2014.01.033.

Coley, D. A., and J. M. Penman. 1996. "Simplified Thermal Response Modelling in Building Energy Management. Paper III: Demonstration of a Working Controller." *Building and Environment* 31 (2): 93–97.

Dimitriou, Vanda, Steven K. Firth, Tarek M. Hassan, Tom Kane, and Michael Coleman. 2015. "Data-driven Simple Thermal Models: The Importance of the Parameter Estimates." *Energy Procedia*78: 2614–2619. http://dx.doi.org/10.1016/j.egypro.2015.11.322.

Dodier, Robert H., and Gregor P. Henze. 2004. "Statistical Analysis of Neural Networks As Applied to Building Energy Prediction." *Journal of Solar Energy Engineering, Transactions of the ASME* 126 (1): 592–600.

Doiron, Matt, William O'Brien, and Andreas Athienitis. 2011. "Energy Performance, Comfort, and Lessons Learned From a Near Net Zero Energy Solar House." *ASHRAE Transactions* 117 (2): 585–596.

Dong, Bing, and Khee Poh Lam. 2014. "A Real-Time Model Predictive Control for Building Heating and Cooling Systems Based on the Occupancy Behavior Pattern Detection and Local Weather Forecasting." *Building Simulation* 7 (1): 89–106. http://link.springer.com/10.1007/s12273-013-0142-7.

ecobee Inc. 2020. "Donate Your Data." https://www.ecobee.com/donateyourdata/.

Hossain, Md Monir, Tianyu Zhang, and Omid Ardakanian. 2019. "Evaluating the Feasibility of Reusing Pre-Trained Thermal Models in the Residential Sector." In *UrbSys 2019 – Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization, Part of BuildSys 2019*, New York, USA, 23–32.

Hu, Sanhe. 2019. "uszipcode." https://uszipcode.readthedocs.io/index.html.

Huang, Joe, James Hanford, Fuqiang Yang, Environmental Energy, Technologies Division, and Lawrence Berkeley. 1999. "Residential Heating and Cooling Loads Component Analysis." Technical Report November. Lawrence Berkeley National Laboratory.

Huchuk, Brent, William O'Brien, and Scott Sanner. 2018. "A Longitudinal Study of Thermostat Behaviors Based on Climate, Seasonal, and Energy Price Considerations Using Connected Thermostat Data." *Building and Environment* 139: 199–210. http://www.sciencedirect.com/science/article/pii/S0360132318302634.

Huchuk, Brent, Scott Sanner, and William O'Brien. 2019 May. "Comparison of Machine Learning Models for Occupancy Prediction in Residential Buildings Using Connected Thermostat Data." *Building and Environment* 160: 106–177. https://doi.org/10.1016/j.buildenv.2019.106177.

Jin, Xin, Kyri Baker, Dane Christensen, and Steven Isley. 2017. "Foresee: A User-centric Home Energy Management System for Energy Efficiency and Demand Response." *Applied Energy* 205: 1583–1595. https://linkinghub.elsevier.com/retrieve/pii/S0306261917311856.

Klein, S. A. 2017. "TRNSYS 18: A Transient System Simulation Program." *Solar Energy Laboratory*, University of Wisconsin, Madison. http://sel.me.wisc.edu/trnsys.

Ławryńczuk, Maciej, and Paweł Ocłoń. 2019. "Model Predictive Control and Energy Optimisation in Residential Building with Electric Underfloor Heating System." *Energy* 182: 1028–1044.

Madsen, H., and J. Holst. 1995. "Estimation of Continuous-Time Models for the Heat Dynamics of a Building." *Energy and Buildings* 22 (1): 67–79.

Missaoui, Rim, Hussein Joumaa, Stephane Ploix, and Seddik Bacha. 2014. "Managing Energy Smart Homes According to Energy Prices: Analysis of a Building Energy Management System." *Energy and Buildings* 71: 155–167. https://www-sciencedirect-com.myaccess.library.utoronto.ca/science/article/pii/S0378778813008335.

Molina, Diogenes, Coby Lu, Viktoriya Sherman, and Ronald G. Harley. 2013. "Model Predictive and Genetic Algorithm-based Optimization of Residential Temperature Control in the Presence of Time-Varying Electricity Prices." *IEEE Transactions on Industry Applications* 49 (3): 1137–1145.

Moon, Jin Woo, and Jong-Jin Kim. 2010. "ANN-Based Thermal Control Models for Residential Buildings." *Building and Environment* 45 (7): 1612–1625. http://linkinghub.elsevier.com/retrieve/pii/S0360132310000211 http://www.sciencedirect.com/science/article/pii/S0360132310000211.

National Renewable Energy Laboratory. 2018. "National Solar Radiation Database."

Natural Resources Canada. 2019. "Canada's Secondary Energy Use."

New, Joshua, Mark Adams, Piljae Im, Hsuihan Lexie Yang, Joshua Hambrick, William Copeland, Lilian Bruce, and James A Ingraham. 2018. Oak Ridge National Lab. Oak Ridge, USA.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. "Scikit-Learn: Machine Learning in Python." Journal of Machine Learning Research 12: 2825–2830.

Peffer, Therese, Marco Pritoni, Alan Meier, Cecilia Aragon, and Daniel Perry. 2011. "How People Use Thermostats in Homes: A Review." *Building and Environment* 46 (12): 2529–2541. http://linkinghub.elsevier.com/retrieve/pii/S0360132311001739.

Prívara, Samuel, Jiří Cigler, Zdeněk Váňa, Frauke Oldewurtel, Carina Sagerschnig, and Eva Žáčeková. 2013. "Building Modeling As a Crucial Part for Building Predictive Control." *Energy and Buildings* 56: 8–22.

Reddy, T. A., L. K. Norford, and W. Kempton. 1991. "Shaving Residential Air-Conditioner Electricity Peaks by Intelligent Use of The Building Thermal Mass." *Energy* 16 (7): 1001–1010.

Safa, Amir A., Alan S. Fung, and Rakesh Kumar. 2015. "Performance of Two-Stage Variable Capacity Air Source Heat Pump: Field Performance Results and TRNSYS Simulation." *Energy and Buildings*94: 80–90. http://dx.doi.org/10.1016/j.enbuild.2015.02.041.

Sanyal, Jibonananda, Joshua New, and Richard Edwards. 2013. "Supercomputer Assisted Generation of Machine Learning Agents for the Calibration of Building Energy Models." In *Proceedings of the conference on extreme science and engineering discovery environment: gateway to discovery*. New York, USA.

Surles, William, and Gregor P. Henze. 2012. "Evaluation of Automatic Priced Based Thermostat Control for Peak Energy

Reduction Under Residential Time-of-use Utility Tariffs." *Energy and Buildings* 49: 99–108. http://dx.doi.org/10.1016/j.enbuild.2012.01.042.

Taylor, Ross. 2016. "PyFlux: An Open Source Time Series Library for Python." https://pyflux.readthedocs.io/en/latest/.

U.S. Department of Energy. 2020. "EnergyPlus." https://energyplus.net.

U.S. Energy Information Administration. 2015. "Residential Energy Consumption Survey (RECS)." Washington, DC. https://www.eia.gov/consumption/residential/data/2015/.

U.S. Energy Information Administration. 2019a. "Use of Energy Explained." https://www.eia.gov/energyexplained/use-of-energy/.

U.S. Energy Information Administration. 2019b. "Use of Energy Explained: Energy Use in Homes." https://www.eia.gov/energyexplained/use-of-energy/homes.php.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, P. Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, Ilhan Polat,Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17: 261–272.

Wang, Zequn, and Yuxiang Chen. 2019. "Data-driven Modeling of Building Thermal Dynamics: Methodology and State of the Art." *Energy and Buildings* 203: 109405.

Wang, Junke, Choon Yik Tang, Michael R. Brambley, and Li Song. 2019. "Predicting Home Thermal Dynamics Using a Reduced-Order Model and Automated Real-time Parameter Estimation." *Energy and Buildings* 198: 305–317. https://doi.org/10.1016/j.enbuild.2019.06.002.

Wang, Junke, Choon Yik Yik Tang, and Li Song. 2020. "Design and Analysis of Optimal Pre-Cooling in Residential Buildings." *Energy and Buildings* 216: 109951. https://doi.org/10.1016/j.enbuild.2020.109951.

Zeifman, Michael, Amine Lazrak, and Kurt Roth. 2019. "Residential Retrofits At Scale: Opportunity Identification, Saving Estimation, and Personalized Messaging Based on Communicating Thermostat Data." *Energy Efficiency* 13: 393–405.

Appendix

Table A1. Tabulated average MAE_r and standard deviations (in brackets) for each model and parameter configuration.

	Baseline	Grey box	Lasso regression	Ridge regression	Random forest	ARX
TR_t	0.59 (0.20)	0.69 (0.36)	0.45 (0.18)	0.45 (0.18)	0.57 (0.18)	0.51 (0.19)
TRS_t	–	0.59 (0.35)	0.43 (0.17)	0.43 (0.17)	0.57 (0.18)	0.44 (0.18)
TRF_t	–	–	0.48 (0.19)	0.49 (0.21)	0.58 (0.19)	0.52 (0.24)
$TRSF_t$	–	–	0.46 (0.19)	0.47 (0.20)	0.58 (0.19)	0.48 (0.20)
$TRSF_{t-1:t}$	–	–	0.43 (0.17)	0.46 (0.20)	0.56 (0.18)	0.55 (0.37)
$TRSF_{t-2:t}$	–	–	0.40 (0.20)	0.42 (0.25)	0.55 (0.18)	0.54 (0.33)
$TRSF_{t-3:t}$	–	–	0.38 (0.21)	0.41 (0.39)	0.53 (0.17)	0.49 (0.29)
$TRSF_{t-4:t}$	–	–	0.36 (0.14)	0.40 (0.33)	0.52 (0.17)	0.51 (0.56)

Table A2. Tabulated average $RMSE_r$ and standard deviations (in brackets) for each model and parameter configuration.

	Baseline	Grey box	Lasso regression	Ridge regression	Random forest	ARX
TR_t	0.69 (0.24)	0.81 (0.42)	0.53 (0.20)	0.53 (0.21)	0.67 (0.22)	0.60 (0.22)
TRS_t	–	0.69 (0.40)	0.50 (0.20)	0.51 (0.20)	0.67 (0.22)	0.52 (0.21)
TRF_t	–	–	0.56 (0.23)	0.57 (0.24)	0.67 (0.22)	0.61 (0.28)
$TRSF_t$	–	–	0.54 (0.22)	0.55 (0.24)	0.67 (0.22)	0.57 (0.24)
$TRSF_{t-1:t}$	–	–	0.51 (0.20)	0.54 (0.24)	0.66 (0.21)	0.65 (0.51)
$TRSF_{t-2:t}$	–	–	0.48 (0.32)	0.50 (0.42)	0.65 (0.21)	0.65 (0.41)
$TRSF_{t-3:t}$	–	–	0.45 (0.35)	0.49 (0.74)	0.64 (0.20)	0.60 (0.37)
$TRSF_{t-4:t}$	–	–	0.43 (0.17)	0.49 (0.62)	0.63 (0.20)	0.64 (0.90)