A thermal control methodology based on a machine learning forecasting model for indoor heating

Makram Abdellatif, Julien Chamoin, Jean-Marie Nianga, Didier Defer

A Thermal Control Methodology Based on a Machine Learning Forecasting Model for Indoor Heating

Makram ABDELLATIF[1], Julien CHAMOIN[1], Jean-Marie NIANGA[2], Didier DEFER[1]

[1] *Yncréa, ULR 4515, Laboratoire de Génie Civil et géo-Environnement (LGCgE), Lille, F-59000, France*

[2] *Yncréa, ULR 7512, Unité Mécanique de Lille (UML), Lille, F-59000, France*

makram.abdellatif@yncrea.fr

Abstract

To optimize use of the data generated from buildings, this document proposes a methodology based on a machine learning model in order to improve thermal comfort and energy efficiency. This methodology relies on measured data (e.g. indoor/outdoor temperature, relative humidity) and forecasted data (e.g. meteorological data) to train a multiple linear regression model to forecast the indoor temperature of the space under study. Applying the genetic algorithm optimization method, this model is then used to evaluate the various heating strategies defined. For each strategy, a score is assigned according to user-defined criteria as a means of prioritizing them and selecting the best one. By studying an office building simulated with the TRNSYS software, a multiple linear regression model could be implemented with errors of less than 1% and an adjusted R^2 coefficient close to 0.9. Compared to a conventional heating strategy, this methodology is capable of improving thermal comfort by up to 43%.

I. Introduction

In Europe, buildings are responsible for about 40% of total energy consumption and emit 36% of total CO_2 emissions, according to the European Commission (European Commission 2014). In addition, people spend on average roughly 90% of their time indoors, as reported by the Environmental Protection Agency (EPA 2012). It is therefore essential for buildings to provide acceptable indoor comfort conditions, allowing users to enjoy a healthy environment and optimize their performance while ensuring an efficient use of building energy.

In this context, several research studies in different fields have investigated various ways to improve indoor comfort and energy efficiency throughout the building life cycle. Regarding the design and construction phases, Fontenelle and Bastos (2014) developed a multi-criteria method for specifying the windows of an office building designed for Rio de Janeiro (Brazil), as a compromise between views of the landscape, daylight entering the work space and energy efficiency. In focusing on innovative materials, Saafi and Daouas (2019) studied the benefits of integrating phase change materials into building envelopes under Tunisian climatic conditions and assessed the effects of such materials on thermal insulation. In their search for the best Heating, Ventilation and Air Conditioning system in an energy retrofit operation, Gustafsson *et al.* (2014) ran a Dynamic Thermal Simulation with TRNSYS software to compare the energy performance of three different systems. Other authors have sought more comprehensive methods; Najjar *et al.* (2019) set up a novel framework for integrating mathematical optimization, Building Information Modeling, and Life Cycle Assessment to enhance operating energy efficiency. They concluded that sustainable building decisions can be achieved by optimizing the material selection and assessment of environmental impacts. Saafi and Daouas (2019) developed an integrated decision support tool to evaluate existing office buildings and recommend an optimal set of sustainable renovation measures, in considering tradeoffs between renovation cost, improved building quality and environmental impacts.

Concerning the operational phase, a significant difference has been found between the predicted (as computed during the design phase) and the measured (during the operational phase) energy performance of buildings (De Wilde 2014). According to a UCL Energy Institute report on data compiled from the CarbonBuzz platform, which records the energy consumption of buildings across Europe, in 2013 for office buildings, the gap between the mean design total energy and mean actual total energy was 1.59 for heating and 1.71 for

electricity (UCL Energy Institute 2013). This difference may be due to several factors, including: i) prediction performance: oversimplified and/or unrealistic design assumptions (occupancy behavior, construction quality, management, control, etc.), inaccuracies, restrictive or oversimplified models (not representative of reality); and ii) actual performance: occupancy behavior (e.g. opening windows), building quality and management control (e.g. inappropriate strategies) (Menezes *et al.* 2012). In this context, several authors have worked on optimizing HVAC system control and building data-based modeling. Benzaama *et al.* (2020) proposed a new methodology for modeling the thermal behavior of residential buildings, using the PieceWise AutoRegressive eXogenous (PWARX) model. They concluded that solar radiation and heating power are the main influential inputs for modeling thermal behavior, with an influence index of 20% and 70%, respectively. In order to determine the optimal start time of the setback temperature during the normal occupancy period of a building, Moon and Jung (2016) set up a predictive Artificial Neural Network (ANN). With the aim of maintaining indoor thermal control, Hang and Kim (2018) demonstrated an enhanced predictive control model based on a Multiple Linear Regression (MLR) model applied on an Internet of Things (IoT) smart space, using Predicted Mean Vote (PMV) to express users' comfort satisfaction. Similarly, Brik *et al.* (2019) proposed ThermCont, a MLR-based tool to predict and control occupants' thermal comfort through the PMV model in real time in office buildings. They used a Genetic Algorithm (GA) to optimize the parameter values of thermal comfort.

On the other hand, IoT is increasingly being democratized and introduced in all fields. Within the building sector, the term smart building, or smart city on a larger scale, has become more widespread with the aim of detecting, tracking, locating and monitoring things (e.g. energy consumption, fault detection) (Minoli, Sohraby and Occhiogrosso 2017).

The purpose of this study is to develop a novel anticipatory control approach for heating systems by relying on building-generated data. This approach utilizes a multiple linear regression model capable of forecasting indoor temperature with high accuracy from inputs such as weather data and heating strategy. This model is then coupled with a genetic algorithm to optimize heating system control, while proposing a new strategy that takes the tradeoff between thermal comfort and energy efficiency into account. In the context of this study, as opposed to the studies presented above (which focus on PMV), thermal comfort is represented as an indoor temperature comfort range (defined by the user and/or according to standards) since the study of PMV requires quite extensive measurements and is often more suitable for experimental buildings. This aspect will be addressed in future works.

To achieve this objective, the proposed approach consists of three main levels, as shown in Figure 1: i) *Data Level*: collection and preprocessing of measured data (Tin, relative humidity, etc.) and forecasted data (meteorological data), ii) *Config Level*: definition of user preferences (selection criteria, e.g. comfort range), and iii) *Core Platform Level*: training of the forecasting model with the collected data, testing of the derived heating strategies, and lastly selection of the best heating strategy according to the selection criteria.

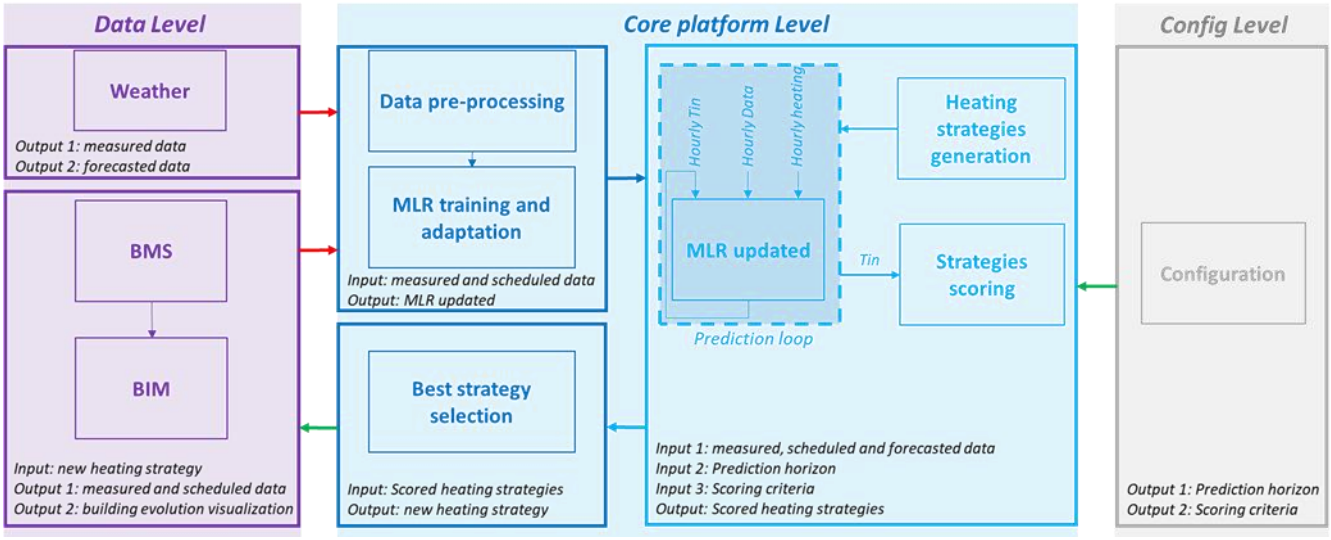

Figure 1: Functioning of the proposed platform

This article is organized as follows. Section 2 presents the development of this novel approach in three parts: description of the indoor temperature forecast model, details on the heating strategy testing process, and

explanation of the new heating strategy selection. Section 3 provides a case study and Section 4 wraps up and concludes the article.

Nomenclature	
ANN	Artificial Neural Network
MLR	Multiple Linear Regression
GA	Genetic Algorithm
BIM	Building Information Model
HVAC	Heating, Ventilation, Air Conditioning
PMV	Predicted Mean Vote
R^2	Coefficient of Determination (0 to 1)
Tin	Indoor temperature (°C)
Tout	Outdoor temperature (°C)
RH	Indoor Relative Humidity (%)
RHout	Outdoor Relative Humidity (%)
Hrad	Horizontal solar radiation (W/m²)
Wvelocity	Wind velocity (m/s)
Wdirection	Wind direction

II. Methodology

This section generates the data input into the TRNSYS software application to simulate an office building located in Nice (France), composed of 8 separately controlled thermal zones as shown in Figure 2. Each zone is equipped with a power-controlled heating system. Initially, all systems are controlled as follows: between 8:00 a.m. and 6:00 p.m., heating power is set to 100% and 0% for the rest of the day (Figure 3) and for weekends. One entire simulation year is considered with a one-hour time step, using a Typical Meteorological Year (TMY) weather file.

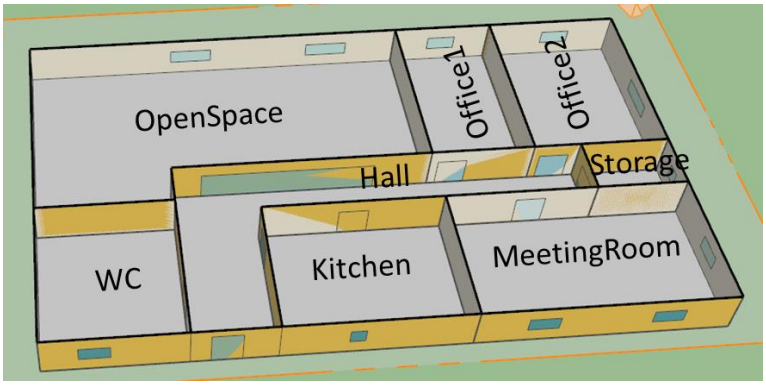

Figure 2: Simulated building layout

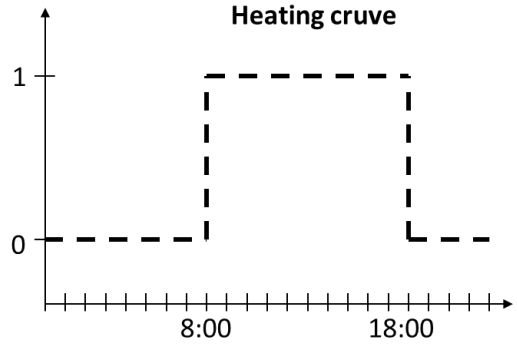

Figure 3: Initial heating system control rule

1. Forecasting model

On the basis of the simulated building, a data-driven model is produced according to a 3-step process yielding highly accurate Tin (indoor temperature) forecasts.

a. Initial models

First of all, the focus lies on selecting the model inputs (predictors). A preselection process was carried out depending on both the availability of data in actual buildings and the physical interpretability of such data. The following data have thus been preselected: heating power demand (Pheat), outdoor temperature (Tout), outdoor relative humidity (RHout), global horizontal radiation (Rad), sky opacity (OpaSky), wind speed (WindS), and wind direction (WindDir). Based on a typical simulation year, three datasets were formed, as shown in Table 1.

Table 1: Dataset splitting

Dataset	Period	Heating state
1	January 1st to May 1st	On

2	May 1st to September 1st	Off
3	September 1st to December 31st	On

A Pearson correlation study was performed on each dataset. The Pearson correlation estimates the linear correlation between two quantitative variables; in other words, it serves as an estimation of the fit of a variable versus another variable by an affine relation obtained through linear regression.

The Pearson correlation coefficient (r) was used to assess the linear relationship between two variables (output-input); it can range from -1 to +1. A positive r indicates a positive linear correlation while a negative r reveals a negative linear correlation. When the value is closer to +1 or -1, the linear correlation is strong (Fu *et al.* 2020). Generally speaking, it is considered that:

- $|r| \leq 0.39$ represents weak correlations,
- $|r|$ between 0.40 and 0.69 represents moderate correlations,
- $|r|$ between 0.70 and 0.90 represents strong or high correlations,
- $|r| \geq 0.9$ represents very high correlations (Fu *et al.* 2020).

Table 2: Correlation table between the variables contained in Dataset 1

	Tin	Tout	Pheat	Rad	RHout	WindS	WindDir	OpaSky
Tin		0.546	0.395	0.452	0.014	-0.059	0.015	-0.002
Tout	0.546		0.103	0.512	-0.525	0.159	0.014	-0.100
Pheat	0.395	0.103		0.510	-0.113	-0.031	0.085	-0.003
Rad	0.452	0.512	0.510		-0.473	0.119	0.083	-0.255
RHout	0.014	-0.525	-0.113	-0.473		-0.281	-0.033	0.479
WindS	-0.059	0.159	-0.031	0.119	-0.281		0.025	-0.135
WindDir	0.015	0.014	0.085	0.083	-0.033	0.025		-0.047
OpaSky	-0.003	-0.100	-0.003	-0.255	0.472	-0.135	-0.048	

-

Table 2 shows that during the first heating period of the year (i.e. Dataset 1), the correlation of Tout with Tin is greater than 0.5 and the correlation of Pheat with Rad is close to 0.5 (0.395 and 0.452, respectively), while the other variables display correlations of close to 0.

Table 3: Correlation table between the variables contained in Dataset 2

	Tin	Tout	Pheat	Rad	RHout	WindS	WindDir	OpaSky
Tin		0.710	NaN	0.061	-0.184	-0.062	0.014	-0.152
Tout	0.710		NaN	0.357	-0.615	0.166	-0.025	-0.142
Pheat	NaN	NaN		NaN	NaN	NaN	NaN	NaN
Rad	0.061	0.357	NaN		-0.590	0.239	-0.024	-0.231
RHout	-0.184	-0.615	NaN	-0.590		-0.364	-0.019	0.359
WindS	-0.062	0.166	NaN	0.239	-0.364		0.042	-0.045
WindDir	0.014	-0.025	NaN	-0.024	-0.019	0.042		-0.046
OpaSky	-0.152	-0.142	NaN	-0.231	0.360	-0.045	-0.046	

Table 3, which presents the results of the correlation study on Dataset 2 (heating turned off), shows that only Tout has a high correlation with Tin, i.e. a correlation coefficient of 0.710. All other variables display a correlation coefficient of close to 0, except for RHout at -0.184.

Table 4: Correlation table between the variables contained in Dataset 3

	Tin	Tout	Pheat	Rad	RHout	WindS	WindDir	OpaSky
Tin		0.799	0.240	0.356	0.207	-0.100	-0.038	0.008
Tout	0.799		0.115	0.463	-0.188	0.001	-0.044	-0.044
Pheat	0.240	0.115		0.500	-0.136	0.022	-0.057	-0.019

Rad	0.356	0.463	0.500		-0.422	0.044	-0.012	-0.202
RHout	0.207	-0.188	-0.136	-0.422		-0.171	-0.029	0.348
WindS	-0.100	0.001	0.022	0.044	-0.171		-0.001	0.082
WindDir	-0.038	-0.044	-0.057	-0.012	-0.029	-0.001		-0.013
OpaSky	0.008	-0.044	-0.020	-0.203	0.348	0.082	-0.014	

Table 4 shows a strong correlation between Tout and Tin during the second heating period (Dataset 3). Rad and RHout have low, but not necessarily negligible, correlation coefficients of 0.356 and 0.207, respectively.

Table 5: Correlation table between variables throughout the year

	Tin	**Tout**	**Pheat**	**Rad**	**RHout**	**WindS**	**WindDir**	**OpaSky**
Tin		0.701	0.125	0.226	0.112	-0.097	0.010	-0.0829
Tout	0.701		-0.171	0.442	-0.201	-0.017	0.025	-0.153
Pheat	0.125	-0.171		0.199	-0.135	0.042	-0.012	0.033
Rad	0.226	0.442	0.199		-0.441	0.103	0.026	-0.242
RHout	0.112	-0.201	-0.135	-0.441		-0.277	-0.020	0.369
WindS	-0.097	-0.017	0.042	0.103	-0.277		0.013	-0.014
WindDir	0.010	0.025	-0.012	0.026	-0.020	0.013		-0.042
OpaSky	-0.083	-0.153	0.033	-0.242	0.369	-0.014	-0.043	

Regarding the full year data (Table 5), it is observed that Tout has a high correlation coefficient with Tin, i.e. 0.701, and the correlation coefficient of Rad with Tin is qualified as non-negligible, i.e. 0.226. In contrast, all other variables have correlation coefficients with Tin of less than 0.2 and for some close to 0 (WindS, WindDir and OpaSky).

The results of these correlation studies show that Tout has a strong correlation with Tin throughout the year. As for Rad and Pheat, their correlations remain quite low (between 0.12 and 0.45), which can be explained by the intermittency of their effect since Rad depends on the presence of sunlight and Pheat on the building occupancy. The other variables have very low correlations (< 0.1), except for RHout with a maximum correlation of 0.2 when the correlation of Tout is 0.799, reflecting the strong correlation between Tout and RHout. It was thus decided to retain Tout, Rad and Pheat.

Next, in examining the correlation between predictors, a high value between two variables indicates their collinearity, hence they share part of the variance. At the model level, this does not necessarily mean that each predictor's effects are cumulative, but rather that the common part plus the specific part of each predictor is cumulative. Subsequently, an analysis of model residuals will confirm (or refute) whether the predictors effectively explain the dependent variable. In other words, the dependent variable is explained by the dependent variables plus a residual part attributed to errors; it is therefore preferable for the residual part to be small.

Based on the above analysis, a model using Tout, Rad and Pheat as inputs can be set up as shown in Figure 4.

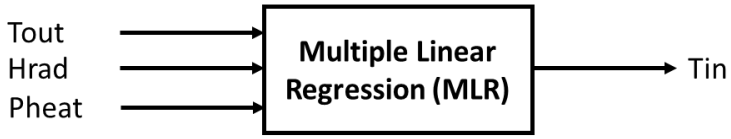

Figure 4: MLR model architecture

The selected data are normalized and model parameters then estimated using the Least Mean Square method (Ciulla and D'Amico 2019). The model equation can be written according to the following general formula (Eq 1):

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad \textit{Eq 1}$$

With, for i = n observations:

- y: dependent variable
- x_i: explanatory variables

- β_0: constant term
- β_p: slope coefficients for each explanatory variable
- ε: model error (residuals)

To estimate the performance of each model, three statistical metrics are employed:

- Adjusted R² (R^2_{adj}): the R² determination coefficient provides an estimate of the quality of model prediction, by measuring the fit between observed and obtained data. The advantage of R^2_{adj} compared to R² is the ability to consider the number of predictors;
- Mean Absolute Percentage Error (MAPE): average of the absolute deviations from the observed values. This is a percentage and therefore a practical indicator for purposes of comparison;
- Root Mean Square Error (RMSE): the square root of the MSE, which is the arithmetic mean of the errors in squares of the differences between model predictions and observations.

To compare models and evaluate performance, the cross-validation method is used. Two techniques are generally available for this type of validation technique: i) the original dataset is divided into two parts, with the first one called the training part and the second one the test part (the model is built on the training part and its performance calculated on the test part, with the split typically being 2/3 training set and 1/3 test set); ii) the test part is left for the final evaluation and the training part divided into k smaller parts (folds), with the model being trained on k-1 parts and tested on the remaining part. This latter operation is repeated k times and k results are obtained. An average is then calculated to determine model performance. This method is referred to as "k-fold cross-validation".

In this study, the 10-fold cross-validation method has been used (i.e. k=10 folds). Results obtained with the model developed on Datasets 1, 2 and 3 are depicted in Figure 5, Figure 6 and Figure 7, respectively. Their average scores are collated in Table 6.

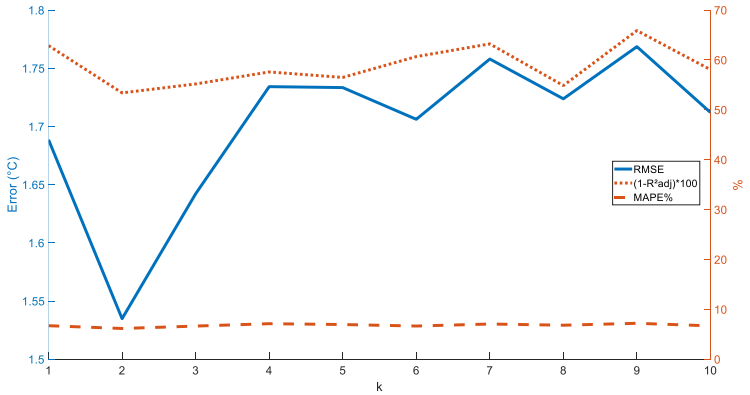

Figure 5: Evolution of model performance vs. K (folds) using Dataset 1

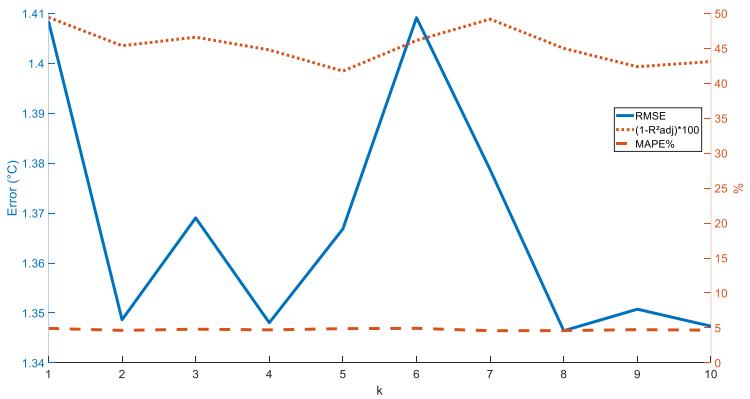

Figure 6: Evolution of model performance vs. K (folds) using Dataset 2

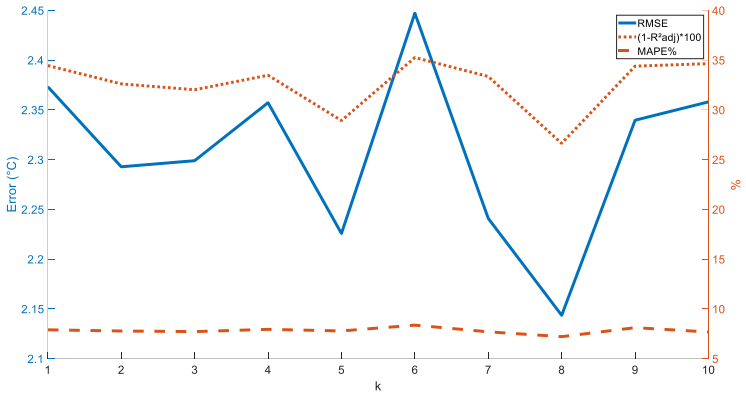

Figure 7: Evolution of model performance vs. K (folds) using Dataset 3

Table 6 lists the average performances obtained by testing the model on the 3 datasets. From a close inspection of the model, all predictors have p-values below 5%, which is considered to be the threshold of significance; therefore, all predictors are significant (Kolasa-Wiecek 2015). This model will thus be used for the remainder of the study.

Table 6: Average Model 1 performance using the 3 datasets

		R^2_{adj}	MAPE (%)	RMSE (°C)
Model 1	**Dataset 1**	0.41	6.80	1.70
	Dataset 2	0.55	4.76	1.37
	Dataset 3	0.67	7.81	2.31

b. *Model optimization*

The selected model features an R^2_{adj}, MAPE and RMSE lying between 0.41 and 0.67, between 4.76 and 7.81 and between 1.37 and 2.31, respectively. Such performances however are insufficient for purposes of this study. The objective of this section is to improve model performance by maximizing R^2 and minimizing both MAPE and RMSE.

To achieve this goal, the authors have opted to enhance each of the initial datasets with historical values. The dataset is considered as a statistical population composed of n static individuals; in turn, each individual is composed of one (or more) outputs (Tin) and predictors (Tout, Rad and Pheat). Adding historical data consists of expanding each individual with its previous data series. Individual i, which contains data measured at time t, will be enhanced by data measured at time t-D (during the previous hour if D=1). To test this approach, only the Tin of the past hour has initially been added. The resulting model (Model 2) can be written in the form of Eq 2 and exhibits the performance reported in Table 7.

$$\widehat{Tin}_i = \beta_0 + \beta_1 Tout_i + \beta_2 Rad_i + \beta_3 Pheat_i + \beta_4 Tin_{i-1} \qquad \textit{Eq 2}$$

Table 7: Average Model 2 performance using the 3 datasets

		R^2_{adj}	MAPE (%)	RMSE (°C)
Model 2	**Dataset 1**	0.97	1.05	0.34
	Dataset 2	0.99	0.12	0.04
	Dataset 3	0.99	0.84	0.34

By only adding the Tin of the past hour, the average MAPE over the 3 datasets improves from 6.45 to 0.67.

A parametric study was conducted by varying the D of each predictor from 0 to 3, as follows:

$$\widehat{Tin}_i = \beta_0 + \beta_1 Tout_i + \beta_2 Rad_i + \beta_3 Pheat_i + \sum_1^4 \beta_i Tin_{i-D} + \sum_1^4 \beta_i Tout_{i-D} + \sum_1^4 \beta_j Rad_{i-D} + \sum_1^4 \beta_k Pheat_{i-D} \qquad \textit{Eq 3}$$

Next, a parametric study was conducted by varying the D of each predictor from 0 to 4 (Eq 3), in order to identify the best delay for each parameter. A total of 32,768 models on each dataset were studied, yielding Eq 4, Eq 5 and Eq 6, corresponding to Datasets 1, 2 and 3, respectively.

$$\widehat{Tin}_i = \beta_0 + \beta_1 Tout_i + \beta_2 Pheat_i + \beta_3 Hrad_i + \beta_4 Tin_{i-1} + \beta_5 Tout_{i-1} + \beta_6 Pheat_{i-1} + \beta_7 Tin_{i-2} + \beta_8 Tout_{i-2} + \beta_9 Pheat_{i-2} + \beta_{10} Tin_{i-3} + \beta_{11} Tout_{i-3} + \beta_{12} Pheat_{i-3} \quad \textit{Eq 4}$$

$$\widehat{Tin}_i = \beta_0 + \beta_1 Tout_i + \beta_2 Tin_{i-1} + \beta_3 Tout_{i-1} + \beta_4 Tin_{i-2} + \beta_5 Tout_{i-2} + \beta_6 Tin_{i-3} + \beta_7 Tout_{i-3} + \beta_8 Hrad_{i-3} \quad \textit{Eq 5}$$

$$\widehat{Tin}_i = \beta_0 + \beta_1 Tout_i + \beta_2 Pheat_i + \beta_3 Tin_{i-1} + \beta_4 Tout_{i-1} + \beta_5 Pheat_{i-1} + \beta_6 Tin_{i-2} + \beta_7 Tout_{i-2} + \beta_8 Pheat_{i-2} + \beta_9 Tin_{i-3} + \beta_{10} Tout_{i-3} + \beta_6 Pheat_{i-3} + \beta_{12} Hrad_{i-3} \quad \textit{Eq 6}$$

By comparing the equations obtained, it turns out, unsurprisingly, that Eqs 4 and 6 are nearly identical by virtue of the fact that these two models are driven over two heating periods. Eq 4 assumes the instantaneous value of Rad, while Eq 6 takes its value with a 2-hour delay. As for Eq 5, heating power is excluded since this is a non-heating period. Since the objective of this work is to implement a scalable model for use throughout the year and since the goal is to optimize heating control, the model based on Eq 5 has been eliminated from consideration.

Table 8: Average performance of Models 3 and 4 using the 3 datasets

		R^2_{adj}	MAPE (%)	RMSE (°C)
Model 3	**Dataset 1**	0.99	0.07	0.02
	Dataset 2	0.99	0.07	0.02
	Dataset 3	1	0.05	0.02
Model 4	**Dataset 1**	0.99	0.07	0.02
	Dataset 2	0.99	0.07	0.02
	Dataset 3	1	0.05	0.02

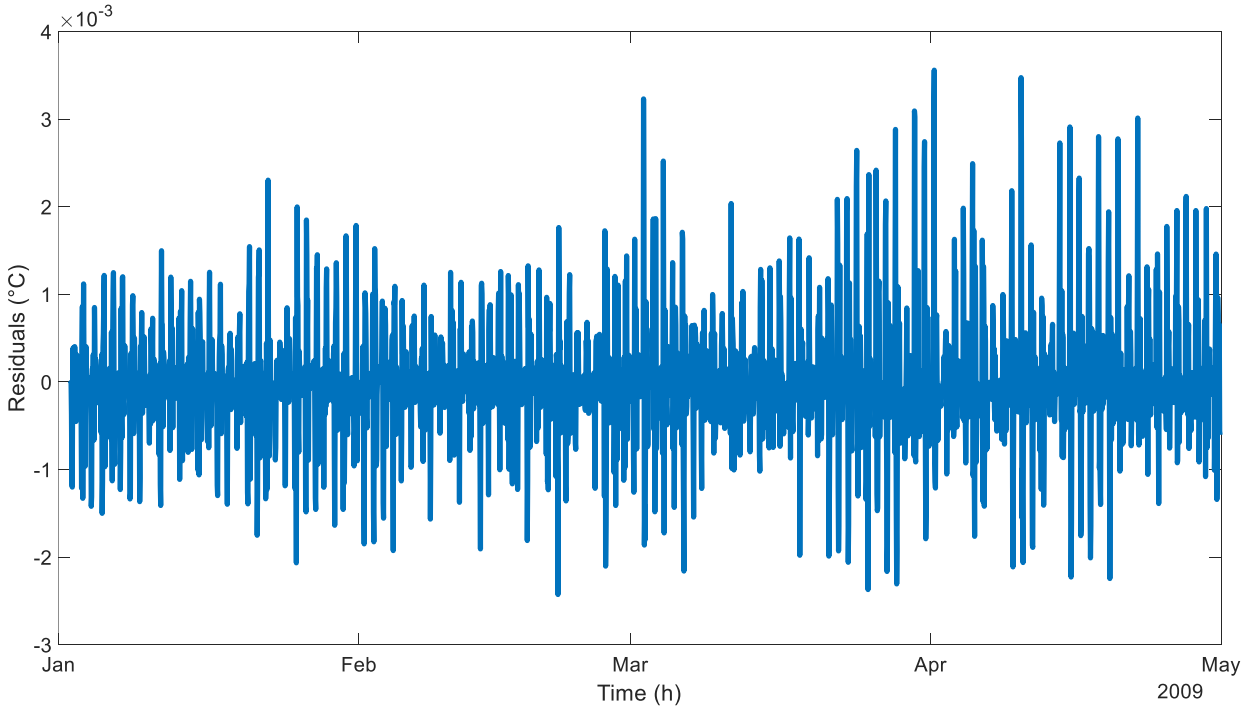

Figure 8: The model's raw residuals

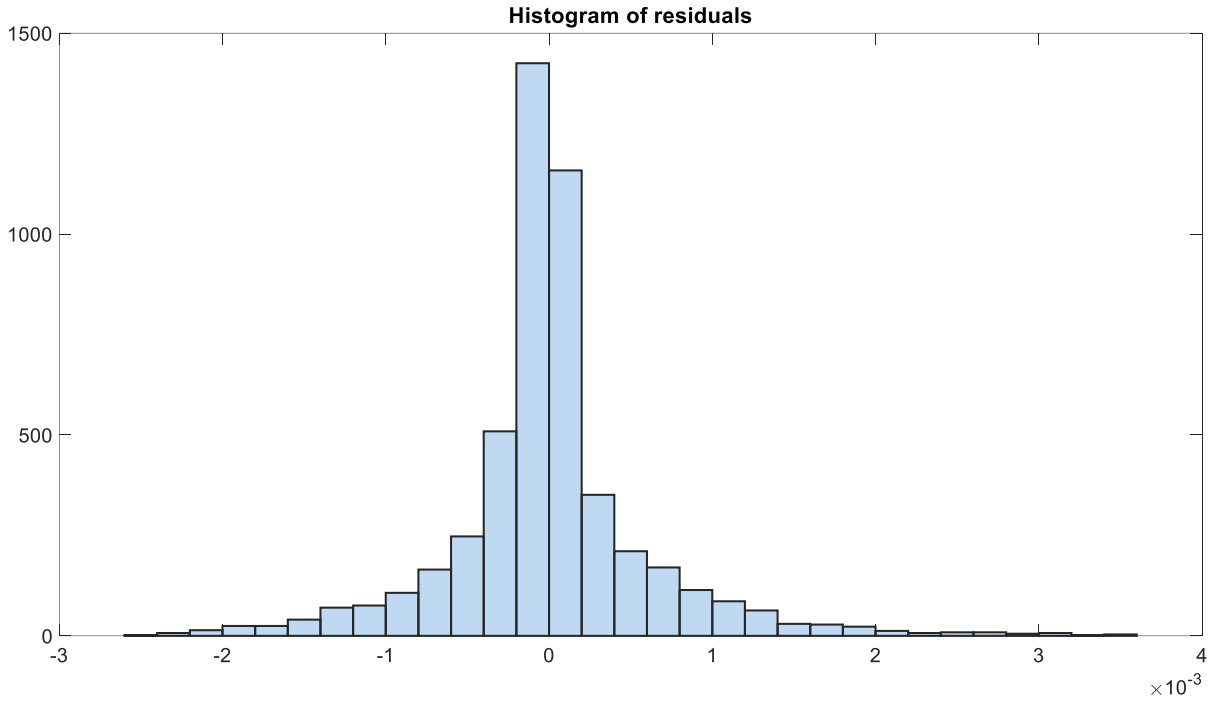

Figure 9: The model's residuals distribution

In examining the residuals presented in Figure 8 and Figure 9, it can be noticed that:

- After fitting the model on the training dataset, its residual errors are independent and identically distributed (i.e. random variables),
- The residual errors are normally distributed (Figure 9),
- The residual errors reveal constant variance (i.e. homoscedastic).

Table 8 displays the performance of Models 3 and 4, as based on Eq 4 and Eq 6 respectively. The highly accurate performances of these two models are nearly identical. Assuming the effect of the sun to be non-instantaneous, the authors selected Model 4 for the remainder of the study. Figure 10 shows an example of results obtained when testing Model 4 on a portion of Dataset 3. The forecasted Tin (solid line) lies very close to the observed Tin (dashed line). Figure 11 shows a marked linear relationship between the response variable (forecasted Tin) and the observation.

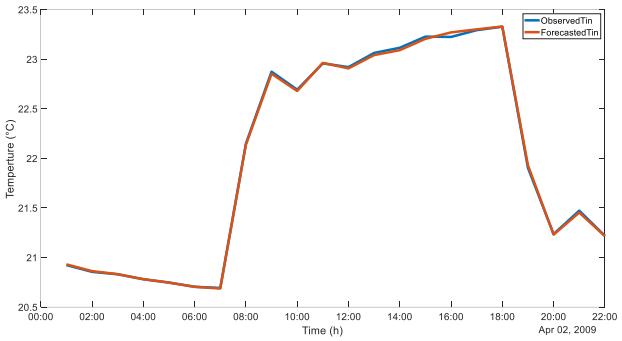

Figure 10: Evolution of the measured Tin and the forecasted Tin between the December 30th at 7 pm and January 1st at 12 am

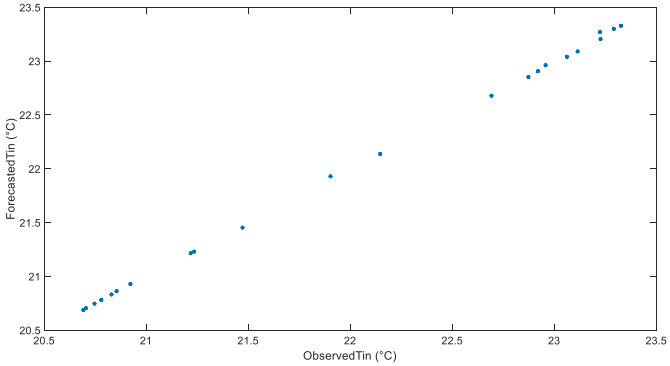

Figure 11: Response variable (forecasted) vs. observed variable

In conclusion, an analysis of the above residuals has enabled verifying the 3 hypotheses necessary for model validation, namely: normality assumption, equality of variances assumption, and independence assumption (Pineda Becerril *et al.* 2019).

2. Selection of heating strategies

In order to improve thermal comfort and energy consumption of the studied zone, an optimal heating strategy, established on the basis of predefined criteria (comfort interval, occupancy period, etc.), must be identified. A

heating strategy is a control sequence composed of n setpoint values, whereby each setpoint corresponds to a defined time step that depends on both building type (inertia) and heating system type. As an example, for a power-controlled heating system with an hourly step, each setpoint corresponds to one hour. To identify the optimal heating strategy, one solution worth considering is the brute force optimization method, which as shown in a previous work (Abdellatif, Chamoin, and Defer 2019) consists of testing all possible solutions and then choosing the best one. However, this approach can become very time-consuming. As an example, for a heating system with 2 modes (Mode 1: 100% power, Mode 2: 0%) and a 24-hour sequence, the number of solutions to be tested equals 2^{24}, i.e. over 16 million. To avoid this kind of constraint, two solutions can be considered: the first one calls for shortening the command sequence, while the second one uses a more comprehensive optimization method.

In this study, the authors opted for the second solution since heating systems often have several modes (e.g. from 0% to 100% with a 1% step); moreover, limiting the control sequence would prevent anticipating potential anomalies. As such, the genetic algorithm optimization method has been chosen.

Genetic algorithms are a family of computational models inspired by evolution. These algorithms encode a potential solution to a specific problem onto a simple chromosome-like data structure and then apply recombination operators to these structures so as to preserve critical information (Whitley 1994). Genetic algorithms are often viewed as function optimizers, although the range of problems suitable for these algorithms is quite broad. The purpose of genetic algorithms is to obtain an approximate solution to a specific optimization problem when no precise method exists to solve it within a reasonable amount of time (or when the solution is unknown); they attempt to simulate the natural evolutionary process according to the Darwinian model in a given environment. The focus here is on individuals in a population. Each individual is represented by a chromosome composed of genes containing the hereditary characteristics of the individual. The principles of selection, crossing and mutation are inspired by natural processes (Barnier *et al.* 2014); its general mode of operations is diagrammed in Figure 12.

In this work, the individuals constituting the population are in fact heating strategies. This population then evolves according to three operators: crossover, mutation, and selection. Crossover contributes to renewing the population by combining two parent solutions to yield daughter solutions that form a new population of the same size. Mutation consists of changing (mutating) a single element of the population solutions. Selection is critical to keeping the best individuals for the fitness function. Algorithm performance depends heavily on the stopping criterion (e.g. iteration number) (Janbain 2020). For this study, the fitness function is composed of the forecasting model and score calculation function. Each heating strategy generated is tested using the forecasting model, in deriving its corresponding Tin. According to the predefined comfort criteria, a score is assigned to each heating strategy (calculated on both Tin) and energy consumption level.

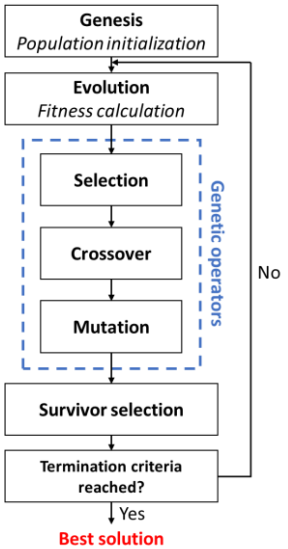

Figure 12: General flowchart of the genetic algorithm

III. Results

To test the logic developed herein, the simulated building has been cloned. Heating of the initial building is controlled according to a deterministic scenario, as described in Section 2, while heating of the cloned building is controlled according to the adaptive logic described above. The authors have defined the comfort criteria listed in Table 9.

Table 9: Comfort criteria relative to building occupancy

		Weekdays	Weekend
Occupied	8:00 am to 6:00 pm	[21 ; 24]	16
Unoccupied	6:00 pm to 8:00 am	16	16

This study focuses on two distinct time periods: from February 17 to March 3, and from December 6 to December 16. Two databases were initially built: one of measured parameters and a second of forecasted parameters (e.g. weather). The weather forecast data herein stem from the weather file used in the TRNSYS simulation. Data from these two databases have been cleaned and preprocessed by means of removing outliers, replacing missing values and performing normalization to standardize the effect of all parameters.

At the first iteration, the MLR model receives a 24-hour data matrix composed of the weather forecast data, the heating strategy over the next 24 hours, and historical data relevant to each parameter. The historical Tin data are extracted from the database of measurement values and replaced by forecasted data whenever the measurements are unavailable, as explained in Figure 13. For example, after 3 iterations, to forecast Tin at i+2 (Tin_{i+2}), the Tin measured three hours prior (Tin_{i-1}), two hours prior (Tin_i) and one hour prior (Tin_{i+1}) are all used. Since the loop was started at time i, Tin_{i-1} is available, while Tin_i and Tin_{i+1} are not. These values are therefore replaced by their forecasted values, respectively $\widehat{Tin}_i$ and $\widehat{Tin}_{i+1}$. This protocol runs the risk of propagating errors, especially prediction errors. To mitigate this problem, the heating strategy is recalculated at each time step.

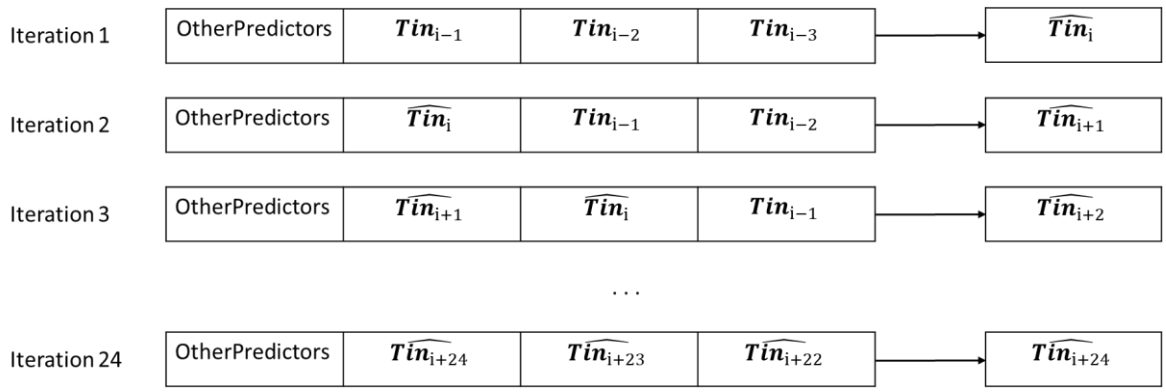

Figure 13: Schematic explanation of the evolution of predictors and predictions during the prediction loop

Using the forecasting model in the fitness function of the genetic algorithm, several heating strategies have been tested. The best strategy is then selected according to the set of predefined criteria, as diagrammed in Figure 14.

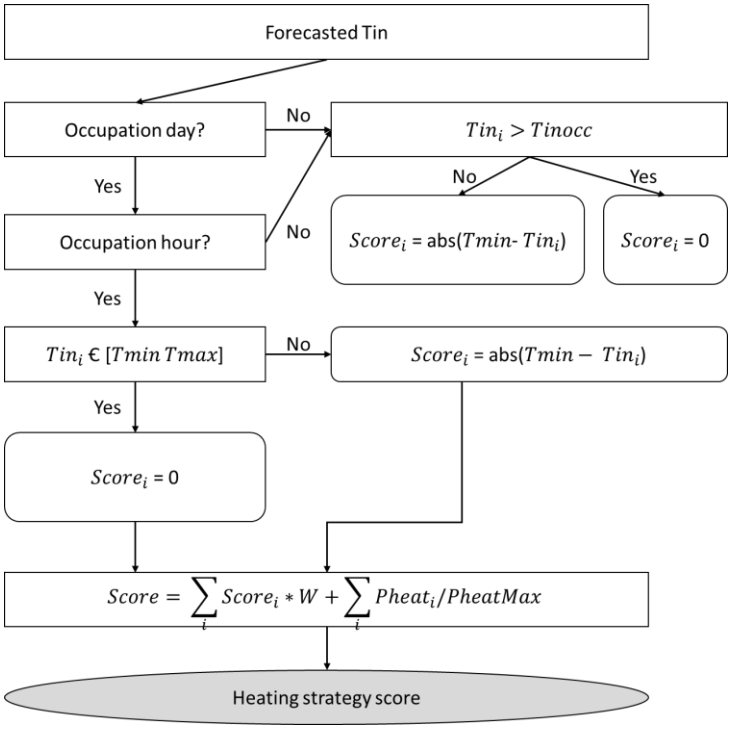

Figure 14. Schematic diagram of the scoring process

Accordingly, the final score of each strategy is the sum of: a first part calculated on the Tin vs. comfort criteria comparison, and a second part calculated on energy consumption (normalized). The first part is multiplied by a weighting coefficient W in order to favor strategies satisfying the comfort criteria. At the end of the optimization process using the applicable genetic algorithm, the heating strategy with the lowest score is selected; this process serves to identify the optimal heating strategy for the next 24 hours, in taking both the weather forecast and the future behavior of the study zone into account.

Figure 15 and Figure 18 present the evolution of Tin both before (initial building) and after (clone building) application of the methodology developed for the periods of February 17 to March 3 and December 6 to December 16, respectively. Moreover, Figure 16 and Figure 19 display the forecasted Tin vs. the Tin of the clone building. This graph follows a near-perfect straight line; it reflects the model's strong performance (less than 1% error). Lastly, Figure 17 and Figure 20 chart the evolution of heating activation and deactivation for these same periods, where 1 corresponds to the heating switch being turned on and 0 to turned off. Two distinct periods are analyzed, one at the beginning of the year the other at the end of the year, in order to demonstrate that the proposed logic works throughout the year.

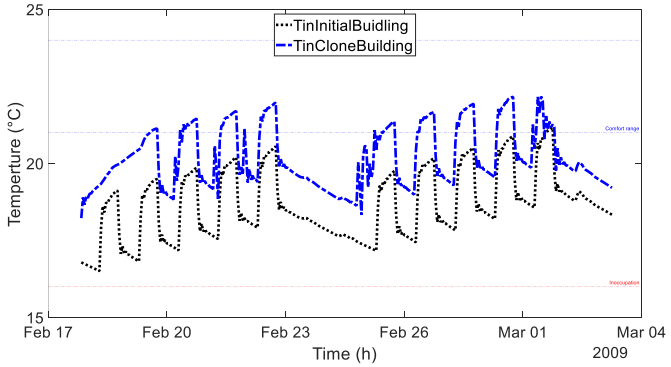

Figure 15: Evolution of Tin for both the initial building and clone building, after the February 17th and March 3rd corrections

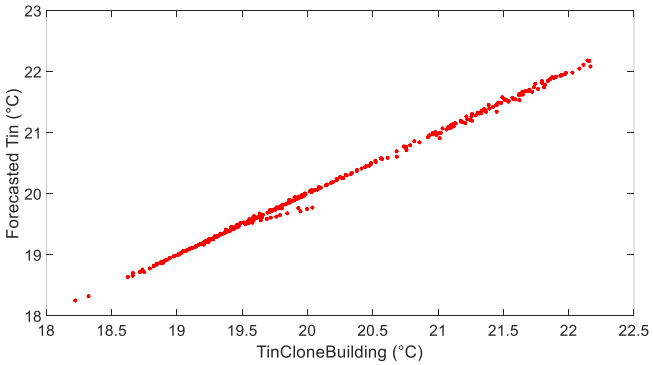

Figure 16: Forecasted Tin vs. clone building Tin, after the February 17th and March 3rd corrections

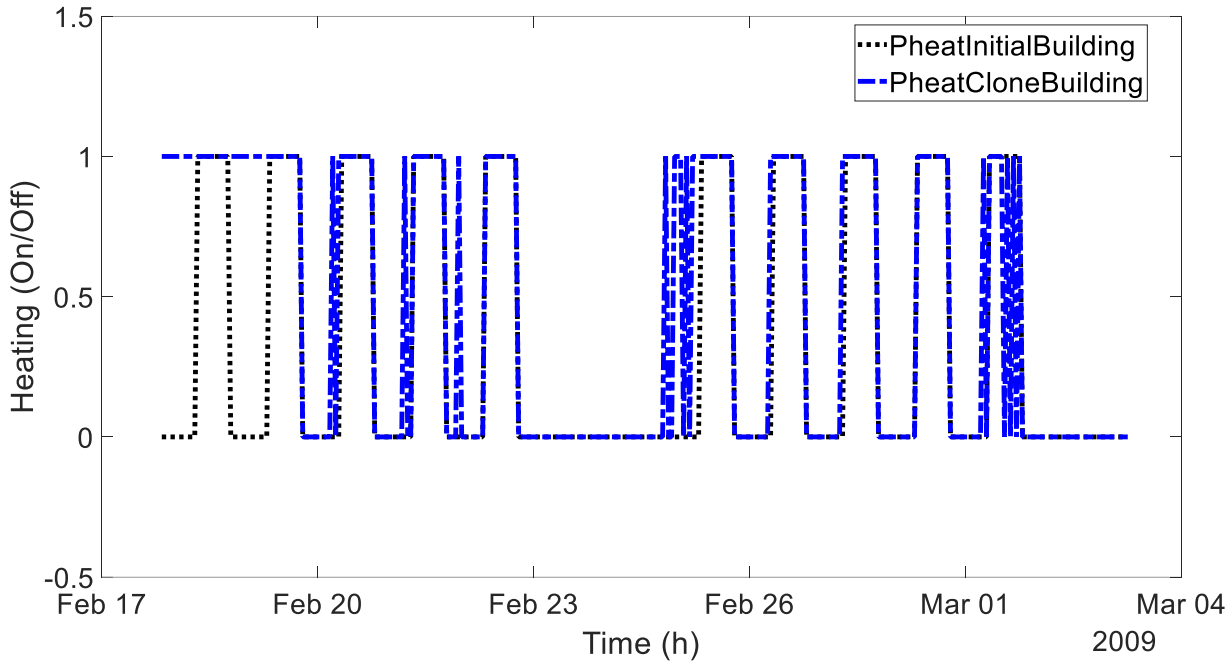

Figure 17: Evolution of heating in both the initial and clone buildings, after the February 17th and March 3rd corrections

In Figure 15 and Figure 18, the black curves show the evolution of Tin for the initial building using a deterministic heating scenario: heating is turned on between 8:00 am and 6:00 pm and turned off outside of this time period, as described in Section 2. This type of heating scenario is commonly used for actual buildings. The blue curves correspond to the Tin of the clone building using the new heating strategies suggested by the logic proposed in this paper.

In this case study, the comfort interval ranges from 21°C to 24°C during occupied hours and above 16°C during unoccupied hours. In Figure 15, the Tin of the initial building (black curve) does not at all meet these comfort criteria. However, by applying the customized heating strategy, the Tin of the clone building fully satisfies these criteria; this could be achieved through anticipating the switching-on and switching-off of the heating, as shown in Figure 17.

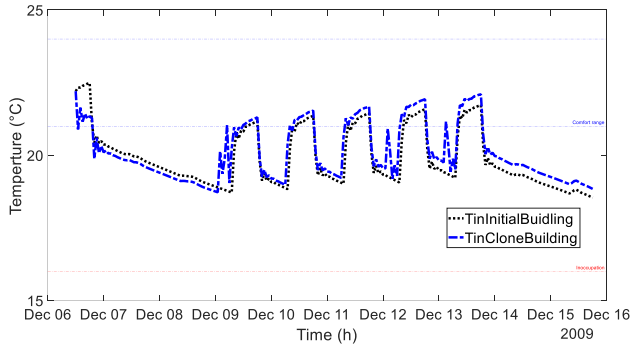

Figure 18: Evolution of Tin for both the initial and clone buildings, after the December 6th and December 16th corrections

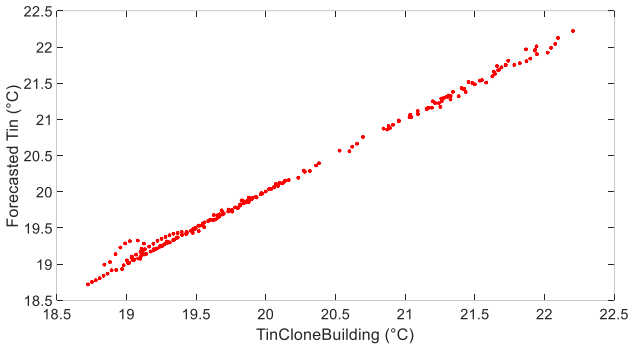

Figure 19. Forecasted Tin vs. clone building Tin, after the December 6th and December 16th corrections

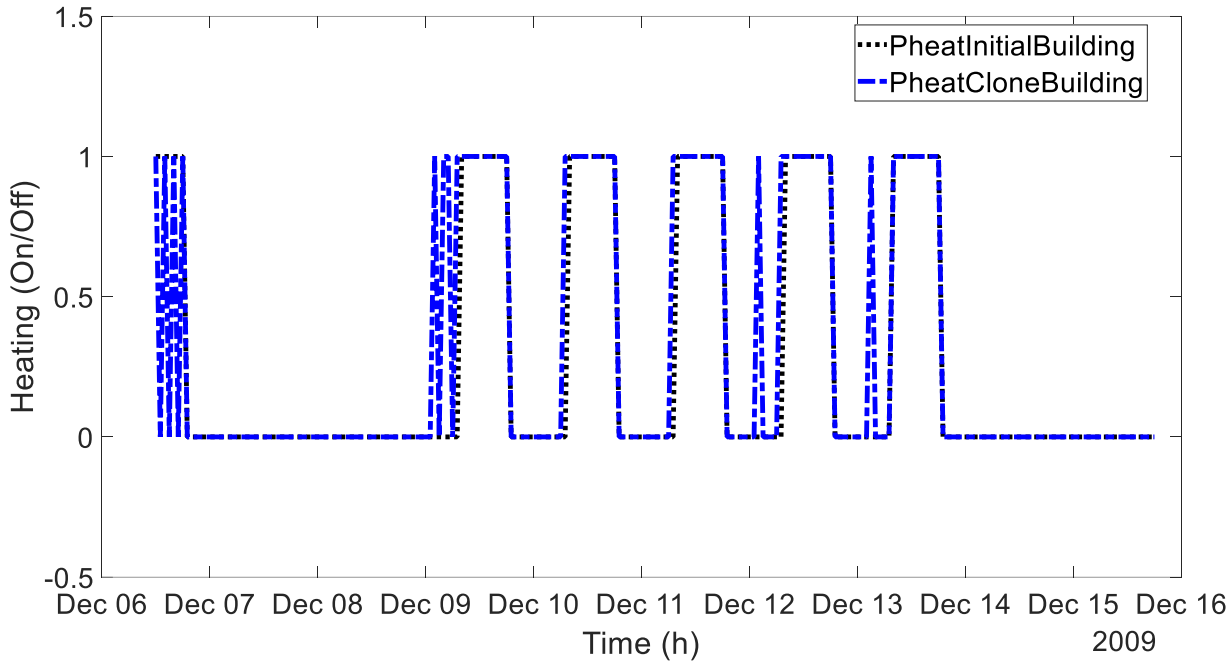

Figure 20: Evolution of heating in both the initial and clone buildings, after the December 6th and December 16th corrections

To the same extent, for the period between December 6th and December 16th (Figure 18), the new heating strategy does allow for the thermal comfort criteria to be met.

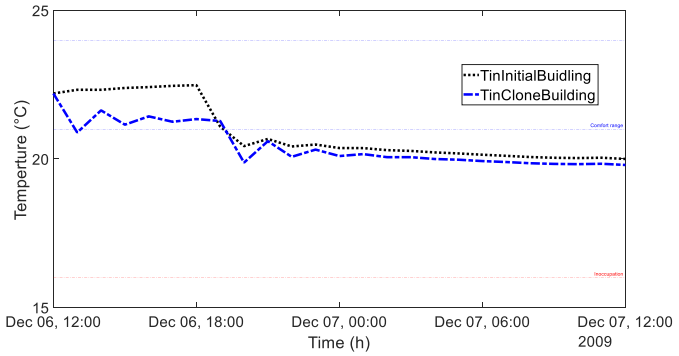

Figure 21: Evolution of Tin for both the initial and clone buildings, after the December 6th and December 7th corrections

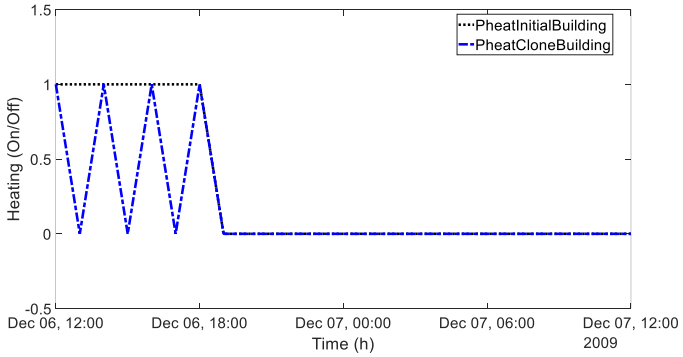

Figure 22: Evolution of heating in both the initial and clone buildings, after the December 6th and December 7th corrections

Heating strategies are chosen to ensure thermal comfort as a priority and to save energy whenever possible. Figure 21 and Figure 22 provide close-up views on the first day of Figure 18 and Figure 20, respectively. It can be noticed that the thermal comfort criteria are being respected, and energy savings have been realized. Indeed, the heating is initially switched on continuously between 12:00 pm and 6:00 pm, while the new heating strategy proposes switching it off for 3 hours. Over this period, an energy savings of nearly 43% could thus be achieved.

IV. Conclusion

In this paper, we have proposed a new thermal control methodology based on a machine learning forecasting model and a genetic algorithm optimization method for indoor heating. This methodology is mainly based on two modules: a MLR model to forecast indoor temperature with an error of less than 1%, and a second module to define new heating strategies in order to improve thermal comfort while seeking to save energy whenever possible. The second module uses a genetic algorithm that integrates the MLR model in order to identify the best heating strategy according to the fitness function.

The MLR forecast model was developed according to a statistical approach implemented herein. By comparing the initial model, which only uses raw data, to the optimized model using both raw data and their historical values, the error was reduced from 8% to less than 1%. This approach is applicable to any type of instrumented building; the model derived is self-adjusting and can be used throughout the year.

The best heating strategy was selected by running a genetic algorithm. Based on the MLR model, this algorithm allowed testing several heating strategies, with each strategy consisting of n heating setpoints corresponding to n time steps. By performing tests, it was found that setting n to 24 is a good compromise between computational time and physical effect. Next, according to the predefined comfort criteria, a score was assigned to each heating strategy.

As part of our future works, an actual building located in northern France will be studied. A propagation error analysis will be conducted during this study. Thereafter, the scoring function will be enhanced with features to encourage energy consumption when least expensive, as well as the integration of parameters that take the comfort perceived by users into account.

V. Acknowledgments

This research work has been supported by VINCI Construction France, Cegelec, Projex and Yncréa under the aegis of the Smart Building as nodes of Smart Grids (SBnodesSG) industrial chair.

References

Abdellatif, Makram, Julien Chamoin, and Didier Defer. 2019. "A Thermal Control Methodology Based on a Predictive Model for Indoor Heating Management." *MATEC Web of Conferences* 295: 01001.

Barnier, Nicolas et al. 2014. "Optimisation Par Algorithme Génétique Sous Contraintes To Cite This Version : Optimisation Par Algorithme Génétique Sous Contraintes." 1(1): 366–74.

Benzaama, M.H., L.H. Rajaoarisoa, B. Ajib, and S. Lecoeuche. 2020. "A Data-Driven Methodology to Predict Thermal Behavior of Residential Buildings Using Piecewise Linear Models." *Journal of Building Engineering* (May): 101523.

Brik, Bouziane, Moez Esseghir, Leila Merghem-Boulahia, and Hichem Snoussi. 2019. "ThermCont: A Machine Learning Enabled Thermal Comfort Control Tool in a Real Time." *2019 15th International Wireless Communications and Mobile Computing Conference, IWCMC 2019*: 294–300.

Ciulla, G., and A. D'Amico. 2019. "Building Energy Performance Forecasting: A Multiple Linear Regression Approach." *Applied Energy* 253(April).

EPA. 2012. *An Introduction for Health Professionals*. http://www.epa.gov/iaq/pdfs/indoor_air_pollution.pdf.

European Commission. 2014. *2030 Climate and Energy Goals for a Competitive, Secure and Low-Carbon EU Economy*. https://ec.europa.eu/clima/news/articles/news_2014012202_en.

Fontenelle, Marília Ramalho, and Leopoldo Eurico Gonçalves Bastos. 2014. "The Multicriteria Approach in the Architecture Conception: Defining Windows for an Office Building in Rio de Janeiro." *Building and Environment* 74: 96–105.

Fu, Ting et al. 2020. "Correlation Research of Phase Angle Variation and Coating Performance by Means of Pearson's Correlation Coefficient." *Progress in Organic Coatings* 139(June 2019): 105459. https://doi.org/10.1016/j.porgcoat.2019.105459.

Hang, Lei, and Do Hyeun Kim. 2018. "Enhanced Model-Based Predictive Control System Based on Fuzzy Logic for Maintaining Thermal Comfort in IoT Smart Space." *Applied Sciences (Switzerland)* 8(7).

Janbain, Ali. 2020. "Utilisation d ' Algorithmes Génétiques Pour l ' Identification Systématique de Réseaux de Gènes Co-Régulés . To Cite This Version : HAL Id : Tel-02464921 DE L ' UNIVERSITÉ DE M ONTPELLIER Utilisation d ' Algorithmes Génétiques Pour l ' Identification Syst."

Kolasa-wiecek, Alicja. 2015. "ScienceDirect Stepwise Multiple Regression Method of Greenhouse Gas Emission Modeling in the Energy Sector in Poland." *JES* 30: 47–54. http://dx.doi.org/10.1016/j.jes.2014.09.037.

Menezes, Anna Carolina, Andrew Cripps, Dino Bouchlaghem, and Richard Buswell. 2012. "Predicted vs. Actual Energy Performance of Non-Domestic Buildings: Using Post-Occupancy Evaluation Data to Reduce the Performance Gap." *Applied Energy* 97: 355–64.

Minoli, Daniel, Kazem Sohraby, and Benedict Occhiogrosso. 2017. "IoT Considerations, Requirements, and Architectures for Smart Buildings-Energy Optimization and Next-Generation Building Management Systems." *IEEE Internet of Things Journal* 4(1): 269–83.

Moon, Jin Woo, and Sung Kwon Jung. 2016. "Algorithm for Optimal Application of the Setback Moment in the Heating Season Using an Artificial Neural Network Model." *Energy and Buildings* 127: 859–69.

Najjar, Mohammad, Karoline Figueiredo, Ahmed W.A. Hammad, and Assed Haddad. 2019. "Integrated Optimization with Building Information Modeling and Life Cycle Assessment for Generating Energy Efficient Buildings." *Applied Energy* 250(April): 1366–82.

Pineda Becerril, Miguel, Omar García, Armando Aguilar, and Frida León. 2019. "Use of Ict for the Use of the Residual Values in Anova." *EDULEARN19 Proceedings* 1(July): 9035–39.

Saafi, Khawla, and Naouel Daouas. 2019. "Energy and Cost Efficiency of Phase Change Materials Integrated in Building Envelopes under Tunisia Mediterranean Climate." *Energy* 187.

UCL Energy Institute. 2013. "Summary of Audits Performed on CarbonBuzz by the UCL Energy Institute." : 1. http://www.carbonbuzz.org/downloads/PerformanceGap.pdf.

Whitley, Darrell. 1994. “A Genetic Algorithm Tutorial.” *Statistics and Computing* 4(2): 65–85.

De Wilde, Pieter. 2014. “The Gap between Predicted and Measured Energy Performance of Buildings: A Framework for Investigation.” *Automation in Construction* 41: 40–49.