NRC·CNRC

NRC Publications Archive
Archives des publications du CNRC

Climate resilience buildings: overheating detection in houses-investigating the feasibility of using smart thermostat data

Bahiraei, F.; Laouadi, A.

For the publisher's version, please access the DOI link below./ Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

https://doi.org/10.4224/40002659

NRC Publications Archive Record / Notice des Archives des publications du CNRC :
https://nrc-publications.canada.ca/eng/view/object/?id=de0938a3-0b18-4e20-b279-a6909ef479d2
https://publications-cnrc.canada.ca/fra/voir/objet/?id=de0938a3-0b18-4e20-b279-a6909ef479d2

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at
https://nrc-publications.canada.ca/eng/copyright
READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site
https://publications-cnrc.canada.ca/fra/droits
LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

National Research Council Canada | Conseil national de recherches Canada

Canada

NRC·CNRC CONSTRUCTION

Climate Resilience Buildings:
Overheating detection in houses-
Investigating the feasibility of using smart thermostat data

Authors: Bahiraei F., Laouadi A.

Report No.: A1-017320.1
Report Date: 1 April 2021
Contract No.: A1-01732010-10
Agreement Date: 15 June 2020

National Research Council Canada | Conseil national de recherches Canada

Canada

Climate Resilience Buildings:

Overheating detection in houses-Investigating the feasibility of using smart thermostat data

Author

Farid Bahiraei. Research associate

Co-authors Laouadi, Abdelaziz. Senior Research Officer

Approved

Fiona Hill

Report No: A1-017320.1
Report Date: 1 April 2021
Contract No: A1-01732010-10
Agreement date: 15 June 2020
Program: BRAMA

40 pages

Copy no. 1 of X

This report may not be reproduced in whole or in part without the written consent of the National Research Council Canada and the Client.

(This page is intentionally left blank)

TABLE OF CONTENTS

Table of Contents i
List of Figures ii
List of Tables iii
Executive Summary 1
Section A. Analysis of overheating in residential buildings using smart thermostat data 2
A.1 Introduction 2
A.1.1 Background and previous work 2
A.1.2 Motivation and objectives 4
A.2 Indoor overheating analysis methodology 5
A.2.1 Definition of overheating events 6
A.2.2 selection of threshold values of t-SET 7
A.2.3 Body water loss estimation 8
A.3 Building Data collection and analysis 8
A.3.1 Data description 8
A.3.2 Daytime and nighttime indoor conditions 9
A.3.3 Building characteristics 11
A.3.4 Data availability 12
A.4 Results and discussion 14
A.4.1 Indoor temperature variation 14
A.4.2 overheating analysis 18
A.4.3 Limitations of the current study 21
A.5 Conclusions and future work 22
Section B. A data-driven indoor overheating warning framework for residential buildings 23
B.1 Introduction 23
B.2 Data description 24
B.3 Forecasting models 25
B.4 Indoor temperature models 26
B.5 Parameter estimation for gray box models 27
B.6 Indoor relative humidity models 28
B.7 Results 29

B.8 Discussion 35
B.9 Conclusions 36
Acknowledgments 36
References 37

LIST OF FIGURES

Figure 1. Distribution of number remote sensors for the selected homes 10
Figure 2. Sample of the hourly thermostat temperature data (top) and sensor temperatures (bottom) for a single thermostat 11
Figure 3. Distribution of buildings construction age 12
Figure 4. Distribution of building type 12
Figure 5. Distribution of building-days for each area 14
Figure 6. Maximum daily indoor control temperatures based on the 7-day prevailing outdoor temperature for homes without AC (top) and with AC (bottom) in all metropolitan areas. The 90% and 80% comfort acceptability limits from ASHRAE-55 adaptive comfort model are shown for comparison 16
Figure 7. Outdoor monthly average temperature in the selected areas 17
Figure 8. Variation of indoor temperature, outdoor temperature and SET during two observed heat events (2018) 20
Figure 9. Percentage of houses with at least one overheating event that is associated with health risks to elderly occupants 21
Figure 10. Selected houses distribution 25
Figure 11. MAE distribution of a) 2R2C; and b) 3R2C grey box models for various training windows 29
Figure 12. MAPE distribution of 2R2C (left) and 3R2C (right) based forecast models for various training windows 31
Figure 13.Indoor and outdoor hourly temperatures based on the selected model for various houses 33
Figure 14. MAE and R-squared distribution of indoor relative humidity model 34
Figure 15. Variation of actual and forecasted hourly indoor relative humidity for various houses 35

LIST OF TABLES

Table 1. Suggested threshold values of t-SETd and t-SETn for vulnerable occupants 7
Table 2.Number of selected homes in each area 9
Table 3. Total number of building-days in different areas 14
Table 4. Air conditioning use in selected Canadian provinces 17
Table 5. Frequency of overheating in homes with and without AC units 19
Table 6. Performance of 2R2C-based model with 96-h training window 31
Table 7. Characteristics of the selected houses 32

EXECUTIVE SUMMARY

This report is the final deliverable of the project A1-017320.1. The main purpose of this report is to investigate the feasibility of using one of the emerging data sources, i.e. smart thermostats, along with machine learning approaches to detect and quantify overheating in buildings. This report is organized in two sections. Section A represents the results of the indoor overheating analysis in residential buildings based on smart thermostat data. In section B, a data-driven indoor conditions warning framework for residential buildings is developed and assessed.

In section A, indoor temperature and humidity measurements from more than 3,000 connected thermostats, during summers of 2016-2019 and from homes located in five major metropolitan areas in Canada were used. Furthermore, a robust procedure to evaluate the frequency and severity of indoor overheating based on the heat-related health outcomes for older people was utilized. In particular, the overheating occurrence in homes with and without central air conditioning (AC) units was comparatively evaluated. The results of the first section showed that 12% of houses under study experienced at least one overheating event. Furthermore, the extreme indoor overheating events with potential health risks to the vulnerable occupants were more common amongst houses without central air conditioning units across the selected cities.

In section B, a feasibility study was done on more than 180 houses located in Ontario that do not use central AC units. In this approach, two gray box models were developed based on the hourly indoor and outdoor temperatures. The gray box models were implemented in a recursive forecasting strategy using sliding training and forecast time periods to provide the forecasted hourly indoor temperatures. The proposed models were able to forecast 12-hour ahead hourly indoor conditions in 92% of houses with less than 5% mean absolute percentage error. This indicated the potential for leveraging data from IoT devices and machine learning, as part of an overheating detection and warning system in buildings. Although this work was based on data from residential buildings, the methodologies developed here can be used for other buildings.

SECTION A. ANALYSIS OF OVERHEATING IN RESIDENTIAL BUILDINGS USING SMART THERMOSTAT DATA

A.1 Introduction

Indoor overheating may occur in different built environments during extreme heat events. Excess indoor heat affects the health and wellbeing of occupants, especially the vulnerable subjects. It is shown that the elderly are less sensitive to thirst and heat, and therefore are less likely to drink enough water which puts them at a higher risk of dehydration during long exposures to high indoor temperatures [1]. The 2003 heatwave in in eight European countries caused around 15,000 premature deaths [2] and during the July 2010 heatwave in Québec, Canada the crude death rates showed a significant increase of 33% in deaths [3]. The crude death rate was considered as the number of deaths occurring among the population of a given geographical area during a given year, per 1,000 mid-year total population of the given geographical area during the same year.

In Canada, the need to implement mitigation strategies have been driving the demand for more energy efficient residential buildings. As a result, highly-insulated and airtight houses are currently being built. Additionally, existing dwellings are being retrofitted to reduce winter space heating demand and associated greenhouse gas emissions. However, there is evidence that new and highly insulated dwellings in Europe are more disposed to overheating than older homes [4], [5]. Furthermore, as Canada is shifting towards a super-aged country [6] and older people are being increasingly encouraged to 'age in place' and remain in the family home, the risk of indoor heat exposure to the seniors is increasing. In addition, temperatures and heatwave frequency are predicted to increase due to climate change. These factors have contributed to a growing concern about an increased occurrence of overheating during the warm summer weather in Canadian homes.

A.1.1 Background and previous work

Studies of heat-related illness and death commonly use outdoor conditions in order to determine the association between ambient heat exposure and illness or deaths. Outdoor heat exposure is often estimated using an airport monitoring station and applied to residents of an entire community. However, people spend up to 90% of their time indoors, especially elderly people who have been shown to be more vulnerable to excess heat [7]. Hence, indoor

conditions are likely to better represent heat exposure of vulnerable individuals [8]. There have been several studies in which indoor condition measurements have been conducted to evaluate overheating events in different types of residential buildings. Loughnan et al. [9] performed indoor measurements and occupant surveys in 20 houses with elderly occupants in a regional town in Australia during the summer of 2012. They found that the indoor maximum temperature was not significantly correlated with the outdoor temperature in houses with air-conditioning (AC) units. Surveys showed residents were comfortable at indoor temperatures up to 26.6°C. Significant differences in the indoor temperature were observed and related to the house characteristics and operation. White-Newsome et al. [10] measured indoor temperatures during the summer of 2009 in 30 residential buildings occupied by seniors in Detroit, Michigan. In this study, about 50% of the homes had air conditioning. They reported that the average maximum indoor temperature was 34.9°C, which was 13.8°C higher than the average maximum outdoor temperature. The authors used four different linear regression models to investigate variables explaining indoor temperature trends and found strong correlation with outdoor temperature, solar radiation, and dew point temperature. Beizaee et al. [4] conducted an indoor conditions study during the cool summer of 2007 in 207 homes across England. Among the dwellings, 193 were not air conditioned. They found that despite the cool summer, 21% of the bedrooms were overheated based on the adaptive thermal comfort standard of BS/EN15251. More specifically, bedrooms of homes built after 1990 were significantly warmer and those built before 1919 were significantly cooler.

Quinn et al. [11] conducted indoor temperature measurements during the summer of 2003 to 2011 in 285 homes in New York City, USA. Information on the use of air conditioning was not collected, but the New York City surveys indicated that 13% of households did not use air conditioning in their homes. The Heat Index was used as a measure of heat-related stress. The results showed that there was a direct link between the indoor and outdoor conditions, but the indoor Heat Index metric showed a significant inconsistency among the studied homes. It was found that the buildings characteristics and how the occupants managed the indoor conditions, may pose additional risks to vulnerable people during the heat events. Van Loenhout et al. [12] performed surveys and field measurements in 113 homes for the elderly in Netherlands during the summer of 2012. The study found that the indoor temperature varied among the homes, and was strongly related to the known health problems of the elderly occupants. During the warmest week of the study, the temperatures of living rooms and bedrooms were higher than the outdoor temperature despite the use of air conditioning.

More than half of the occupants perceived their indoor spaces as too warm, and reported symptoms such as thirst, sleep disturbance and excessive sweating.

In another study, Vellei et al. [13] conducted surveys and measured indoor temperature and air quality data over two years (2014-2015) for vulnerable and non-vulnerable occupants. The measurements were performed in 68 homes located in Exeter, UK, where overheating is not common. The surveys showed that vulnerable occupants felt cooler compared to non-vulnerable occupants. The study found that overheating, according to the criteria of CIBSE – Guide A [14], was found to occur, particularly in buildings with vulnerable occupants. Pathan et al. [15] conducted occupant surveys and indoor temperature measurements to assess overheating in 122 dwellings in London, UK during the summer of 2009 and 2010. No information was provided on the use of air conditioning in the buildings. The overheating analysis was performed using ASHRAE Standard 55 adaptive thermal comfort [16]. The study found that the London dwellings faced a significant risk of overheating under the current climate. The results indicated that 29% (37%) of the living rooms and 31% (49%) of the bedrooms exceeded the overheating criteria of the ASHRAE standard during the summer of 2009 (2010). Baborska et al. [17] performed field measurements to evaluate the overheating risk in 18 retrofitted flats in one tower block in northern England that were known to have a summer overheating issue. It was discovered that during the hottest summer months in 2013 and 2014, the average indoor temperature was 8°C higher than the outdoor temperature. The measurements showed that overheating occurred in 44% of bedrooms and 28% of living rooms according to the CIBSE's fixed temperature overheating criterion [14]. Similar results were obtained using the overheating criterion based on the adaptive thermal comfort given in the BS/EN 15251 standard. Touchie et al. [18] conducted field measurements in seven mid and high-rise retrofitted multi-unit residential buildings located in Toronto, Canada. The outdoor and indoor conditions were collected over three years: prior (2015), during (2016), and after retrofit (2017). The results showed that indoor temperatures increased with the outdoor conditions in un-conditioned suites not facing north. In those suites, the indoor temperatures were higher than the adaptive thermal comfort and outdoor temperatures during the heatwave periods. Furthermore, the occupants were exposed to medium to high levels of heat stress, particularly during nighttime.

A.1.2 Motivation and objectives

Based on the previous literature overview, common limitations and issues appear in relation to the overheating analysis in residential buildings:

- Most efforts to understand residential indoor conditions from real household measurements have historically relied on small samples (commonly 10s and rarely 100s), during relatively short periods of time and in a localized geographic area. Currently, there are no studies that address all three data collection limitations.
- Another issue is that most commonly used thermal comfort standards, which are used to quantify the severity and frequency of overheating, have not been derived from direct assessments of homes. The CIBSE overheating recommendation is based on BS EN152511 adaptive thermal comfort model [19] which has been derived from data mainly obtained from office buildings where people have less opportunity to adapt. Another widely used thermal comfort standards, AHSRAE-55 adaptive model of thermal comfort, is also derived from a global database of 21,000 measurements taken primarily in office buildings [16].

To address these issues, we use indoor temperature and humidity measurements from more than 3,000 connected thermostats, during the summers of 2016-2019 and from homes located in five major metropolitan areas in Canada namely Montreal, Ottawa, Greater Toronto Area, Edmonton and Calgary. Furthermore, the current study uses a robust procedure to evaluate the frequency and severity of indoor overheating based on the heat-related health outcomes for older people. In particular, we comparatively evaluate overheating occurrence in homes with and without a central AC unit. In Canada, cooling and ventilation during summer months are not mandated in building energy codes and regulations, and the use of air conditioning in residential buildings varies among provinces with an average of 58% [20]. However, the aforementioned factors i.e. climate change, an aging population and highly-insulated homes raise concerns about the health of occupants, especially vulnerable people. Furthermore, in areas where air-conditioning is already used, there is a concern about survivability and health during unpredictable circumstances where the mechanical cooling equipment or the electricity supply fails.

This section is structured to present the indoor overheating analysis methodology first. Subsequently, the connected thermostat data is described. The results and discussion section summarizes our findings about the frequency and severity of overheating events in the homes under study. Finally, limitations of the current work and future research needs are discussed.

A.2 Indoor overheating analysis methodology

This section presents the principles of the used overheating assessment methodology for vulnerable subjects**.** More information on the methodology may be found in [21].

A.2.1 Definition of overheating events

The current method for the assessment of overheating risk uses the heat stress index of transient standard effective temperature (t-SET) and severity and intensity indices of overheating events (SETH, IOH), and body water loss and core temperature criteria to limit the dehydration and maximum core temperature of building occupants exposed to overheating events. In residential buildings, where an occupant can occupy different spaces during the daytime and the nighttime, the severity and intensity of an overheating event of duration N continuous days, is expressed as follows:

$$\mathrm{SETH} = \sum_{\mathrm{N}}(\mathrm{SETH_n} + \mathrm{SETH_d}) \qquad (1)$$

$$\mathrm{IOH} = \mathrm{SETH}/(\mathrm{N}*24) \qquad (2)$$

with:

$$\mathrm{SETH_d} = \sum_{\mathrm{daytime}}(\text{t-SET}_t - \text{t-SET}_{\mathrm{d}})^{+} \cdot \Delta\mathrm{t} \qquad (3)$$

$$\mathrm{SETH_n} = \sum_{\mathrm{nighttime}}(\text{t-SET}_t - \text{t-SET}_{\mathrm{n}})^{+} \cdot \Delta\mathrm{t} \qquad (4)$$

where:

SETH: severity of an overheating event (°C·h)

IOH: intensity of an overheating event (°C)N: duration of overheating event (days)t-$\mathrm{SET_t}$: hourly value of t-SET of the space being occupied at hour (t) during day or night time exposure (°C)

t-SETn : threshold value of t-SET for a sleeping occupant during nighttime exposure (°C)

t-SETd : threshold value of t-SET for an active (wakeful) occupant during daytime exposure (°C)

$\mathrm{SETH_d}$: magnitude of a heat event occurring over daytime exposure (°C·h)

$\mathrm{SETH_n}$: magnitude of a heat event occurring over the preceding nighttime exposure (°C·h)

Δt: calculation time step (h)

The symbol ($^+$) in equations (2) and (3) indicates that only positive values are considered.

Based on equation (1), a daily overheating event is counted only if its daytime magnitude ($SETH_d$) exceeds an arbitrarily minimum value of 4°C·h ($SETH_{d,min}$=4°C·h). The preceding nighttime magnitude ($SETH_n$) is counted in equation (1) only if the daytime magnitude ($SETH_d$) exceeds $SETH_{d,min}$. Therefore, sleep disturbance during the preceding nighttime only affects active occupants under daytime heat stress. Multiple overheating events are separated by a recovery period of at least one day with $SETH_d \leq SETH_{d,min}$.

The nighttime exposure is assumed between 10:00 pm to 7:00 am local time for all homes. Note that, in the context of overheating, a complete day of exposure starts from the sleep start hour (10 pm) of the previous night and ends at the sleep start hour of the current day. Occupants are assumed to dwell in the living room during the daytime and in the bedroom during the nighttime.

A.2.2 selection of threshold values of t-SET

The daytime and nighttime threshold values of t-SET ($t\text{-}SET_d$ and $t\text{-}SET_n$) for vulnerable occupants are chosen to ensure their health is not affected by long exposure to heat events. Table 1 lists the suggested thresholds values of t-SET for healthy older people in the residential buildings. The supporting literature of each value may be found in [21].

Table 1. Suggested threshold values of t-SETd and t-SETn for vulnerable occupants

$t\text{-}SET_d$ (°C) for free-running buildings		**$t\text{-}SET_d$ (°C) for air-conditioned buildings**		**$t\text{-}SET_n$ (°C)**
Acclimatized healthy older occupants (June-September) free to adapt to heat	Un- acclimatized healthy older occupants (May) free to adapt to heat	Acclimatized occupants with limited adaptation to heat	Un-acclimatized occupants with limited adaptation to heat	All occupants

27.7	26.5	26.7	25.5	30

A.2.3 Body water loss estimation

The selection of overheating criteria is important to ensure the health of building occupants is protected during extreme heatwaves. Dehydration and heat stroke (increasing body core temperature above 40°C) were the leading causes of mortality during heatwaves [22] and the current method is based on linking the severity and intensity indices of indoor overheating events to heat-related body dehydration and maximum core temperature during the heat exposure period. The physiological model for older people of Itani et al. [23], [24] is used to estimate the elderly body dehydration and core temperature under indoor conditions. For each building, the indoor conditions are used to calculate t-SET value and determine the overheating events through equations (1)-(4). For each overheating event, body water loss and maximum core temperature are calculated.

A.3 Building Data collection and analysis

A.3.1 Data description

The data used for this analysis were obtained from the 'Donate Your Data' (DYD) program provided by the connected thermostat manufacturer, ecobee Inc. The data consists of five-minute interval data that is measured by the thermostat (temperature, humidity, and motion) and any connected remote sensors (temperature and motion).

The original DYD dataset contains 104,690 thermostats from 49 countries. In this study, we use the data from Canadian houses. Therefore, the dataset, in the context of this paper, refers to the thermostats that are located in Canada. The data from five metropolitan areas i.e. Calgary (CLG), Edmonton (EDM), Greater Toronto Area (GTA), Montreal (MTL), and Ottawa (OTT) is used. The names of cities and towns in each area are listed in Appendix A.

In each metropolitan area, we selected thermostats that had been online for at least one month during the summer years of 2016-2019. For each year, the summer season was defined as the first day of May until the end day of September, except for 2019 where data was available till the end of August. For each thermostat, the runtime of the heating equipment

controlled by the thermostat was checked and the days that had heat runtime greater than zero were flagged as unusable for the overheating analysis. Furthermore, the overheating analysis framework uses hourly data, therefore, the thermostat interval data, which is five minutes, was processed into one-hour average intervals. The DYD dataset also provides information on the number of installed cooling stages which we used to divide the selected homes into two categories of dwellings; those with an installed central Air-conditioning (with AC), and without central air-conditioning (without AC). Table 2 shows the number of selected homes in each metropolitan area.

Table 2.Number of selected homes in each area

Area	Number of houses without AC	Number of houses with AC
CLG	274	284
EDM	108	254
MTL	35	147
OTT	16	635
GTA	63	1247
Total	496	2567

A.3.2 Daytime and nighttime indoor conditions

The smart thermostat connects to the home's HVAC equipment through the same wires as a traditional thermostat. Thermostat placement is based on where these wires are installed and is unique to each home. However, the remote sensors are wireless and battery operated and can be placed anywhere within 13.7m of the thermostat. The manufacturer recommends remote sensors to be kept away from direct sunlight and heating and cooling sources and placed in important rooms to improve comfort [25]. Based on these observations and in order to have a homogenous methodology for all homes, the thermostat temperature readings are assigned to daytime exposure (daytime t-SET_t in equation (3)) and, in houses with remote sensors, the average of all remote sensor temperature readings are used as the nighttime

temperature (nighttime t-SET_t in equation (4)). In houses with no remote sensors, the thermostat readings are used for both daytime and nighttime analysis.

DYD metadata contains the user-provided number of sensors for each house. However, the homeowners can add remote sensors or stop using them at any time. Therefore, for each of the selected buildings, the sensor readings were scanned and the real number of sensors was determined. Figure 1 shows the distribution of the number of sensors for the selected buildings with and without AC units.

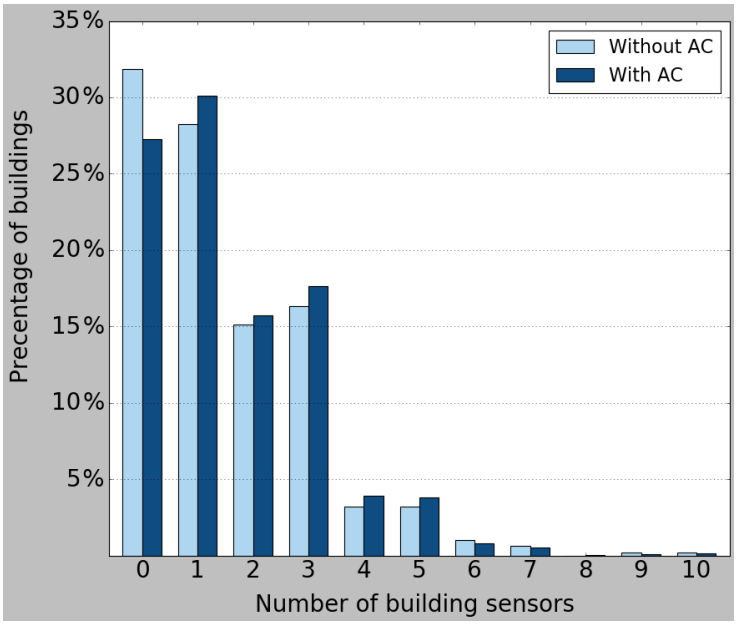


Figure 1. Distribution of number remote sensors for the selected homes

Figure 2 shows an example of the temperature readings from a randomly selected home with three remote sensors during the summer of 2019. There is evidence of two main types of missing data observed in the dataset. The first type can happen when power outages occur or if the thermostat loses connection with the user's wireless internet network (e.g. May 20-22). The second type of missing data, which is more common, happens due to the radio frequency connection lost between the remote sensor(s) and the thermostat (e.g. June 8 to 14). Furthermore, homeowners may choose to use or to stop using sensors at any time (e.g. sensor 2 has been used from June 25 to August 17). The following algorithm was used to handle various missing data scenarios:

- If the temperature readings (from either thermostat or remote sensors) are missing for

less than 6 consecutive hours, a linear interpolation was used to fill the missing values. Otherwise:

- If data is available from remote sensors, the average of readings from all available remote sensors is used as the nighttime temperature and the thermostat reading will be used for the daytime conditions.
- If data is not available from any of the sensors, the thermostat readings will be used for both daytime and nighttime analysis.
- If data is not available from either of thermostat and sensor(s) for more than 6 consecutive hours, that period of time will be flagged as unusable for the overheating analysis.

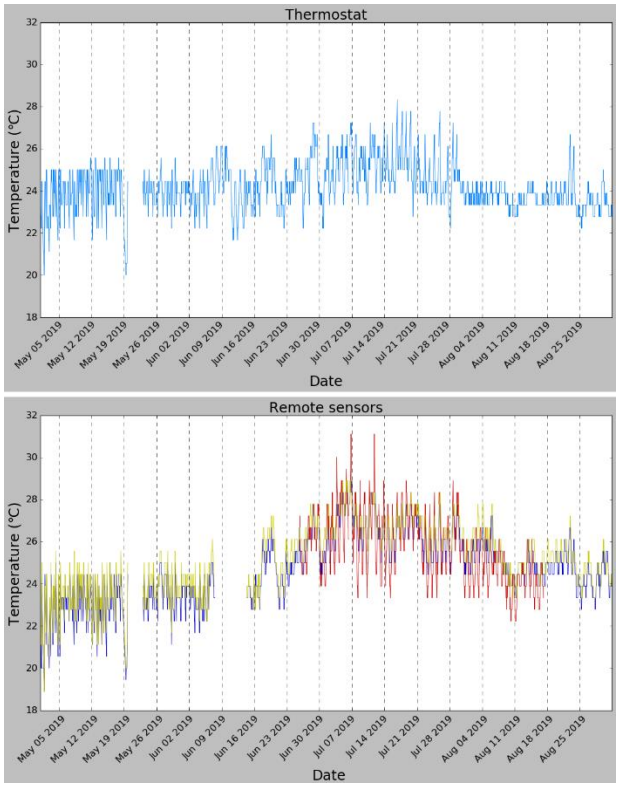


Figure 2. Sample of the hourly thermostat temperature data (top) and sensor temperatures (bottom) for a single thermostat

A.3.3 Building characteristics

Figure 3 compares the distribution of the user-provided construction age (as 2020) between homes with and without AC units. This figure shows that the construction age distribution broadly matches between the two home categories. The building type distribution of houses with and without AC units are compared in Figure 4. Note that in this figure 'other' refers to missing data and typographical errors. Alike Figure 3, the dwelling type distribution is similar between houses with and without AC units.

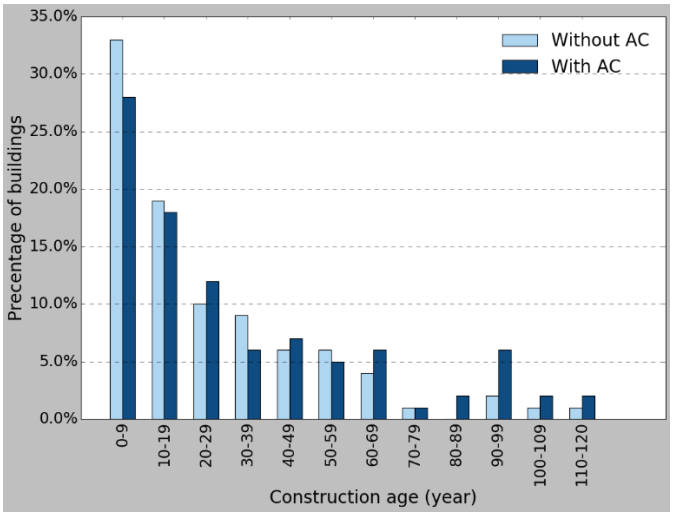


Figure 3. Distribution of buildings construction age

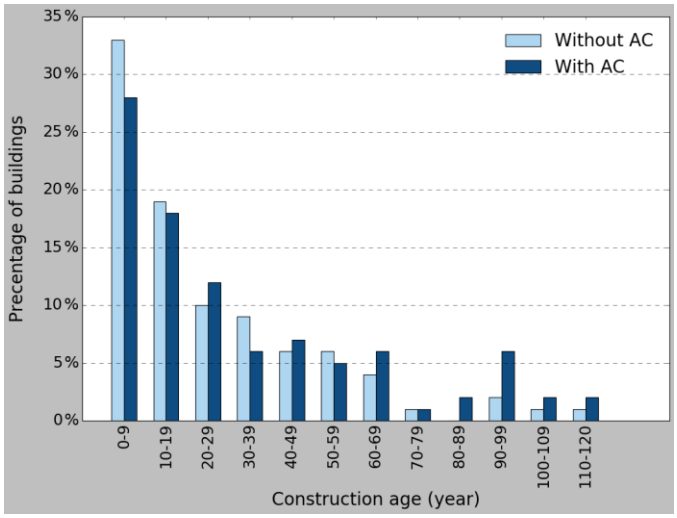
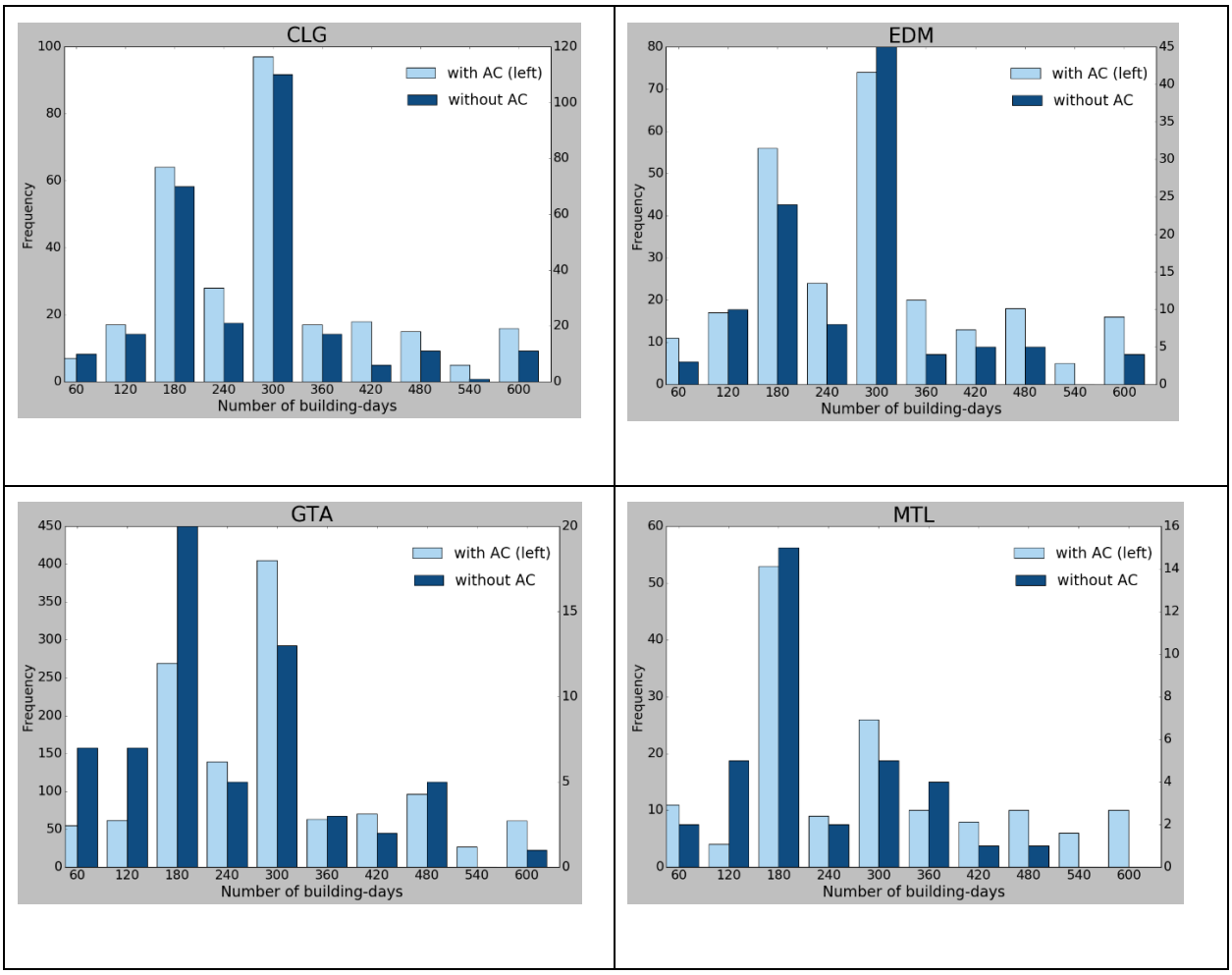


Figure 4. Distribution of building type

A.3.4 Data availability

For each building, the DYD dataset contains data beginning from the first connection (or from January 2015 if actual online status occurred before that date) and ending at the release of data on August 30, 2019. However, in some cases, the data collection has stopped shortly after the first connection. Therefore, the number of building-days for buildings with and without AC was defined. Multiple overheating events are, by definition, separated by a recovery period of at least one day. Therefore, we cannot have multiple overheating events per day. Moreover, based on equations (1-4) a single hourly data point is adequate to determine overheating during a day. Therefore, for each building, the number of days is defined as the number of days during which we have at least one hourly indoor temperature and humidity reading from the thermostat. Figure 5 shows the building-days distribution for houses with and without AC units. The total number of building-days in each area is presented in Table 3. The effects of data availability on the overheating analysis are discussed in the next section.



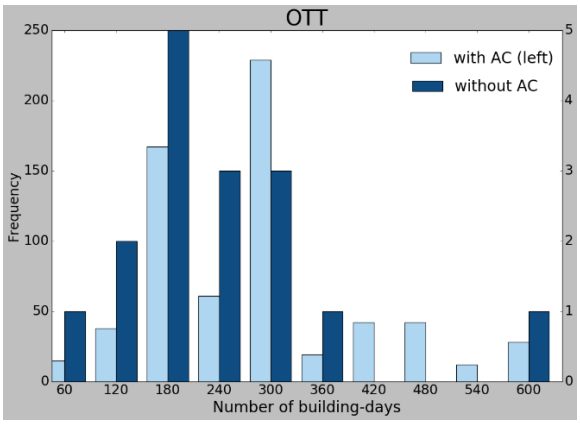


Figure 5. Distribution of building-days for each area

Table 3. Total number of building-days in different areas

Area	Number of building-days	
	With AC	Without AC
CLG	74,167	65,296
EDM	65,532	26,095
GTA	321,408	12,725
MTL	36,251	6,375
OTT	163,948	3,205
Total	661,306	113,696

A.4 Results and discussion

A.4.1 Indoor temperature variation

For each selected thermostat, the maximum daily indoor temperature and the prevailing outdoor temperature were calculated. To determine the prevailing outdoor temperature, the previous seven days' outdoor temperature was averaged with equal weighting [16]. Figure 6 shows the heat maps of maximum daily indoor temperature versus prevailing outdoor temperature for houses with and without AC. The 80% and 90% comfort acceptability limits

from the ASHRAE-55 adaptive comfort model are shown for houses without AC. Generally, the observed maximum indoor temperatures lie between the 90% limit at prevailing temperatures higher than 10^{o}C. However, about 16.5% of maximum daily indoor temperature data exceed the 90% acceptability upper bound threshold in homes without units. Similarly, ASHRAE-55 comfort temperature limits for the human occupancy model used in this study (air velocity of 0.1 m/s, metabolic rate of 1 met, clothing level of 0.5 clo and relative humidity of 50%) are shown for houses with AC units. Note that in houses with mechanical ventilation, the comfort temperatures limits are independent of the outdoor temperature. The results indicate that around 72.5% of maximum temperature readings are smaller than the lower limit while 17.3% of data lie between the limits.

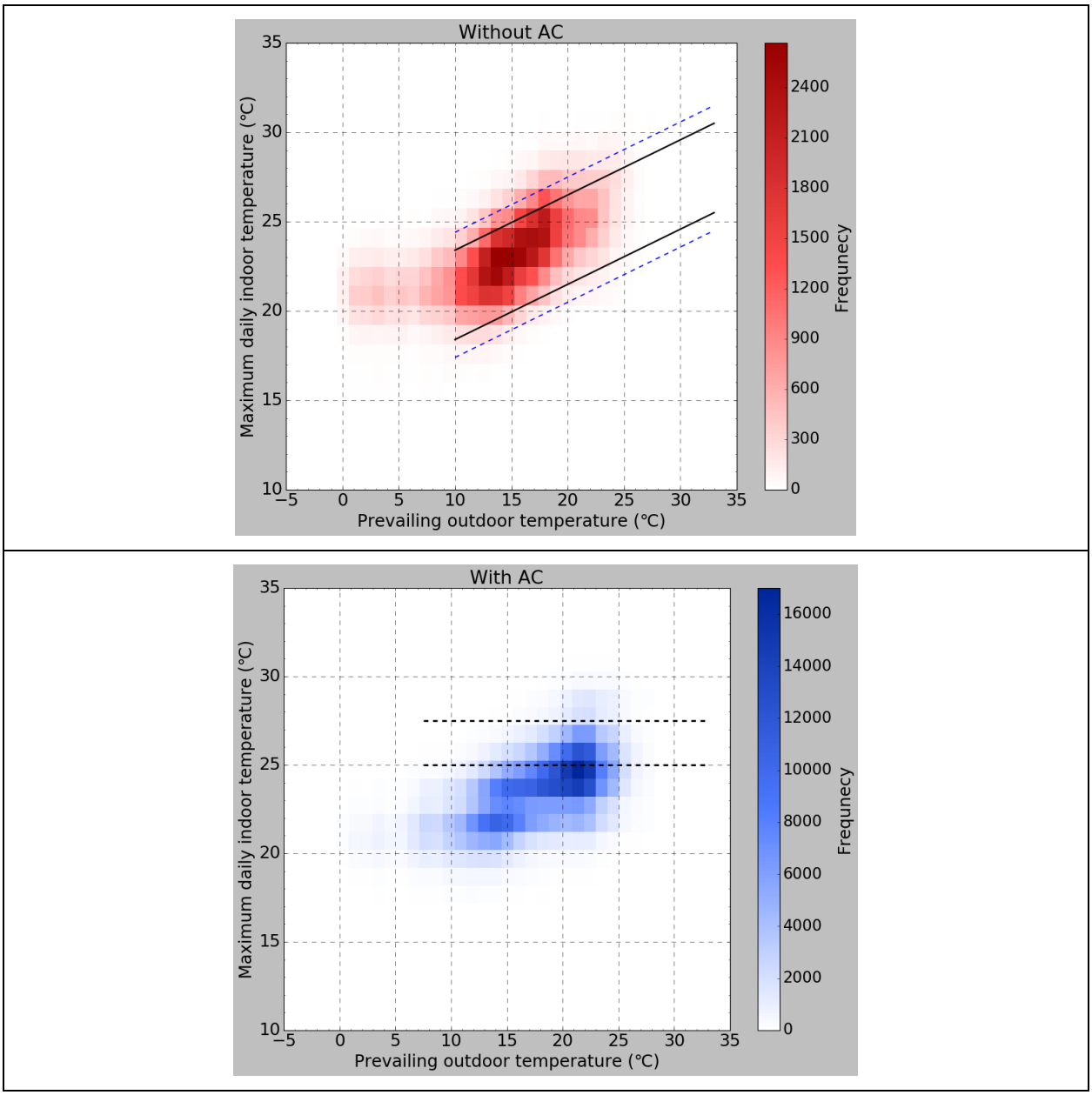


Figure 6. Maximum daily indoor control temperatures based on the 7-day prevailing outdoor temperature for homes without AC (top) and with AC (bottom) in all metropolitan areas. The 90% and 80% comfort acceptability limits from ASHRAE-55 adaptive comfort model are shown for comparison

As can be seen in figure 6, the most observed prevailing temperature in the without AC homes is about 15°C while this value is considerably higher for houses without AC (21°C). This is mainly due to the non-uniform geographical distribution of houses in this work. As presented in Table 1, 77% of homes without AC are located in the province of Alberta (CLG and EDM areas) which has a milder temperature during the cooling season (Figure 7). Furthermore,

Table 4 shows that in all provinces a considerably higher percentage of houses in the DYD dataset use AC units compared to the 2016 Statistics Canada data on the air conditioning use [26]. This is probably because homeowners who purchase a smart thermostat are more concerned about the thermal comfort and energy efficiency of their homes and are more likely to invest in a central AC unit. This is an inherent limitation of the current study and may result in a conservative estimation of overheating in warmer areas such as GTA.

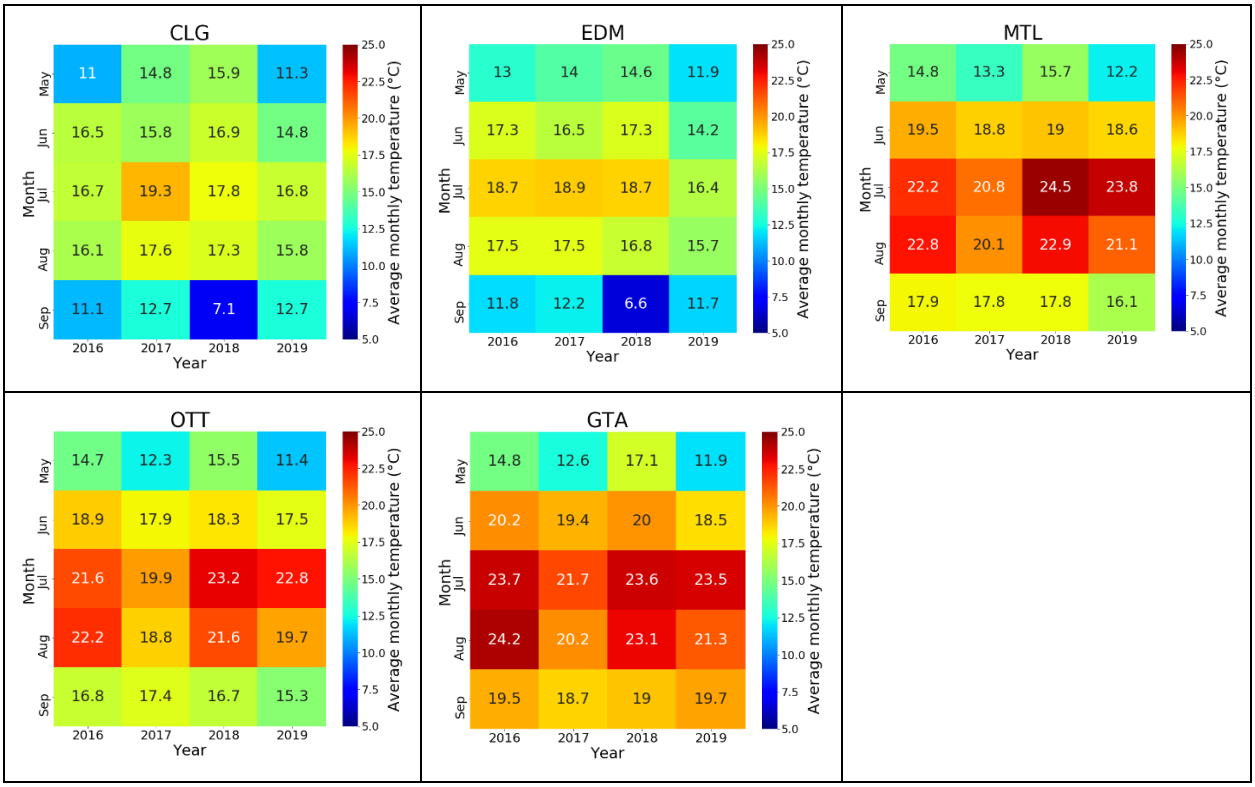


Figure 7. Outdoor monthly average temperature in the selected areas

Table 4. Air conditioning use in selected Canadian provinces

Province	Percentage of homes without central AC from DYD dataset	Percentage of homes without central AC in DYD dataset
Alberta (CLG and EDM areas)	78%	40%
Ontario (GTA and OTT areas)	37%	3.4%
Quebec (MTL area)	80%	14%

A.4.2 overheating analysis

For each home, the hourly indoor and outdoor temperature and relative humidity values were used to calculate t-SET values to be used in equations (1) to (4). For calculating the t-SET values the occupants are assumed to be in a sedentary seated position (metabolic rate of 1 met for the young and 0.79 met for older people) and wearing typical summer clothing (0.5 clo) during day time exposure. During nighttime exposure, the reference occupant is assumed in a sleeping mode (metabolic rate 0.7 met for the young and 0.55 met for older people), wearing a short sleeved top and bottom pajamas with a total clothing insulation of 1.38 clo [27], [28]. Furthermore, the mean radiant temperature is assumed to be equal to the indoor air temperature and air velocity is fixed at 0.1 m/s for all homes.

The total number of overheating events for each area was calculated based on the described overheating criteria and the frequency of overheating occurrence was found as follows:

$$f_{OHE} = \frac{\text{total number of overheating events}}{(\text{number of buildingdays}/100)} \qquad (5)$$

Note that some of the overheating events could be a result of homeowners who use the building as a secondary property or who take extended vacations. In these cases, the building is unoccupied and measures such as opening the windows or using AC are not used to lower the indoor temperature. Therefore, for each overheating event the hourly motion detection data (from the thermostat and, if available, remote sensors) during the overheating event and an arbitrary window of 9 hours after the end of the event was checked. If no activity was detected during that period of time the overheating event was neglected. The 9-hour window was used to lower the false negatives caused by occupants' inactivity (e.g. during sleep hours).

Table 5 presents the frequency of overheating occurrence in each area for houses with and without AC units. The homes in the CLG and EDM areas are less exposed to the risk of overheating. This is mainly due to the milder summers in these areas (the monthly average temperature in various areas is compared in Figure 7). Moreover, for each area, the use of AC unit considerably reduced the risk of overheating in the homes. This is particularly obvious in the warmer areas of MTL, OTT, and GTA.

Table 5. Frequency of overheating in homes with and without AC units

Area	OHE in buildings without AC	OHE in buildings with AC
CLG	3.7%	2%
EDM	1.1%	0.1%
MTL	14.1%	6.1%
OTT	21.8%	4.2%
GTA	15.7%	5.8%

Figure 8 shows the variation of indoor temperature and t-SET in a house with the largest observed overheating severity (SETH = 335^{o}C*h). This building is located in OTT area and is not equipped with a central air conditioner. During the period of time shown in this figure, the home undergoes two overheating events. In both events, the outdoor temperature remains higher than 18^{o}C which results in indoor temperatures higher than 26^{o}C throughout the overheating event. During the first event, the average daytime and nighttime indoor temperatures are 30.3^{o}C and 29.4^{o}C, respectively while the indoor temperature reaches 10.0°C higher than the outdoor temperature. The second overheating event shows a similar trend, however, the outdoor temperature falls after 72 hours and therefore, the overheating does not persist which results in a smaller SETH value.

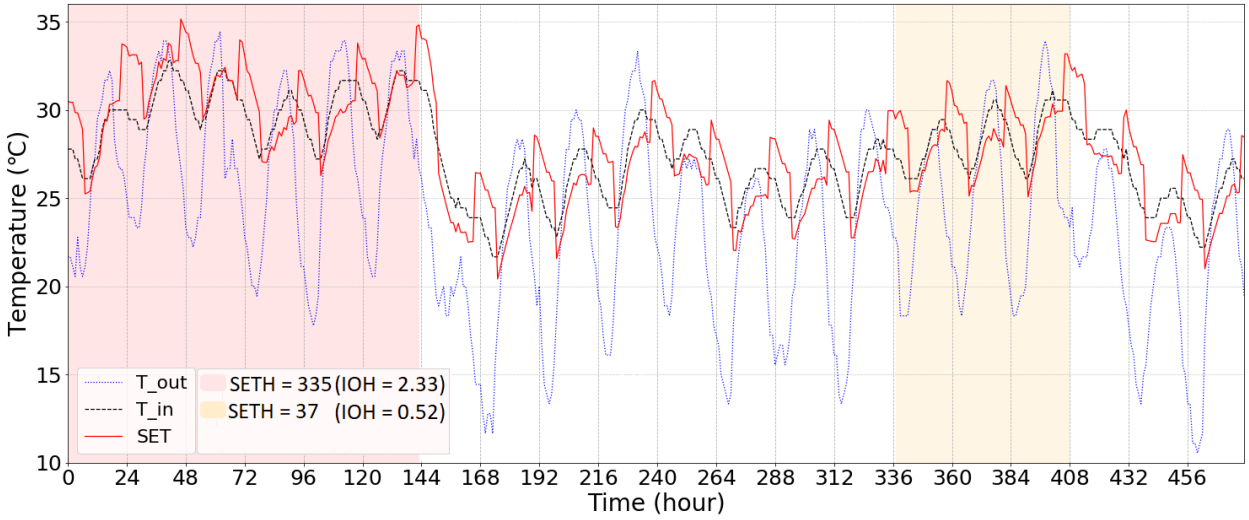


Figure 8. Variation of indoor temperature, outdoor temperature and SET during two observed heat events (2018)

Figure 9 compares the frequency of homes with and without AC units that are exposed to at least one overheating event, which led to a body water loss higher than 10% of body weight (houses with multiple overheating events were counted as one). Note that based on the body water loss limiting criterion as previously specified, 10% is the maximum allowable body water loss for the healthy older occupants assuming occupants replace 80% of water loss by sweating and respiration during heat events. The frequencies in Figure 9 are calculated similar to Equation (5):

$$f_{WH} = \frac{\text{number of homes with a body water loss higher than } 10\%}{\text{number of buildingdays}/100} \qquad (6)$$

As shown in Figure 9, the use of air conditioning considerably reduces the frequency of indoor overheating events that are associated with health risk to elderly occupants in all areas. However, this is especially the case for overheating events in the warmer areas of GTA, MTL, and OTT.

Note that the overheating analysis methodology proposed by Laouadi et al. [21] also uses core body temperature as a criterion for determining indoor overheating events. Based on this method, the maximum allowable core body temperature for older people corresponds to an overheating intensity (IOH) of 3.8°C. However, the overheating intensity did not reach the

threshold value in the houses under study, and therefore, the body water loss was considered as the only overheating creation.

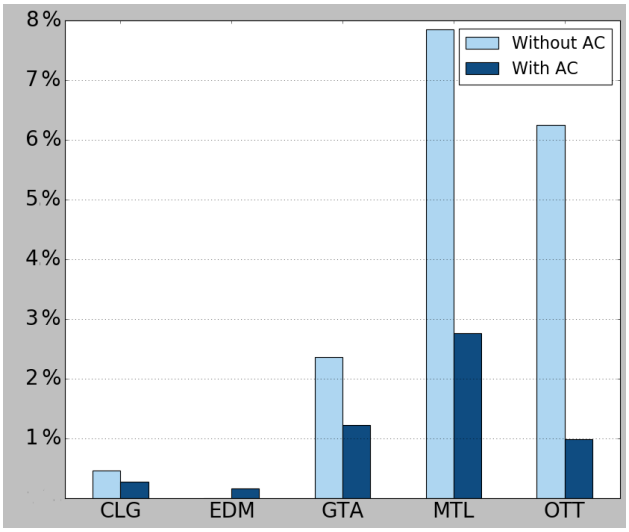


Figure 9. Percentage of houses with at least one overheating event that is associated with health risks to elderly occupants

A.4.3 Limitations of the current study

While this study has shown that connected thermostat data could be used to conduct large-scale studies on the health risks of indoor overheating to the vulnerable occupants, there are several limitations to this approach that need to be considered:

- The DYD dataset does not provide information on the use of stand-alone (portable or window) air conditioners. Therefore, some of the homes that are classified as without AC in this study, may use stand-alone air conditioning. This may result in a conservative estimation of the overheating risks in houses that do not use any type of AC unit.
- As shown in Figure 4, less than 5% of the total homes in this study are apartments However, around 28% of total Canadian private dwellings are Apartments [29]. One possible explanation for this discrepancy is that renters are less likely to purchase a smart thermostat for their residence and landlords/mass builders commonly use less expensive thermostats in this type of building. These factors, along with the potential compatibility issues between the connected thermostat and the building HVAC system,

pose a barrier to the market penetration of connected thermostats in apartments. Therefore, the results of this study are generally applicable to other types of home shown in Figure 4.

A.5 Conclusions and future work

This study shows how the time series data collected from connected thermostats, could be used as an effective approach for investigating the health risks of indoor overheating in various types of dwellings. The data source is, in particular, leveraged to solve the common issues, i.e. relatively short study length and small sample size, which have limited similar studies in the past. The hourly indoor temperature and humidity data from more than 3,000 homes during four summer years were used to assess the health outcomes of overheating on the older occupants. Two categories of home, i.e. with and without central air conditioner were considered to investigate the effects of AC usage on the frequency of overhearing events in residential buildings. It was shown that both home categories share similar building age and type distribution. The results show that both home categories undergo overheating, however, the frequency of severe overheating events, which may lead to serious health issues for seniors, can be significantly reduced by the use of central AC units. This could be of interest to the policymakers as the cooling and ventilation in summer are not mandated in Canadian building codes.

One main limitation of the current work is the uneven distribution of connected thermostats in the dataset which results in lower data availability from houses that are located in the warmer areas and are more exposed to extreme heat events. However, the growing market of household internet of things devices promises new streams of data that can be used to examine the residential buildings' indoor conditions.

One future direction of this work will seek to investigate the correlation between the frequency and severity of overheating and the basic dwelling characteristics such as the age of construction and building type. Furthermore, the dataset can be used to establish a relationship between indoor and outdoor conditions. The results of such a study can be utilized to assess the impacts of future climate change on the indoor overheating in Canadian homes.

SECTION B. A DATA-DRIVEN INDOOR OVERHEATING WARNING FRAMEWORK FOR RESIDENTIAL BUILDINGS

B.1 Introduction

Excess indoor heat and prolonged periods of hot weather affect the health and wellbeing of occupants and increase the risk of mortality and morbidity. It is shown that the elderly are less sensitive to thirst and heat, and therefore are at a higher risk of dehydration during long exposures to high indoor temperatures [1]. In addition, temperatures and heatwave frequency are predicted to increase due to climate change [30]. These factors have contributed to a growing concern about an increased occurrence of overheating during the warm summer weather in Canadian homes.

Based on the previous literature overview, most efforts to understand residential indoor conditions from real household measurements have historically relied on small samples, during relatively short periods of time and in a localized geographic area due to practical constraints. The growth in the use of connected thermostats has allowed for such studies to be conducted without the limitations of traditional data collection methods. In this work, the indoor temperature and humidity measurements from more than 180 connected thermostats, during the summer of 2019 and from homes located in Ontario, Canada are used to investigate the feasibility of developing a predictive warning framework.

The smart thermostat data is, in particular, was leveraged to develop an indoor overheating forecast framework tailored to the dwelling's thermal conditions and occupant behaviour. Two of the common options to mitigate high indoor temperatures are using air conditioning and opening windows. However, in Canada, cooling and ventilation of residential buildings are not mandated in building energy codes, and regulations regarding the use of air conditioning in residential buildings varies among provinces with an average of 58% [20] of homes adopting ACs across Canada. Opening the windows is not always feasible due to problems associated with pollution, noise and crime [33]. Moreover, when elevated outdoor temperatures exceed the indoor overheating limits, the use of natural ventilation as a heat purging strategy becomes ineffective or even counter-productive. Regardless of the indoor overheating mitigation strategy, it has been shown that heat-related health risks can be decreased by using a predictive indoor overheating warning system [34], [35].

B.2 Data description

The original 2019 DYD dataset contains data from 104,690 thermostats from 49 countries. In this study, we use the data from houses located in Ontario, Canada. Therefore, the dataset, in the context of this section of the report, refers to the thermostats that are located in Ontario that do not use a central air condition (AC) unit. For each building, the DYD dataset contains indoor and outdoor temperature and relative humidity data from the first connection (or from January 2015 if actual online status occurred before that date) until the release of data on August 30, 2019. However, in some cases, the data collection has stopped shortly after the first connection. We selected thermostats that had been online for at least one month during the summer of 2019. For each thermostat, the runtime of the heating equipment controlled by the thermostat was checked and the days that had heat runtime greater than zero were flagged as unusable for the overheating analysis. The DYD dataset also provides information on the number of installed cooling stages which we used to select homes without central air-conditioning.

The smart thermostat connects to the home's HVAC equipment through the same wires as traditional thermostats. The thermostat placement is based on where these wires are installed and is unique to each home. However, as mentioned in section A, the remote sensors are wireless and battery operated and can be placed anywhere within 13.7m of the thermostat but the exact location of remote sensors is not known. Based on these observations and in order to have a homogenous methodology for all homes, only the thermostat temperature readings were used in this study. Furthermore, if the temperature readings were missing for less than 3 consecutive hours, a linear interpolation was used to fill the missing values. Otherwise, that period of time was flagged as unusable for the overheating analysis.

The geographical distribution of the selected houses is shown in figure 10. Most homes are located in the highly populated areas in southern Ontario with around 25% of houses in the Greater Toronto Area.

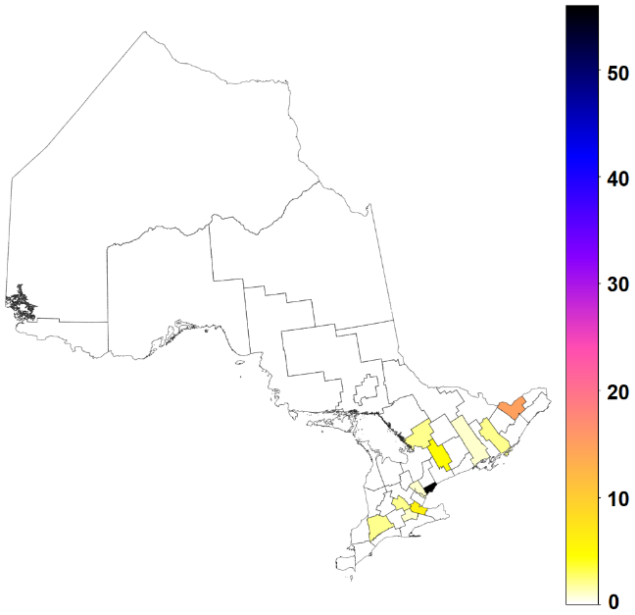


Figure 10. Selected houses distribution

B.3 Forecasting models

Prediction of indoor temperature has been attempted using white, grey, and black box models. White box models are purely theoretical, physics-based models that typically use empirical equations. A Gray box model combines a simple theoretical model structure with parameter estimation from observed data. Black box models do not assume any model formulations and knowledge of building properties and characteristics. The performance of these models for indoor temperature and relative humidity has been investigated in various studies [36]–[40]. White-box models provide detailed building energy performance simulations, however, their development is time-consuming and they require detailed building information which is not available in the current dataset. Some of the shortcomings of white-box models are addressed by gray and black box approaches. These are usually easier to calibrate and more scalable than white box models [41] which makes them favorable for large-scale studies. In particular, Yu et al. [42] used smart thermostat data to evaluate the accuracy of two black box models for indoor temperature prediction based on artificial neural network (ANN) and generalized regression neural network (GRNN) algorithms. They reported GRNN to be a better predictor with a mean square error of 0.26^{o}C. Arendt et al. [43] showed that black-box models were able to predict indoor temperature with high accuracy. Despite the advantages of black box models, they are inherently interpolation algorithms and are not suitable for making

predictions with unseen data. Gray box models structure is based on physical grounds, and therefore, these models typically require less experimental data than black-box models and are less sensitive to the data quality. Moreover, they should be more robust for extrapolation on operating conditions outside of those encountered in the training dataset [44]. Baasch et al. [45] compared the performance of three different grey box models to estimate the dwelling thermo-physical properties from smart thermostat data. They developed a simplified 1R1C gray box model based on the thermal energy balance. Although this model was easy to implement, it resulted in a large number of outliers. Hossain et al. [46], [47] developed various Bayesian neural network-based gray box thermal models by using smart thermostat data. They showed that a simple 1R1C model is not suitable for all houses and higher order models result in a smaller number of outliers.

B.4 Indoor temperature models

In this study, two gray box models were developed based on the hourly indoor and outdoor temperatures. The gray box models were implemented in a recursive forecasting strategy using sliding training and forecast time periods to provide the forecasted hourly indoor temperatures. Note that gray box models developed in this work are linear which, due to their lower computational cost, make them more suitable for large-scale studies. Furthermore, it has been shown that more complex non-linear prediction models, as well as the use of additional predictors does not necessarily increase the forecasting accuracy obtained by linear models [48].

The gray box models are based on an equivalent electrical circuit (RC model). In this approach, temperature difference is the driving potential, and the flow through the circuit is the heat flow [38], [49]. Various RC network models have been proposed in the literature. As shown in [45]–[47], the low-order 1R1C model usually results in a large number of outliers. High-order RC models are generally more accurate but they are more computationally expensive to use, which makes them less suitable for large-scale time-series analysis. Therefore, in this study, we developed a 2R2C and a 3R2C model to examine if these relatively simple models can offer a reasonable trade-off between performance and computational cost.

In this work, the solar and internal heat gains were neglected since this data is not collected by smart thermostats and is therefore not provided in the data set. The governing equations from

the thermal network models can be derived with, for example, Kirchhoff's node potential law. For the 2R2C model:

$$\frac{\partial T_{in}}{\partial t} = -\frac{1}{R_w C_i} T_{in} - \frac{1}{R_b C_b} T_b + \frac{1}{R_w C_i} T_{out} \quad (7)$$

$$\frac{\partial T_b}{\partial t} = -\frac{1}{R_b C_w} T_{in} \quad (8)$$

For 3R2C model:

$$\frac{\partial T_{in}}{\partial t} = -\left(\frac{1}{R_b C_b} + \frac{1}{R_g C_b}\right) T_{in} + \left(\frac{1}{R_b C_b}\right) T_w + \left(\frac{1}{R_g C_b}\right) T_{out} \quad (9)$$

$$\frac{\partial T_w}{\partial t} = \left(\frac{1}{R_b C_w}\right) T_{in} - \left(\frac{1}{R_w C_w}\right) T_w + \left(\frac{1}{R_w C_w}\right) T_{out} \quad (10)$$

In these equations, T_{in} and T_{out} represent indoor and outdoor temperatures, R_w is the thermal resistance of walls and C_i and C_b show internal and building envelop thermal capacitance, respectively. R_g represents the thermal resistance of parts of the building envelope that are not included in R_w, such as windows and the doors. C_w is the capacitance associated with walls. Rb, Tb, Cb, Tw?

How these model (C, R) parameters are calculated

B.5 Parameter estimation for gray box models

To estimate the model parameters, the differential equations (7)-(10) were converted into a linear state-space representation, as follow:

$$\frac{\partial x}{\partial t} = Ax + Bu \quad (11)$$

$$y = Cx + Du \quad (12)$$

In equations (11) and (12), the temperatures form the state vector (x), the indoor temperature is the output represented by y, and outside temperature forms the input vector denoted by u.

The state matrices are as follows:

For 2R2C model:

$$A_2 = \begin{bmatrix} -\frac{1}{R_w C_i} & \frac{1}{R_b C_b} \\ -\frac{1}{R_b C_w} & 0 \end{bmatrix}, B_2 = \begin{bmatrix} \frac{1}{R_w C_i} \\ 0 \end{bmatrix}, u = [T_{out}], x_2 = \begin{bmatrix} T_{in} \\ T_b \end{bmatrix}$$

For 3R2C model:

$$A_3 = \begin{bmatrix} -\left(\frac{1}{R_b C_b} + \frac{1}{R_g C_b}\right) & \frac{1}{R_b C_b} \\ \frac{1}{R_b C_w} & -\frac{1}{R_w C_w} \end{bmatrix}, B_2 = \begin{bmatrix} \frac{1}{R_w C_i} \\ 0 \end{bmatrix}, u = [T_{out}], x_2 = \begin{bmatrix} T_{in} \\ T_w \end{bmatrix}$$

$$C = [0 \quad 1], D = [0 \quad 0]$$

The response to the system described in (11)-(12) can be written as [50]:

$$x_{(t)} = (F.I - \Phi)^{-1}\left(q.\Gamma_2 + \Gamma_1 - \Gamma_2\right)u_{(t)} \quad (13)$$

where q denotes the forward shift operator:

$$\Phi = e^{A\delta}, \Gamma_1 = A^{-1}(\Phi - I)B, \Gamma_2 = A^{-1}\left(\frac{\Gamma_1}{\delta} - B\right)$$

By substituting equation (13) in (12):

$$y_{(t)} = C(q.I - \Phi)^{-1}\left(q.\Gamma_2 + \Gamma_1 - \Gamma_2\right)u_{(t)} \quad (14)$$

Assuming that inverse of $(q.I - \Phi)$ is non-singular:

$$y_{(t)} = \sum_{j=0}^{2}(S_j u_{t-j\delta}) - \sum_{j=1}^{2}\left(e_j y_{t-j\delta}\right) \quad (15)$$

where:

$$S_0 = C\Gamma_2$$

$$S_1 = C[R_0(\Gamma_1 - \Gamma_2) + R_1\Gamma_2] + e_1 D$$

$$S_2 = CR_1(\Gamma_1 - \Gamma_2) + e_2 D$$

where R matrices can be found based on Leverrier's algorithm as follow:

$$R_0 = I, \; R_i = \Phi R_{i-1} + e_i I \quad \text{where} \quad e_i = \frac{Trace(\Phi R_{i-1})}{n} \quad \text{for } i = 1, 2$$

Using this approach, as shown in equation (15), the indoor temperature ($y_{(t)}$) can be expressed as a linear function of indoor temperature at the previous time step and the outside temperature at the two previous time steps. The smart thermostat data, along with equation (15), was utilized to estimate the state matrices by using a non-linear least squares method.

B.6 Indoor relative humidity models

The recorded outdoor relative humidity (RH_{out}) was checked for all selected houses. In all cases, RH_{out} values reached 100% in the training dataset. Therefore, unlike indoor temperature, extrapolation is not required for the indoor relative humidity models (RH_{in}). It is

shown that RH_{in} can be estimated with a releasable accuracy by using linear models [51]. Hence, in this study, a multiple linear regression model was used to estimate hourly RH_{out} based on the outdoor temperature and relative humidity. In order to select the most appropriate model, various training windows between 1 to 5 hours was used as shown in equation (16):

$$RH_{in}{}^{t} = f\left(RH_{out}{}^{t-i\delta}, T_{out}{}^{t-i\delta}\right) \qquad 0 < i < 5 \qquad (16)$$

A model with the minimum mean absolute error was selected as the RH_{out} forecast model.

B.7 Results

In this study, the forecasting accuracy was evaluated using two scale-dependent error metrics: mean absolute percentage error (MAPE) and mean absolute error (MAE).

As mentioned in section B.4, a recursive forecasting strategy using sliding training and forecast windows was utilized to forecast the indoor temperature. Five training windows namely 48, 72, 96, 120 and 144 hours were used to train gray box models. The trained models were implemented in a recursive forecasting strategy to predict the hourly indoor temperature during the next 12, 24 and 48 hours.

Figures 11a and 11b show the MAE of indoor temperature predictions of 2R2C and 3R2C models, respectively. As can be seen, the 2R2C models perform noticeably better than the 3R2C model for all training windows. The 2R2C model is able to predict indoor temperature with a median MAE less than 0.2℃. However, both models result in a number of outliers.

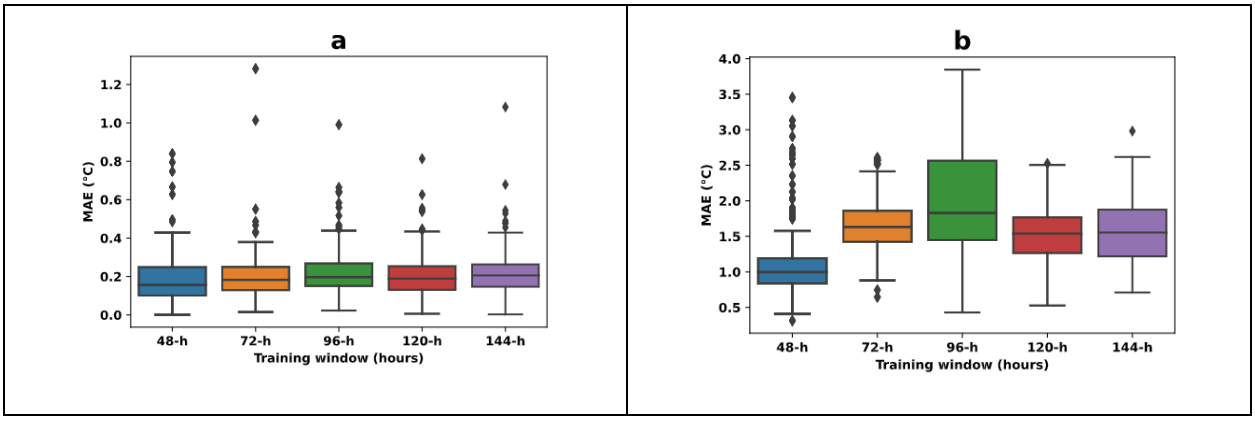
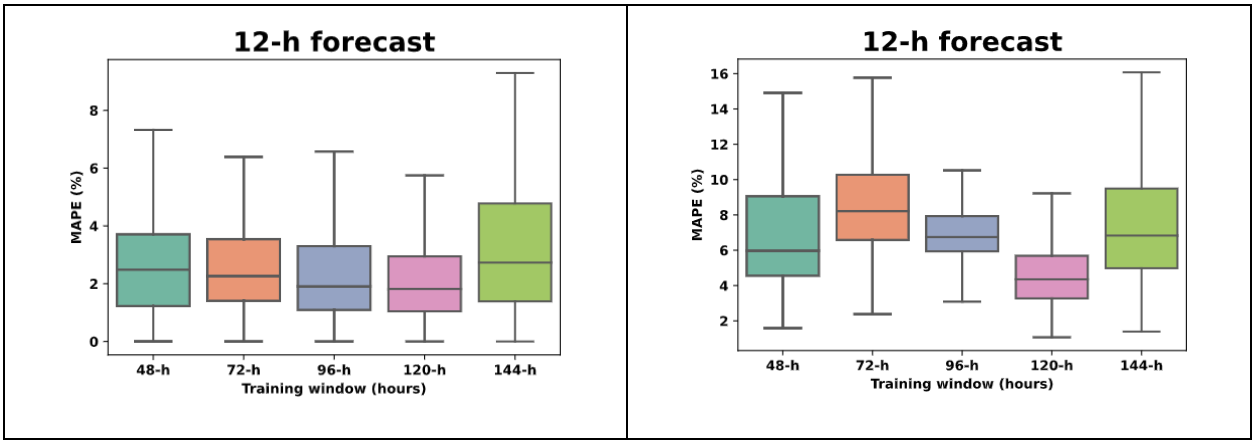


Figure 11. MAE distribution of a) 2R2C; and b) 3R2C grey box models for various training windows

Although figure 11 can provide useful information about the gray box models performance, the accuracy of the forecasting models can only be evaluated based on the recursive strategy. In a recursive strategy, the prediction from a one-step-ahead is used as an input for future prediction horizons. An important advantage of recursive approaches is that only one model is required, which considerably reduces the computational time, particularly in large-scale studies when a model needs to be trained for each house. Conversely, recursive forecasting typically produces biased predictions. This is because any error generated in the first step of prediction accumulates with each subsequent prediction, making accurate predictions at longer forecasting horizons more challenging [52].

The MAPE values of the models were used to evaluate the performance of forecast tools for the three forecast horizons, namely 12, 24 and 48 hours (figure 12). The results show that in all cases the 2R2C-based model outperforms the 3R2C-based model. Furthermore, it is evident that forecast accuracy decrease with the forecast horizon. As mentioned above, this is an inherent feature of recursive forecast strategies. In 2R2C-based models, the 96-h training horizon produces the lowest MAPE and therefore, this model was selected as the best performing model.



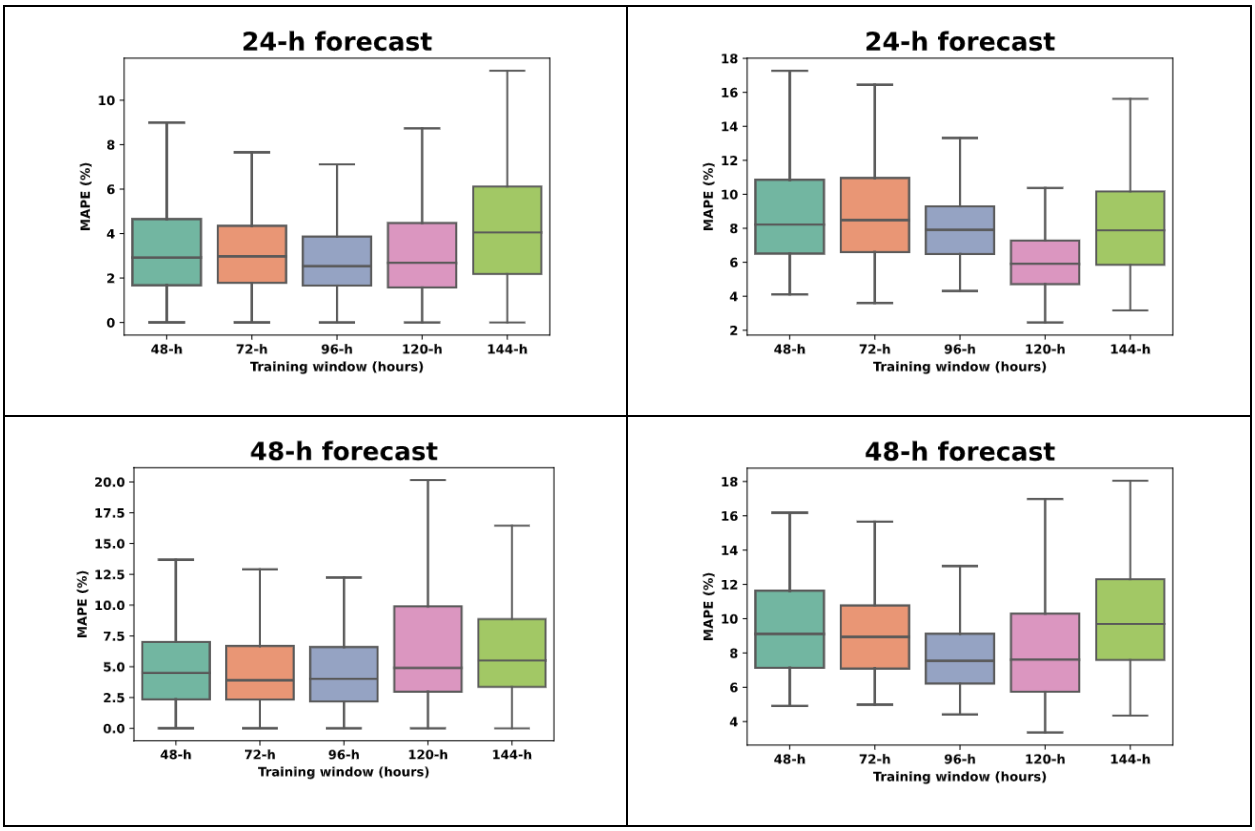


Figure 12. MAPE distribution of 2R2C (left) and 3R2C (right) based forecast models for various training windows

In this study, an MAPE of 5% is used as the threshold for forecast accuracy. Table 6 represents the percentage houses for which the selected forecast model (2R2C-based model with 96-h training window) can achieve an MAPE lower than 5% for various forecast horizons.

Table 6 shows that the selected model is able to forecast the 12-h ahead indoor temperatures with less than 5% error for the majority of houses. As the forecast horizon increases, the model's prediction ability decreases which can be mainly contributed to the error accumulation in the recursive forecast strategy. Other contributing factors are discussed in section B.8.

Table 6. Performance of 2R2C-based model with 96-h training window

Forecast horizon (hours)	Percentage of houses with MAPE lower than 5%
12	92

24	87
48	61

To provide a more detailed overview of the selected model performance, three houses were selected to represent the best, typical and the worst acceptable forecast accuracy based on the MAPE criterion discussed in the previous section. The characteristics of houses are presented in Table 7.

Table 7. Characteristics of the selected houses

Characteristics	House A (minimum MAPE or the best accuracy)	House B (median MAPE or the typical accuracy)	House C (threshold MAPE or the worst acceptable accuracy)
Location	Cambridge	Drayton	Toronto
Type	Detached	Detached	Condominium
Floor area (m^2)	150	200	150
Age (years)	90	70	30
MAPE (24-h horizon)	0.04%	2.53%	4.90%

The indoor and outdoor temperatures variation for all three houses is shown in figure 13. For House A, the forecasts are very accurate with the maximum error of 0.4°C, while the model tends to forecast a smooth temperature profile. For House B, the model overestimates the minimum indoor temperature by about 0.9°C, while at higher outdoor temperatures the model is relatively accurate and error does not exceed 0.3°C. For house C, the model struggles to adapt to the rapid outdoor temperature increase between 5am and 12am and underestimates the indoor temperature with a maximum error of 1.2°C. As the outdoor temperature becomes more

stable during the rest of the day, the model accuracy significantly increases (the mean absolute error between 13pm and 23:59pm is 0.3°C).

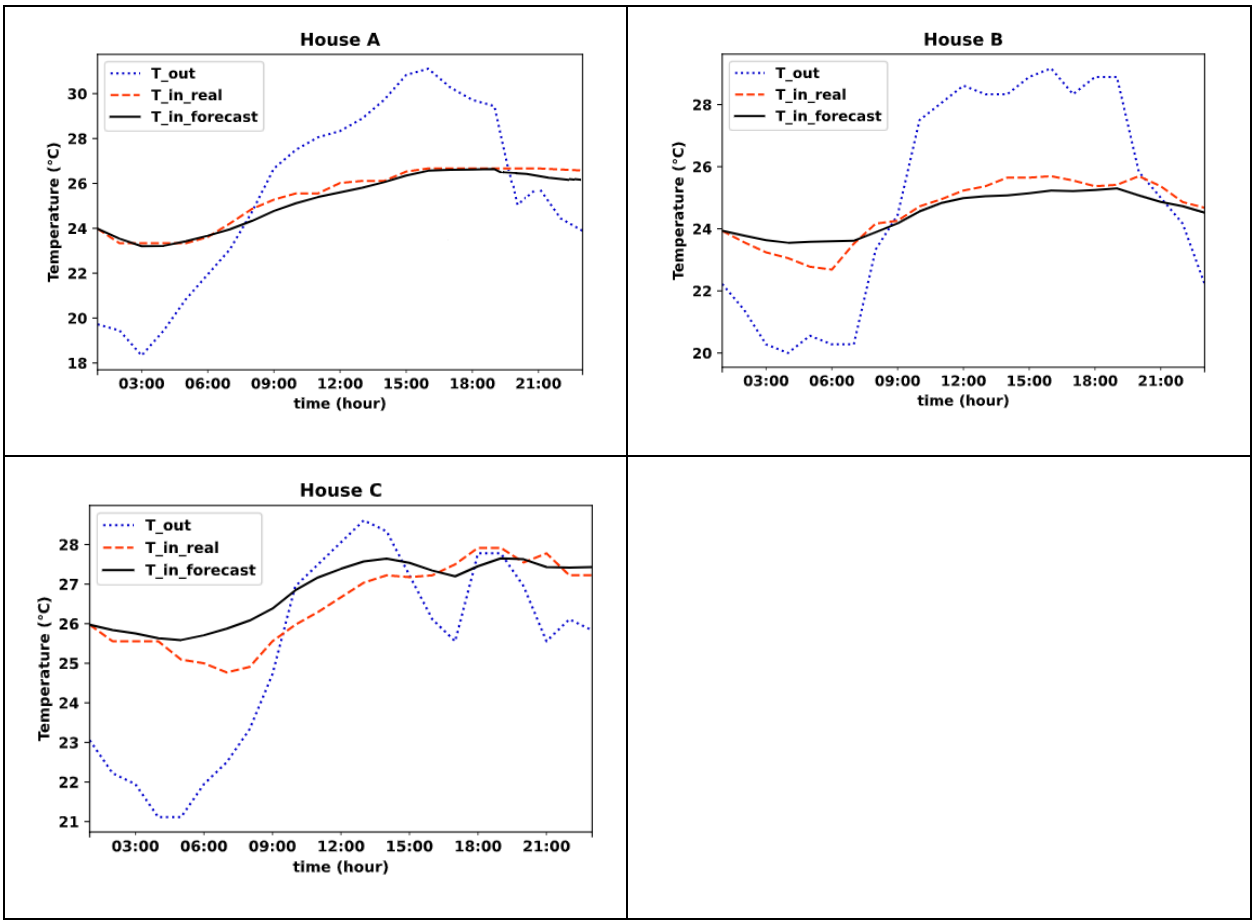


Figure 13.Indoor and outdoor hourly temperatures based on the selected model for various houses

The performance of the proposed relative humidity forecast models is shown in figure 14. This figure shows that a model with a four-hour training window can forecast indoor relative humidity with a median MAE of 4.1% in all houses under study. Therefore this model was selected to estimate the indoor relative humidity.

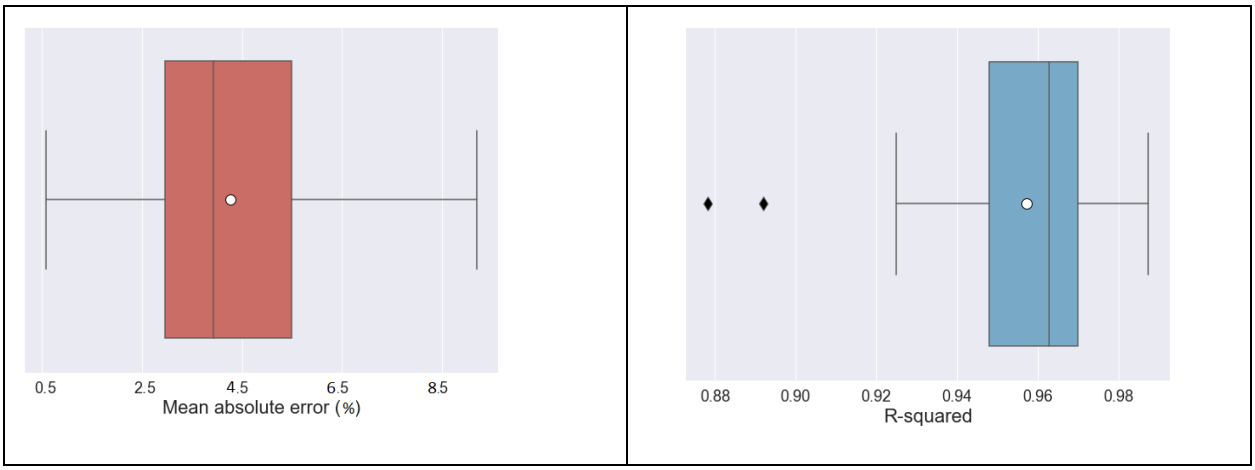


Figure 14. MAE and R-squared distribution of indoor relative humidity model

The variation of indoor relative humidity for the three houses described in the previous section is shown in figure 15. In general, the models tend to underestimate the indoor relative humidity while predicting a smoother profile compared to the actual values. Maximum absolute errors were 3.3%, 4.2% and 4.8% for houses A, B and C, respectively.

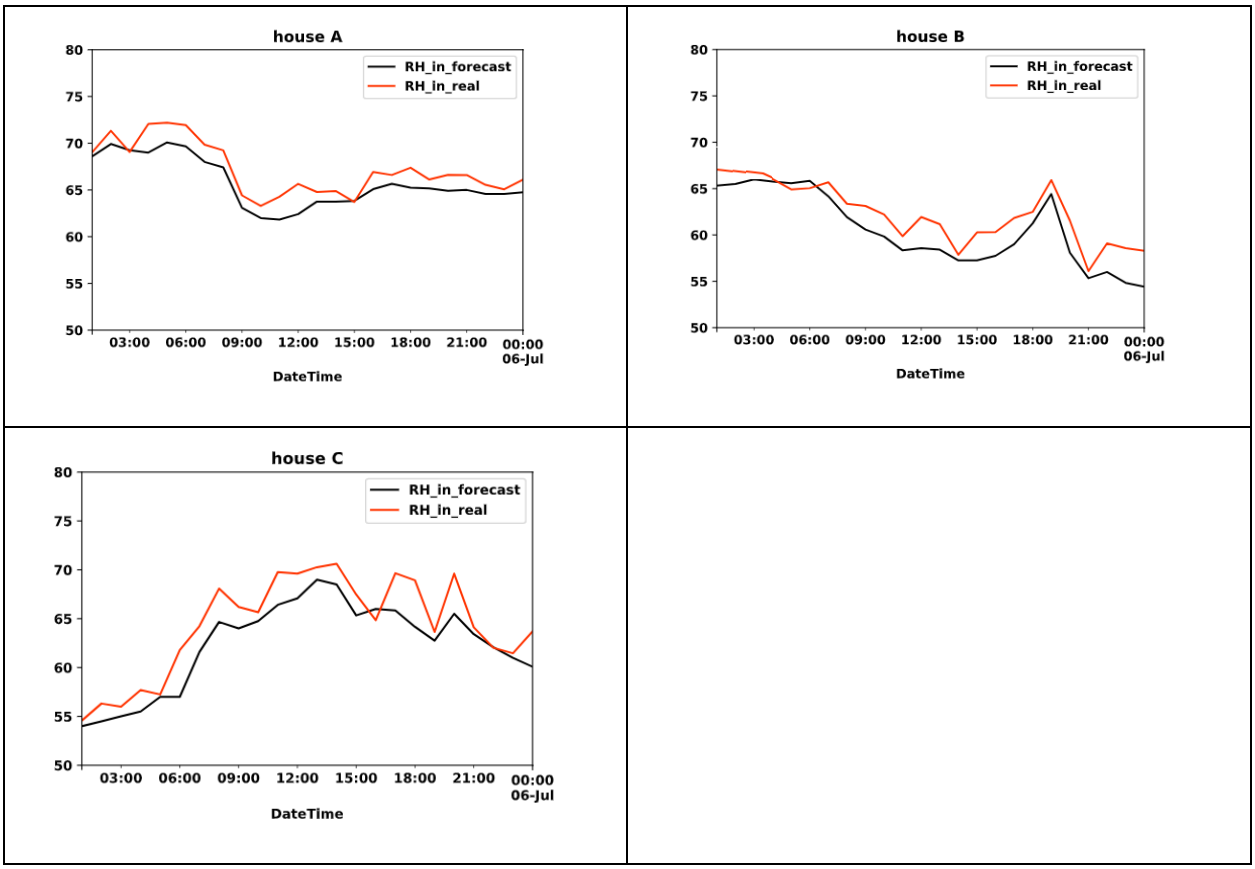


Figure 15. Variation of actual and forecasted hourly indoor relative humidity for various houses

B.8 Discussion

The aim of this work was to investigate the feasibility of using data collected from smart thermostats for developing indoor conditions forecast models. Such a model can provide an early warning of likely elevated temperatures and indoor overheating events. First, two gray box models were developed to capture the house-specific response to the outdoor conditions. Then, the effects of training window and forecast horizon on the models performance were studied to identify the best performing forecast strategy. The global strategy is an important feature of the forecast models since it removes the need for manual model identification for each house. However, it was shown that no model was able to forecast indoor conditions of all houses. For a 12-h forecast horizon, the selected model was able to forecast indoor temperature of 92% of houses with less than 5% MAPE, while that number dropped to 61% for a 48-h forecast horizon. Three main reasons may contribute to these observations. First, in this work, a recursive strategy was used and because of the error accumulation in each subsequent prediction, the model performance decreases at larger forecast horizons. Second, the DyD dataset offers no detailed information about the building characteristics (e.g. house orientation, shading, and thermostat location). Therefore the effects of solar radiation on indoor temperature cannot be modeled. Furthermore, the dataset only contains data about the central cooling system. Therefore, some houses may use window units that are not available in the dataset. Third, in residential buildings, the occupants' behaviour can have strong effects on the indoor conditions, especially during warmer days. Mitigating actions during a heatwave could include opening windows and even doors and closing the curtains during the day. There are no data on these actions and therefore, models cannot learn the effects of mitigation actions. In this study, the effects of mitigation actions were decreased by training the gray box models over relatively short periods of time, as well as, using a sliding approach.

However, the forecast models can be viewed as a forecast of what will happen if no one takes mitigation actions (beyond the established patterns of operation). From a health perspective, this can be useful information since it allows the occupants (or their caregivers) to proactively take action to lower the indoor temperature in order to maintain it within safe levels. Because at shorter forecasting horizons (12-h) the predictions are more accurate and consistent, a warning framework with a 12-h horizon could be set to alert the occupants or caregivers ahead of dangerous indoor conditions. This is especially beneficial for elderly homeowners and long-term care homes in which occupants are more vulnerable to indoor heat events.

B.9 Conclusions

Early prediction of impending high temperatures in buildings could play a vital role in reducing heat-related morbidity and mortality. In this work, the potential for leveraging data from smart thermostats to predict indoor conditions has been investigated using hourly data from 189 houses in Ontario during the summer of 2019. A combination of gray box and black box models along with a recursive multi-step-ahead approach with sliding training windows and forecast horizons was used to estimate indoor temperature and relative humidity. These provided predictions for forecasting horizons of 12, 24 and 72 hours. The accuracy of the predictions over the real measurements was evaluated using MAPE and MAE over different forecasting horizons.

Comparison between the various forecast horizons showed that the proposed approach can forecast 12-hour ahead indoor temperature and humidity in 92% of houses with less than 5% MAPE. Also, the mechanisms behind lower accuracy in longer forecast horizons were discussed.

Overall the early findings of this work suggest that highly detailed building information is not required to produce reasonable forecasts of indoor conditions in houses that do not use a central AC unit. This indicates the potential for using IoT devices and machine learning as part of an overheating detection and warning system in buildings. Although this work was based on data from residential buildings, the approaches developed can also be used for long-term care homes.

Future work will explore alternative forecast approaches such as ARX and ARMAX. Finally, testing of the forecasting models on simulated climate change data will be carried out in order to examine the reliability of forecasts in houses across Canada.

ACKNOWLEDGMENTS

This work was carried out by the Construction Research Centre of the National Research Council Canada with funding from Infrastructure Canada in support of the Pan Canadian Framework on Clean Growth and Climate Change. The authors were very thankful for their support. The authors would like to extend their gratitude to Saptak Dutta for reviewing this document and providing useful comments and suggestions.

REFERENCES

[1] K. J. Lomas and S. M. Porritt, “Overheating in buildings: lessons from research,” *Build. Res. Inf.*, vol. 45, no. 1–2, pp. 1–18, 2017, doi: 10.1080/09613218.2017.1256136.

[2] P. H. England, “Heatwave plan for England,” *Pharm. J.*, 2015, doi: 10.1211/pj.2015.20068624.

[3] R. Bustinza, G. Lebel, P. Gosselin, D. Bélanger, and F. Chebana, “Health impacts of the July 2010 heat wave in Québec, Canada,” *BMC Public Health*, vol. 13, no. 1, 2013, doi: 10.1186/1471-2458-13-56.

[4] A. Beizaee, K. J. Lomas, and S. K. Firth, “National survey of summertime temperatures and overheating risk in English homes,” *Build. Environ.*, vol. 65, pp. 1–17, 2013, doi: 10.1016/j.buildenv.2013.03.011.

[5] S. M. Tabatabaei Sameni, M. Gaterell, A. Montazami, and A. Ahmed, “Overheating investigation in UK social housing flats built to the Passivhaus standard,” *Build. Environ.*, vol. 92, pp. 222–235, 2015, doi: 10.1016/j.buildenv.2015.03.030.

[6] Government of Canada, “Action for Seniors report,” 2014. [Online]. Available: https://www.canada.ca/en/employment-social-development/programs/seniors-action-report.html.

[7] R. S. Kovats and S. Hajat, “Heat stress and public health: A critical review,” *Annu. Rev. Public Health*, vol. 29, pp. 41–55, 2008, doi: 10.1146/annurev.publhealth.29.020907.090843.

[8] A. Smargiassi, M. Fournier, C. Griot, Y. Baudouin, and T. Kosatsky, “Prediction of the indoor temperatures of an urban area with an in-time regression mapping approach,” *J. Expo. Sci. Environ. Epidemiol.*, vol. 18, no. 3, pp. 282–288, 2008, doi: 10.1038/sj.jes.7500588.

[9] M. Loughnan, M. Carroll, and N. J. Tapper, “The relationship between housing and heat wave resilience in older people,” *Int. J. Biometeorol.*, vol. 59, no. 9, pp. 1291–1298, 2015, doi: 10.1007/s00484-014-0939-9.

[10] J. L. White-Newsome *et al.*, “Climate change and health: Indoor heat exposure in vulnerable populations,” *Environ. Res.*, vol. 112, pp. 20–27, 2012, doi: 10.1016/j.envres.2011.10.008.

[11] A. Quinn *et al.*, “Predicting indoor heat exposure risk during extreme heat events,” *Sci. Total Environ.*, vol. 490, pp. 686–693, 2014, doi: 10.1016/j.scitotenv.2014.05.039.

[12] J. A. F. van Loenhout *et al.*, “The effect of high indoor temperatures on self-perceived health of elderly persons,” *Environ. Res.*, vol. 146, pp. 27–34, 2016, doi: 10.1016/j.envres.2015.12.012.

[13] M. Vellei *et al.*, “Overheating in vulnerable and non-vulnerable households,” *Build. Res. Inf.*, vol. 45, no. 1–2, pp. 102–118, 2017, doi: 10.1080/09613218.2016.1222190.

[14] CIBSE, “CIBSE Guide A: environmental design,” London, 2015.

[15] A. Pathan, A. Mavrogianni, A. Summerfield, T. Oreszczyn, and M. Davies, “Monitoring summer indoor overheating in the London housing stock,” *Energy Build.*, vol. 141, pp. 361–378, 2017, doi: 10.1016/j.enbuild.2017.02.049.

[16] ASHRAE, “ANSI/ASHRAE Standard 55–2017, Thermal Environmental Conditions for Human Occupancy,” Atlanta, GA, 2017.

[17] M. Baborska-Narożny, F. Stevenson, and M. Grudzińska, “Overheating in retrofitted flats: occupant practices, learning and interventions,” *Build. Res. Inf.*, vol. 45, no. 1–2, pp. 40–59, 2017, doi: 10.1080/09613218.2016.1226671.

[18] M. F. Touchie, E. S. Tzekova, J. A. Siegel, B. Purcell, and J. Morier, “Evaluating summertime overheating in multi-unit residential buildings using surveys and in-suite monitoring,” *Therm. Perform. Exter. Envel. Whole Build.*, vol. 2016-Decem, pp. 135–151, 2016.

[19] F. Nicol and M. Humphreys, “Derivation of the adaptive equations for thermal comfort in free-running buildings in European standard EN15251,” *Build. Environ.*, vol. 45, no. 1, pp. 11–17, 2010, doi: 10.1016/j.buildenv.2008.12.013.

[20] *Natural Resources Canada ’ s Office of Energy Efficiency Leading Canadians to Energy Efficiency at Home , at Work and on the Road*. Ottawa: Natural Resources Canada, 2012.

[21] A. Laouadi, M. Bartko, and M. A. Lacasse, “A New Methodology of Evaluation of Overheating in Buildings,” *Energy Build.*, p. 110360, 2020, doi: 10.1016/j.enbuild.2020.110360.

[22] A. Fouillet *et al.*, “Has the impact of heat waves on mortality changed in France since the European heat wave of summer 2003? A study of the 2006 heat wave,” *Int. J. Epidemiol.*, vol. 37, no. 2, pp. 309–317, 2008, doi: 10.1093/ije/dym253.

[23] M. Itani, N. Ghaddar, K. Ghali, and A. Laouadi, “Development of heat stress charts for older people under indoor environmental conditions,” *Energy Build.*, vol. 224, p. 110274, 2020, doi: 10.1016/j.enbuild.2020.110274.

[24] M. Itani, N. Ghaddar, K. Ghali, and A. Laouadi, “Bioheat modeling of elderly and young for prediction of physiological and thermal responses in heat-stressful conditions,” *J. Therm. Biol.*, vol. 88, no. January, p. 102533, 2020, doi: 10.1016/j.jtherbio.2020.102533.

[25] “ecobee SmartSensor.” [Online]. Available: https://www.ecobee.com/en-ca/accessories/smart-temperature-occupancy-sensor/.

[26] “Statistics Canada. Table 38-10-0019-01 Air conditioners,” 2017.

[27] Z. Lin and S. Deng, “A study on the thermal comfort in sleeping environments in the subtropics-Measuring the total insulation values for the bedding systems commonly used in the subtropics,” *Build. Environ.*, vol. 43, no. 5, pp. 905–916, 2008, doi: 10.1016/j.buildenv.2007.01.027.

[28] Z. Lin and S. Deng, “A study on the thermal comfort in sleeping environments in the subtropics-Developing a thermal comfort model for sleeping environments,” *Build. Environ.*, vol. 43, no. 1, pp. 70–81, 2008, doi: 10.1016/j.buildenv.2006.11.026.

[29] “Type of Dwelling Highlight Tables, 2016 Census,” 2016. [Online]. Available: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/td-tl/Table.cfm?Lang=Eng&T=101&SR=1&S=1&O=A&RPP=25.

[30] “Canada’s climate is warming twice as fast as global average,” *Environment and Climate Change Canada*, 2019. [Online]. Available: https://www.canada.ca/en/environment-climate-change/news/2019/04/canadas-climate-is-warming-twice-as-fast-as-global-average.html. [Accessed: 28-Feb-2021].

[31] G. P. Kenny, J. Yardley, C. Brown, R. J. Sigal, and O. Jay, “Heat stress in older individuals and patients with common chronic diseases,” *Cmaj*, vol. 182, no. 10, pp.

1053–1060, 2010, doi: 10.1503/cmaj.081050.

[32] J. Rocklöv, K. Ebi, and B. Forsberg, “Mortality related to temperature and persistent extreme temperatures: A study of cause-specific and age-stratified mortality,” *Occup. Environ. Med.*, vol. 68, no. 7, pp. 531–536, 2011, doi: 10.1136/oem.2010.058818.

[33] R. S. McLeod and M. Swainson, “Chronic overheating in low carbon urban developments in a temperate climate,” *Renew. Sustain. Energy Rev.*, vol. 74, no. February, pp. 201–220, 2017, doi: 10.1016/j.rser.2016.09.106.

[34] M. Gustin, A. Oraiopoulos, R. S. McLeod, and K. J. Lomas, “A new empirical model incorporating spatial interpolation of meteorological data for the prediction of overheating risks in UK dwellings,” *Proc. 33rd PLEA Int. Conf. Des. to Thrive, PLEA 2017*, vol. 3, pp. 3786–3793, 2017.

[35] M. Gustin, R. S. McLeod, K. J. Lomas, G. Petrou, and A. Mavrogianni, “A high-resolution indoor heat-health warning system for dwellings,” *Build. Environ.*, vol. 168, p. 106519, 2020, doi: 10.1016/j.buildenv.2019.106519.

[36] B. Delcroix, J. Le Ny, M. Bernier, M. Azam, B. Qu, and J. S. Venne, “Autoregressive neural networks with exogenous variables for indoor temperature prediction in buildings,” *Build. Simul.*, 2020, doi: 10.1007/s12273-019-0597-2.

[37] G. Mustafaraj, G. Lowry, and J. Chen, “Prediction of room temperature and relative humidity by autoregressive linear and nonlinear neural network models for an open office,” *Energy Build.*, vol. 43, no. 6, pp. 1452–1460, 2011, doi: 10.1016/j.enbuild.2011.02.007.

[38] O. M. Brastein, D. W. U. Perera, C. Pfeifer, and N. O. Skeie, “Parameter estimation for grey-box models of building thermal behaviour,” *Energy Build.*, vol. 169, pp. 58–68, 2018, doi: 10.1016/j.enbuild.2018.03.057.

[39] B. Cui *et al.*, “A hybrid building thermal modeling approach for predicting temperatures in typical, detached, two-story houses,” *Appl. Energy*, vol. 236, no. November 2018, pp. 101–116, 2019, doi: 10.1016/j.apenergy.2018.11.077.

[40] F. Massa Gray and M. Schmidt, “A hybrid approach to thermal building modelling using a combination of Gaussian processes and grey-box models,” *Energy Build.*, vol. 165, pp. 56–63, 2018, doi: 10.1016/j.enbuild.2018.01.039.

[41] S. Prívara, J. Cigler, Z. Váňa, F. Oldewurtel, C. Sagerschnig, and E. Žáčeková, “Building modeling as a crucial part for building predictive control,” *Energy Build.*, vol. 56, pp. 8–22, 2013, doi: 10.1016/j.enbuild.2012.10.024.

[42] D. Yu, V. Street, and A. S. Fung, “Predicting Indoor Temperature From Smart Thermostat,” 2018.

[43] 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,” *Build. Perform. Model. Conf.*, pp. 173–180, 2018.

[44] X. Yu, L. Georges, M. D. Knudsen, I. Sartori, and L. Imsland, “Investigation of the Model Structure for Low-Order Grey-Box Modelling of Residential Buildings NTNU , Dept . of Energy and Process Engineering , Trondheim , Norway AU , Dept . of Indoor Climate and Energy , Aarhus , Denmark SINTEF Building and Infrastruct,” pp. 5076–5083, 2019.

[45] G. Baasch, A. Wicikowski, G. Faure, and R. Evins, “Comparing gray box methods to derive building properties from smart thermostat data,” *BuildSys 2019 - Proc. 6th ACM*

Int. Conf. Syst. Energy-Efficient Build. Cities, Transp., pp. 223–232, 2019, doi: 10.1145/3360322.3360836.

[46] M. Hossain, T. Zhang, and O. Ardakanian, “Identifying Grey-box Thermal Models with Bayesian Neural Networks.”

[47] M. M. Hossain, T. Zhang, and O. Ardakanian, “Evaluating the feasibility of reusing pre-trained thermal models in the residential sector,” *UrbSys 2019 - Proc. 1st ACM Int. Work. Urban Build. Energy Sensing, Control. Big Data Anal. Vis. Part BuildSys 2019*, pp. 23–32, 2019, doi: 10.1145/3363459.3363529.

[48] M. Gustin, R. S. McLeod, and K. J. Lomas, “Can semi-parametric additive models outperform linear models, when forecasting indoor temperatures in free-running buildings?,” *Energy Build.*, vol. 193, pp. 250–266, 2019, doi: 10.1016/j.enbuild.2019.03.048.

[49] 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 Build.*, vol. 82, pp. 263–274, 2014, doi: 10.1016/j.enbuild.2014.07.025.

[50] J. E. Seem, S. A. Klein, W. A. Beckman, and J. W. Mitchell, “Transfer functions for efficient calculation of multidimensional transient heat transfer,” *J. Heat Transfer*, vol. 111, no. 1, pp. 5–12, 1989, doi: 10.1115/1.3250659.

[51] J. L. Nguyen, J. Schwartz, and D. W. Dockery, “The relationship between indoor and outdoor temperature, apparent temperature, relative humidity, and absolute humidity,” *Indoor Air*, vol. 24, no. 1, pp. 103–112, 2014, doi: 10.1111/ina.12052.

[52] R. Chandra, Y. S. Ong, and C. K. Goh, “Co-evolutionary multi-task learning with predictive recurrence for multi-step chaotic time series prediction,” *Neurocomputing*, vol. 243, pp. 21–34, 2017, doi: 10.1016/j.neucom.2017.02.065.