RESEARCH ARTICLE

A comparison between grey-box models and neural networks for indoor air temperature prediction in buildings

ARTICLE HISTORY
Compiled December 29, 2023

ABSTRACT
Model Predictive Control has gained much attention due to its potential to improve building operations by reducing costs, integrating renewable energy sources, and increasing thermal comfort. This paper aims to compare the accuracy of grey-box models based on resistance-capacitance (RC) networks and Long-Short-Term Memory (LSTM) neural networks in the prediction of the buildings' thermal response, which is a key feature for the successful implementation of predictive controllers. Indoor air temperature prediction tests have been performed on simulated and measured data from buildings with different thermal insulation and thermal mass during both heating and cooling seasons. Results show that neural networks have, on average, a better prediction performance than grey-box models. Both modeling approaches are affected by the building characteristics and by the season considered. The grey-box models require less training data, although the latter seems to play a role only in the worse-performing tests. When user setpoint changes in the testing phase, the LSTM neural network shows a significant drop in the root mean square error. In conclusion, although LSTM outperforms grey-box models on average, the reduced training data and higher reliability under normal operating conditions, as well as their linearity, make RC models a strong alternative for predictive controllers.

KEYWORDS
Building simulation; Long Short-Term Memory Neural Networks; Grey-box models; Model Predictive Control;

Symbols

α	Absorption coefficient
C	Capacitance
H	Transmittance coefficient
I	Irradiance
ϕ	heat Load
θ	Temperature

Acronyms

BB	Black Box
BEMS	Building Energy Managment System
DHW	Domestic Hot Water
GB	Grey Box
HVAC	Heating ventilation and Air Conditioning
HW	Heavy Weight
LW	Light Weight
LSTM	Long short-term memory
ML	machine Learning
MPC	Model Predictive Control
MSE	Mean Square Error
nC	n-capacitances grey box model
NN	Neural Network
RMSE	Root Mean Square Error

Sub/superscripts

conv	convective
e	external
eq	equivalent
hg	heat gains
i	indoor air
int	internal
m	building mass
op	opaque
s	indoor surfaces
sol	solar
sup	supply
tr	transmission
ve	ventilation
w	window

1. Introduction

Building Energy Management Systems (BEMS) enable to monitor and efficiently operate heating, ventilation, and air-conditioning (HVAC) systems as well as lighting systems and other energy-consuming equipment present in buildings. In addition, BEMS are expected to play an increasingly important role in the energy system due to the need of having flexible and dynamically responsive electric loads to effectively match demand and supply while maintaining high levels of reliability, comfort and efficiency [1]. Despite the great potential offered by intelligent control systems to improve efficiency, thermal comfort, self-consumption and to support grid operators, the majority of buildings today still adopt simple rule-based control (RBC) techniques with only limited energy saving capabilities [2]. The decreasing costs in computation and sensing devices paves the way to the diffusion of computerized building automation systems that adopt advanced control strategies, like model predictive control (MPC). MPC is a flexible control technique since it can handle time-varying external as well as internal constraints and disturbances [3]. It relies on an optimization problem that can take the

form of tracking error, control effort, energy cost, demand cost, power consumption, or a combination of these objectives [4]. Nowadays, the primary barrier to a substantial adoption of MPC in the building industry is the scalability of the technology, since every building is unique, thus substantial effort accompanies each new controller implementation [5]. In the last decade, MPC has become a dominant control strategy in research on intelligent building operation [2]. An alternative to MPC is represented by Reinforcement Learning (RL). The latter is a form of machine learning that consists of an agent that learns what actions to take depending on the state of the environment. The agent learns by trial and error and is rewarded for taking desirable actions [6]. RL can be either model-free or model-based, while MPC relies on a model that reproduces the heat dynamics of the building and of its technical systems.

1.1. Control-oriented building models

Building energy models are commonly split into three groups: white-box, grey-box and black-box models. Many simulation tools used by the building physics community, such as EnergyPlus [7] or TRNSYS [8], are based on white-box models, i.e. sets of differential equations that represent all the physical phenomena that significantly influence the energy balance of the considered building. As such, white-box models are not suitable to control applications because they do not rely on measured data and calibration processes are very computationally expensive. Therefore, only black-box and grey-box models have been considered in this work.

1.1.1. Black-box models

Black-box models learn the thermal behaviour of buildings by fitting measured data through different statistical models without making any prior assumptions regarding physical relationships. There exist several black-box linear models, such as Auto Regressive (AR), Auto Regressive with eXogenous inputs (ARX), Auto Regressive with Moving average (ARMA), Auto Regressive with Moving Average and eXogenous inputs (ARMAX) and Output Error (OE) models. Artificial Neural Networks (ANN) are parametric non-linear models, whereas k-Nearest Neighbors, Support Vector Machines (SVM), Decision Trees, and Random Forest are non-parametric non-linear models [2]. SVM demonstrated to outperform traditional back-propagation neural networks in hourly cooling load predictions using outdoor temperature and solar radiation as input variables [9]. Extreme gradient boosting (XGBoost), a ML technique based on Decision Trees, was found to produce highly accurate predictions of heating and cooling loads based on synthetic datasets, and outperformed ANN and degree-day regression [10]. Long Short-Term Memory (LSTM) neural networks were able to accurately predict the indoor air temperature of both single and multiple thermal zones of a commercial building in Canada during the winter season with both constant and variable air supply [11]. The relative performance compared to alternative neural network techniques was found to depend on the length of the prediction horizon and on the chosen error indicator. Luo et al [12] simultaneously predicted the heating, cooling, lighting loads and BIPV electrical power output of a building using three black-box models: ANN, SVM and LSTM. The comparison suggested that for multi-output predictions ANN is the most accurate and SVM the computationally fastest, with LSTM showing medium errors and the highest computation time. An online learning approach based on recurrent neural networks for load forecasting achieved higher accuracy than the traditional offline LSTM neural network for five households for all forecasting lengths

and a tuning module adapts ANN hyper-parameters to newly arriving patterns [13]. Schubnel et al [14] compared a linear model with non-linear regressors and a recurrent neural networks and found the linear model to be better with regard to sample efficiency, objective minimization and computation time. RNN was instead better with regards to modelling accuracy and respecting constraints. LSTM neural network was used within an MPC framework in a simulation environment to decide optimal set-point trajectories for energy consumption minimization [15]. In a recent survey [16] on ANN for building energy prediction, almost half of them appeared to use LSTM neural networks. The main open questions are their transferability to different buildings and their combination with suitable optimization techniques for their actual deployment in control applications. Although a vast literature exists about control-oriented black-box models, there are few implementations of such models in MPC-controlled buildings [17, 18].

1.1.2. Grey-box models

Grey-box models represent a trade-off between the mentioned modelling paradigms, as they fit measurements onto simplified models of the building heat dynamics, often expressed in the form of lumped capacitance models based on the electrical analogy. The first examples of such building models based on two [19, 20] or three capacitances [21] were already proposed in the 1980s. A four-state model was formulated using stochastic differential equations to describe the thermal response of two test rooms oriented on the north and south side of a building [22]. Two states for each room were sufficient to obtain an accuracy of 0.4°C. The very different influence of solar radiation in the south- and north-oriented test rooms led to two different radiator models. Bacher and Madsen [23] and Privara et al [24] presented different identification methods for building models to be used for MPC. The importance of building modelling for MPC applications was discussed in several works, e.g. [25, 26, 27]. Different toolboxes are available to perform system identification, i.e. to find a suitable model and to estimate its parameters based on the available data [28, 29]. A study on the dispersion of estimated parameters showed that the latter can be significantly reduced by reducing the degrees of freedom of the calibration process [30]. A recent study quantified the influence of many factors on MPC performance, and found that over-training and improper model structure are the most critical factors, and that the optimal length of training data for single thermal zone models is limited to a few days [31]. There are several simulation-based demonstrations of MPC in buildings using grey-box models, mostly for office and University buildings [32, 33, 35], labs and building demonstrators [34, 36].

1.1.3. Comparison of grey-box and black-box models

The main advantage of the black-box approach is that it requires lower development time -thus cost- compared to physics-based approaches. Neural networks and other kinds of non-linear data-driven approaches are preferred by the machine learning community. On the other hand, black-box models require more training data than grey-box models and are not reliable outside the training range [5, 2]. The main advantage of grey-box models and linear black-box models is that they allow to formulate convex optimization problems, thus guaranteeing an optimal solution of control problems. For this reason, they are often preferred by the control community. In recent years, researchers have tried to merge data-driven and physics-based modelling techniques

to overcome their relative weaknesses [40, 41, 42]. Although the mentioned results on hybrid modelling approaches are promising, little has been said so far on the relative performance between grey-box and black-box models. This fact can be explained by the different reference communities to which these modelling paradigms come from, namely control theory and machine learning. In fact, to the authors knowledge, only two papers that specifically address this topic have been published so far. The first one compared different grey-box models with ANN and ARX using data monitored in the air handling unit of a Canadian house [4]. Data-driven models and in particular ANN outperformed grey-box models. However, the comparison was limited to the HVAC system i.e. it did not include the building envelope and the indoor environment. A second study compared a white-box model implemented in EnergyPlus, three grey-box models based on RC thermal networks and two black-box models including NARX and ANN for indoor air temperature prediction of a University building in Denmark [43]. In this case, the black-box models were found to outperform grey- and white-box models in 7 tests out of 8. It was also found that grey-box models do not need long training periods and that their performance worsens over long time periods due to the error accumulation over time. Moreover, recent findings seem to suggest that, although non-linear black box models are more accurate compared to grey-box models and linear black-box models for indoor air temperature and heating/cooling loads prediction, such advantage does not necessarily reflect in better MPC performance in practice [14, 18].

1.2. Research gap and objective

The literature review shows that over the past two decades, building scientists dealing with MPC have mainly focused on improving RC models and system identification techniques in order to utilise grey-box models for HVAC systems'control. Meanwhile, the machine learning community has emerged, which has developed promising data-driven modelling techniques such as neural networks.

In this context, a research gap has emerged, as only a few articles specifically deal with the comparison between these modelling approaches under different conditions.

The present work provides an original contribution in this direction by comparing grey-box models and the most commonly studied black-box model (i.e. LSTM neural networks [16]), considering different building types, different lengths of the training datasets and using different data sources (simulations, measurements) for the prediction of indoor air temperature in both heating and cooling seasons.

The remainder of this manuscript is organized as follows: Section 2 describes the grey-box and black-box models; Section 3 describes the case-study building and the assumptions used under each test and scenario. Section 4 shows the comparison between the models and a critical discussion on the results. Finally, Section 5 summarizes the key findings of the research.

2. Models

This Section presents a brief description of the models considered for comparison, i.e. grey-box models and LSTM neural networks.

2.1. Grey-box models

The grey-box model considered here is a variant of the well-known ISO 13790 [44], a single-zone building model based on the electrical analogy, shown in Fig. 1. This model describes the transient thermal behavior of the building using one thermal capacitance and five thermal transmission coefficients. These coefficients describe the transmission through glazed and opaque building elements ($H_{tr,w}$ and $H_{tr,op}$ respectively) and the heat exchange due to ventilation and infiltration air change rates H_{ve}. The thermal transmittance of opaque building components, $H_{tr,op}$, is divided into two components $H_{tr,em}$ and $H_{tr,ms}$, while $H_{tr,is}$ represents the coupling conductance between indoor air and internal building surfaces. Indoor air temperature and internal surfaces' temperature nodes are θ_i and θ_s, respectively. In the model proposed by the Standard, the thermal inertia of all building components is lumped into a single thermal capacitance, C_m, whose equivalent temperature is the corresponding node θ_m. The other two nodes represent the temperatures of the outdoor air θ_e and the supply air θ_{su}. Concerning the solar heat gain ϕ_{sol}, the Standard split it into three components: the solar radiation transmitted through glazed surfaces, the solar radiation absorbed by opaque external surfaces, and the negative radiative heat flux exchanged with the surrounding environment. Both internal and solar heat gain (ϕ_{int} and ϕ_{sol}) are distributed to the three temperature nodes θ_i, θ_s, and θ_m according to semi-analytical correlations provided by the Standard.

Two main modifications have been performed to the model described so far, as shown in Fig. 1. First, two further thermal capacitances have been added to consider temperature variations at higher frequencies, i.e. the thermal capacitance of lightweight building components C_s (furniture, internal partitions, doors, etc.) and the thermal capacitance of indoor air volume C_i. Secondly, the solar heat gain here only considers the transmitted radiation through glazed surfaces $\phi_{sol,tr}$. The solar radiation absorbed by external opaque surfaces has been considered through an equivalent sol-air temperature $\theta_{e,eq}$, which is a common assumption for simplified building energy models [45].

$$\theta_t^{e,eq} = \theta_t^e + \frac{\alpha_{se}\, I_t^{sol,op}}{h_{conv,se}} \tag{1}$$

As a result, the model can be described by the following three equations:

$$H_{ve}(\theta_t^{su} - \theta_t^i) + H_{tr,is}(\theta_t^s - \theta_t^i) + \phi_t^{hg,i} + \phi_t^{hc,conv} = \frac{C_i}{\tau}(\theta_t^i - \theta_{t-\tau}^i) \tag{2a}$$

$$H_{tr,w}\left(\theta_t^{e,eq} - \theta_t^s\right) + H_{tr,is}\left(\theta_t^i - \theta_t^s\right) + H_{tr,ms}\left(\theta_t^m - \theta_t^s\right) + \phi_t^{hg,s} + \phi_t^{hc,rad} = \frac{C_s}{\tau}(\theta_t^s - \theta_{t-\tau}^s) \tag{2b}$$

$$H_{tr,em}\left(\theta_t^{e,eq} - \theta_t^m\right) + H_{tr,ms}\left(\theta_t^s - \theta_t^m\right) + \phi_t^{hg,m} = \frac{C_m}{\tau}(\theta_t^m - \theta_{t-\tau}^m) \tag{2c}$$

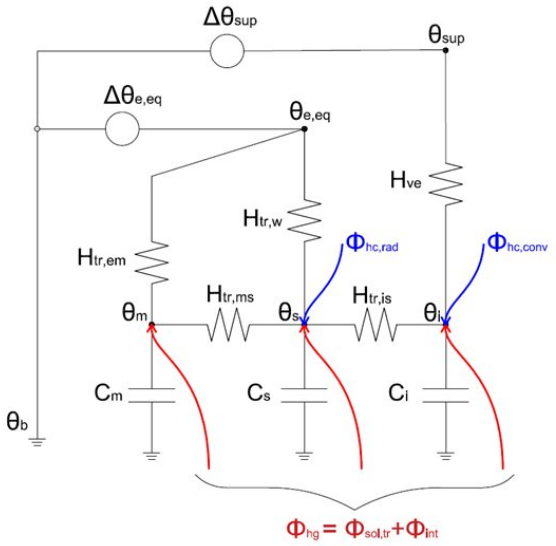


Figure 1. Lumped-capacitance building model.

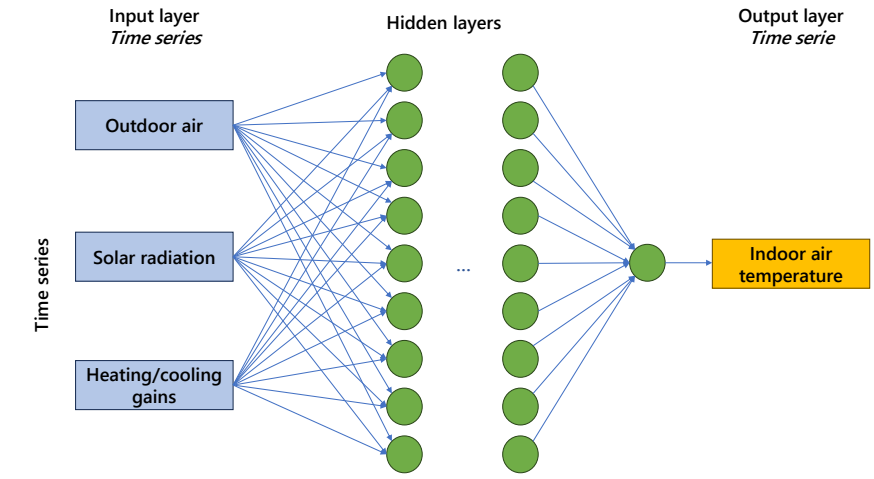
In the present work, three model variants have been considered: a first-order 1C model where $C_s = C_i = 0$ and $C_m \neq 0$, a second-order 2C model where $C_i = 0$ and $C_m \neq 0; C_i \neq 0$, and the third-order 3C model, where all capacitances are not zero. Three variants of this model are considered to investigate the effect of considering more low-frequencies capacitances in grey box models.

2.2. Long-short memory neural networks

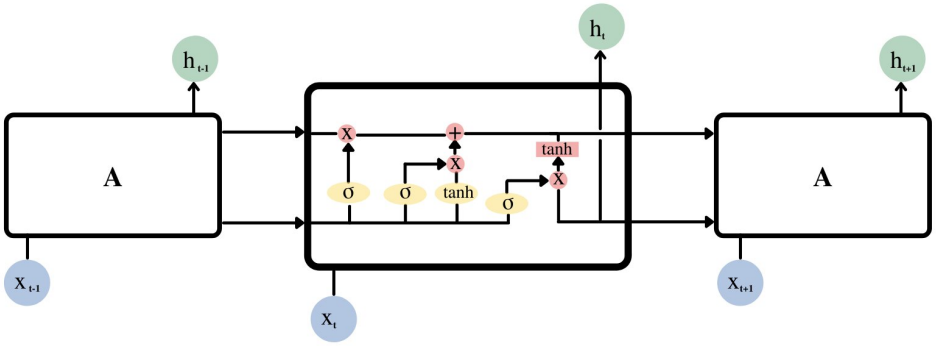
With respect to pure Artificial Neural Networks, Long-Short-Term-Memory (LSTM) neural networks represent the most used modeling technique used for buildings' energy loads prediction [16]. Indeed, the main feature of LSTMs is that they can exploit information for long periods of time, because they are capable of learning long-term dependencies and considering dynamic and time-dependent systems, such as buildings.

As depicted in Figure 2, LSTMs act like a chain of repeating modules of neural networks. The core idea behind the network is the cell state, which is used to transport the information. The latter is added or removed from the cell state with structures called gates. The workflow of the LSTM is the following:

(1) *Forget gate*: The first step in the LSTM is to select which information has to be deleted from the cell state. This step is performed by the first gate (left gate in Figure 2), which is called the "forget gate layer".
(2) *Input gate*: In the second step the LSTM selects which new information has to be stored in the cell state. This means that: first, a gate (called the "input gate layer") decides which values have to be updated, second, a *tanh* layer creates a vector of new candidate values. This vector can be added to the state.
(3) *Update*: Update the old cell state with the values computed in step 2.
(4) *Output gate*: LSTM composes the output based on the new cell state, filtered by another gate multiplying a tanh. This multiplication is performed by pushing the values between -1 and 1.



(a)



(b)

Figure 2. Structure of LSTM network (a) with focus on the repeating module (b).

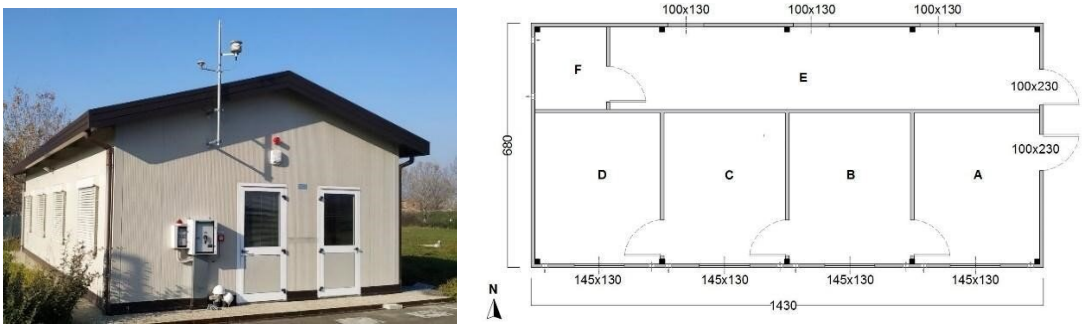
3. Methods

This Section is divided into five parts. The first describes case study building, whereas the second lists the assumptions for creating the synthetic datasets with EnergyPlus. The third and fourth subsections describe the parameter calibration process based on the synthetic and measured data, respectively. The fifth part describes the indicators used to carry out the prediction performance.

3.1. Case-study building

The case study building consists of a 100 m^2 single-storey prefabricated block with 6 rooms, as shown in Fig. 3b. Four similar rooms (A, B, C and D) are oriented to the south, with a total net heated area of 59 m^2. The northern part, i.e. rooms E and F, is separately controlled and used for other research purposes. The laboratory is meant to simulate the performance of an efficient "all-electric" building. To this end, the laboratory has been equipped with (i) an air-to-water heat pump air conditioning system connected to four fan coils with a 300-liter water heat storage tank; (ii) a heat pump water heater with 200-liter storage tank for domestic hot water (DHW) production; (iii) an air/air conditioning system consisting of an outdoor condensing unit and four fan coils with a total nominal cooling capacity of 6.8 kW; (iv) four air extractors; (v) a set of electrical appliances (e.g., washing machine, dryer, dishwasher and combined refrigerator).

The building materials are lightweight: walls are made of 10 *cm* prefabricated polyurethane panels, while the ceiling, placed at 2.70 *m* height, is made of 10 *cm* mineral glass fibre. The ground floor is made of concrete, with internal insulation of 5 *cm* of XPS, below 5 *cm* of screed coupled with a PVC floor layer. One double pane window is installed in each of the four rooms (A, B, C, D). Fig 3 provides a photo and a sketch of the case study building, while the first column in Table 1 provides further data on the building envelope. The building is located in Piacenza (Northern Italy), in a temperate humid climate with hot summers, according to Koppen-Geiger classification (Cfa).



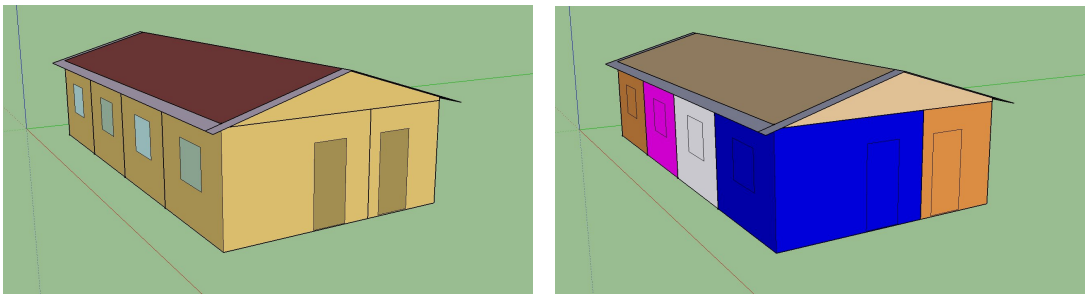
(a) Photo. (b) Floor plan.

Figure 3. Photo and floor plan of the case study building.

3.2. Generation of synthetic datasets with EnergyPlus

A detailed dynamic building model has been developed in EnergyPlus [7] to produce synthetic data to test the calibration process on different building structures. The geo-

metrical sketch of the model is shown in Figure 4. A thermal zone was created for each building room, including an additional thermal zone to model the space between the internal ceiling and the roof tile (see Fig. 4b). Only zones A, B, C, D are conditioned, while free-floating temperature is applied to the remaining ones.



(a) Sketch. (b) Thermal zoning.

Figure 4. EnergyPlus model of the lab.

All the simulations were carried out considering the weather file of Piacenza (Italy), where the building is located. The weather file was downloaded from the EnergyPlus database [46]. Different configurations of materials have been used in the analysis, combining different thermal insulation levels and construction types with different weight as per 1. Two lightweight buildings (LW) were simulated considering prefabricated panels with Polyurethane and XPS insulation. The first one, model B, is highly insulated and represents the actual structure of the Lab. The second one, model Bni, is similar but with a lower thermal insulation level. Concerning heavyweight construction sets (HW), three additional stratigraphies have been used, referring to typical residential Italian buildings of the 1970s (B70), 1990s (B90) and those of recent construction (BN), respectively [47]. The main materials are bricks and concrete, coupled with variable external insulation. Table 1 provides further details on the construction sets.

Table 1. U-value ($W/(m^2K)$) of main building components of the simulated structures.

Building structure	**Lightweight (LW)**		**Heavyweight (HW)**		
Materials	Polyurethane and XPS		Solid/hollow bricks or concrete, variable external insulation		
Building code	**B (real facility)**	**Bni**	**B70**	**B90**	**BN**
Representative for	>2005	-	1976-90	1991-2005	>2005
External wall U-value $W/(m^2K)$	0.28	1.72	1.18	0.69	0.28
Roof U-value $W/(m^2K)$	0.29	0.29	1.29	0.63	0.31
Ground floor U-value $W/(m^2K)$	0.39	0.39	0.75	0.62	0.29
Windows U-value $W/(m^2K)$	2.29	5.68	2.29	2.29	1.53

The internal heat gains have been set within the range of 5 W/m^2 (during the night) and 10 W/m^2 (morning and dinner time) in zone A, B, C, D, resulting in a total average value between 180 W and 370 W. This schedule considers both electric appliances and lights in the building. No vapor generation has been considered inside the building. Infiltration and natural ventilation airflow rate have been set to a constant value of 0.15 vol/h. These assumptions have been set in order to qualitatively represent a real operational condition of a residential building. The temperature setpoint of conditioned zones (A, B, C, D) has been set to 20 $°C$ from 07:00 to 23:00 during the winter season, with a night setback of 17 $°C$. The scheduling time is equal during

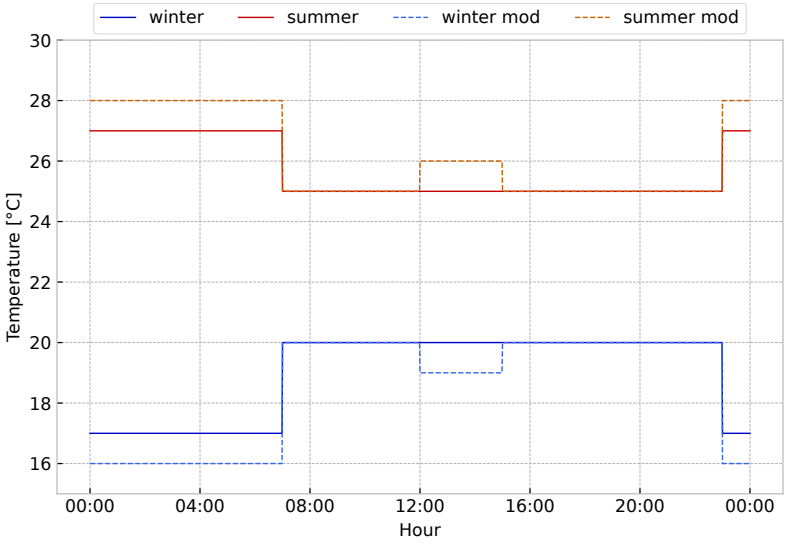


Figure 5. Setpoint schedules applied to the EnergyPlus models.

the cooling season, with a setpoint of 25 $°C$ and a setback of 27 $°C$. In addition, a cut-out temperature difference of 2 $°C$ has been added to the setpoint object in the model. This option provides a deadband to the heating/cooling setpoint, resulting in a behaviour of the HVAC system similar to a standard hysteresis controller. This option has been added to create more realistic datasets and to provide the necessary fluctuations in the disturbances/states needed for model calibration. A second setpoint schedule, called *Mod*, has been used to further test the models in different and more variable conditions. The *Mod* setpoint strategy consists in including an additional setback period in between 12:00 and 14:00 (19 $°C$ and 26 $°C$, for winter and summer period, respectively). Figure 5 shows graphically both setpoints. The time-step of the EnergyPlus simulation has been set to 15 minutes.

After the setup of the building models, the python library *eppy* [48] was used to automatically run EnergyPlus simulations from the corresponding *.idf* files. The profiles of the simulations were later used to train the models. In particular, the following variables, listed in Table 2, were considered to build the synthetic datasets: dry-bulb (external) air temperature and relative humidity, wind speed and direction, diffuse and total solar radiation on the horizontal plane, average internal temperature and relative humidity of the conditioned thermal zones, heating and cooling demand, total internal heat gains (sum between those due to electrical appliances and those linked to occupancy i.e. to the presence of people, as described above).

The datasets were finally prepared by selecting different training and testing periods, as shown in Table 3. As grey-box models and LSTM neural networks perform differently depending on the amount of training data, a preliminary analysis was carried out to set suitable training periods: 14, 24 and 34 days have been chosen for the NN, while 3, 5 and 7 days have been set for the grey-box models. While training days are different, the testing periods (1 or 2 days after the training periods) are the same for both model types.

Table 2. List of variables considered for the training phase.

Time-series	
Outdoor air	Outdoor air drybulb temperature
	Outdoor air relative humidity
	Wind speed
	Wind direction
Solar gains	Horizontal diffuse solar irradiance
	Direct normal solar irradiance
Internal gains	Total internal heat gain
HVAC loads	Heating/cooling load
Zone	Average zone air temperature
	Average zone air humidity

Table 3. Training and testing periods.

Model	Season	Training			Testing	
Black Box		34 days	24 days	14 days	1 day	2 days
	Winter	20 Jan - 23 Feb	30 Jan - 23 Feb	9 - 23 Feb	24 Feb	24 - 25 Feb
	Summer	3 Jul - 6 Aug	13 Jul - 6 Aug	23 Jul - 6 Aug	7 Aug	7 - 8 Aug
Grey Box		7 days	5 days	3 days	1 day	2 days
	Winter	16 Feb - 23 Feb	18 - 23 Feb	20 - 23 Feb	24 Feb	24 - 25 Feb
	Summer	30 Jul - 6 Aug	1 Aug - 6 Aug	3 Aug - 6 Aug	7 Aug	7 - 8 Aug

3.3. Model calibration using simulation output

3.3.1. LSTM neural networks

The LSTM NN was created and trained using *PyTorch* python library [49]. The calibration process consists of three steps: generating the neural network, optimizing its hyperparameters, and training it over the data to make it learn the system behavior. The model was generated by coupling an LSTM layer to a fully connected layer and a sigmoid function. Input data were normalized within the range [0,1] and extended to a 3D array. Indeed, the 2D input dataset was reshaped to consider a 3-hour data history as 3rd dimension, as required by recurrent neural networks. The second step involves hyperparameters analysis. Before training, hyperparameters were optimized to avoid underfitting or overfitting issues. The considered hyperparameters are learning rate, hidden layer size, and optimization algorithm (see Table 4). The goal was to find an optimal combination of them that minimizes a predefined loss function. In order to obtain the optimal results, *sklearn.model_selection.GridSearchCV* was used. The third and last step is model training. In this perspective, a proper number of training iterations (epochs) was selected for each case to avoid overfitting or underfitting issues. To this purpose, the training stopped once the MSE improvement over two consecutive iterations was lower than a predefined threshold value.

Table 4. Optimization hyper-parameters.

Hyper-parameter	Values
Learning rate	[0.01, 0.001, 0.0001]
LSTM hidden layer size	[16, 32, 64, 128]
Optimization algorithm	Adam, RMSprop [50]

3.3.2. Grey-box models

The calibrated parameters in the grey-box models are not limited to the five thermal resistances and three thermal capacitances presented in Section 2.1. In fact, the calibration process determines six other parameters that are needed to scale and distribute the heat load from the HVAC system, as well as internal and solar heat gains to the temperature nodes. These additional parameters are the convective fraction of the heat emission system k_{conv}, the multipliers for the incident solar radiation on glazed and opaque surfaces $k_{s,gla}$ and $k_{s,opa}$, the average internal heat loads $\phi_{int,0}$ and two coefficients, k_s and k_a, that distribute the free heat gains to the three temperature nodes according to the following Equations (adapted from Standard's formulation):

$$\phi_t^{hg,i} = 0.5\ \phi_t^{int} + k_s \phi_t^{sol} \tag{3a}$$

$$\phi_t^{hg,s} = (1 - k_a) \left[0.5\ \phi_t^{int} + (1 - k_s)\, \phi_t^{sol} \right] \tag{3b}$$

$$\phi_t^{hg,m} = k_a\ \left[0.5\ \phi_t^{int} + (1 - k_s)\, \phi_t^{sol} \right] \tag{3c}$$

Therefore, the calibration aims to estimate the value of 12-14 parameters, depending on the number of capacitances of the chosen model. The parameters are initialized using basic information on the geometrical and physical characteristics of the building considered (e.g. floor area, area of windows, thermal transmittance of windows and walls, etc). The calibration algorithm is based on the Trust Region Reflective algorithm contained in the Python library *scipy.optimize.least_squares* [51]. This library is based on algorithms suited to solve constrained optimization problems in which the objective function is quadratic. The parameter calibration stops when at least one of the following exit criteria is met: (i) when the difference in the objective function between two consecutive iterations is lower than 0.0005%, (ii) when the maximum number of iterations (set to 500) is reached. The disturbance ϕ_{sol} was calculated by processing global irradiation data on the horizontal plane, i.e. calculating the sum of the incident solar radiation on oriented external walls using the well-known *pvlib* library [52].

3.4. Model calibration using monitored data

The comparison between grey-box and black-box models was subsequently repeated using data collected from the case study building. The latter is equipped with several sensors that can log internal and external environmental conditions, HVAC systems operating conditions and the electric appliances consumption. For the purpose of this work, the following variables were used:

- *Weather conditions*: outdoor temperature, outdoor relative humidity, global horizontal irradiance, diffuse irradiance;
- *Indoor conditions*: average indoor temperature and humidity;
- *Internal heat gain*: electric consumption of the appliances;
- *Heating and cooling load*: fan-coils heating and cooling power.

The logs, recorded with a sample time of 1 minute, have been averaged over 15-minute time-steps. The testing periods are listed in Table 5. Training periods are 14 and 24 days for LSTM neural networks, 3 and 5 days for grey-box models (See Section 3.2).

Table 5. Training and testing periods with measured data.

Model	Season	Testing	
Black-box		1 day	2 days
	Winter	2 Dec	1 - 2 Dec
	Summer	29 Aug	28 - 29 Aug
Grey-box		1 day	2 days
	Winter	2 Dec	1 - 2 Dec
	Summer	29 Aug	28 - 29 Aug

3.5. Indicators for model performance assessment

The performance of the models was evaluated with two metrics, namely Root Mean Square Error (RMSE) and R-Square (R^2). These metrics were used to evaluate the calculated indoor air temperature both in the training and testing phases. The first indicator was chosen to measure the average error of the models:

$$RMSE = \left(\frac{\sum_{t=1}^{T} \left(\theta_t^i - \theta_t^{i,meas} \right)^2}{T} \right)^{0.5} \qquad [^\circ C] \tag{4}$$

where T is the size of the sample (number of time-steps) of the period considered. The second indicator assesses to what extent the variance in the output variables is explained by the input variance. Graphically, this metric assesses the similarity between calculated and measured indoor air temperature profiles. It is defined as follows:

$$R^2 = \frac{\sum_{t=1}^{T} \left(\theta_t^i - \theta_t^{i,meas} \right)^2}{\sum_{t=1}^{T} \left(\theta_t^i - \overline{\theta}^{i,meas} \right)^2} \qquad [-] \tag{5}$$

where $\overline{\theta}^{i,meas}$ is the average measured indoor air temperature over the entire period considered. A fair accuracy evaluation of energy models like those considered in this work relies on both indicators, as highlighted by Chakraborty and Elzarka [53].

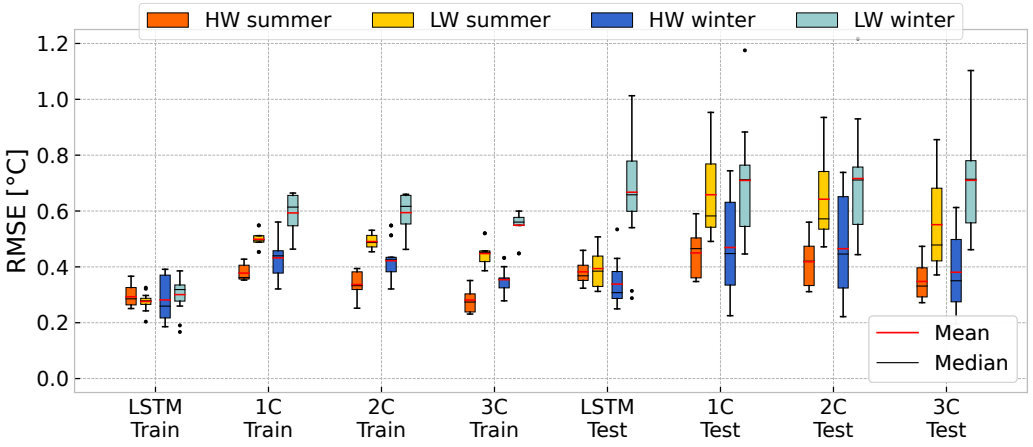
4. Results

The first section compares black-box and grey-box models based on the synthetic datasets generated with EnergyPlus. The second section relies on the monitored data from the case study building.

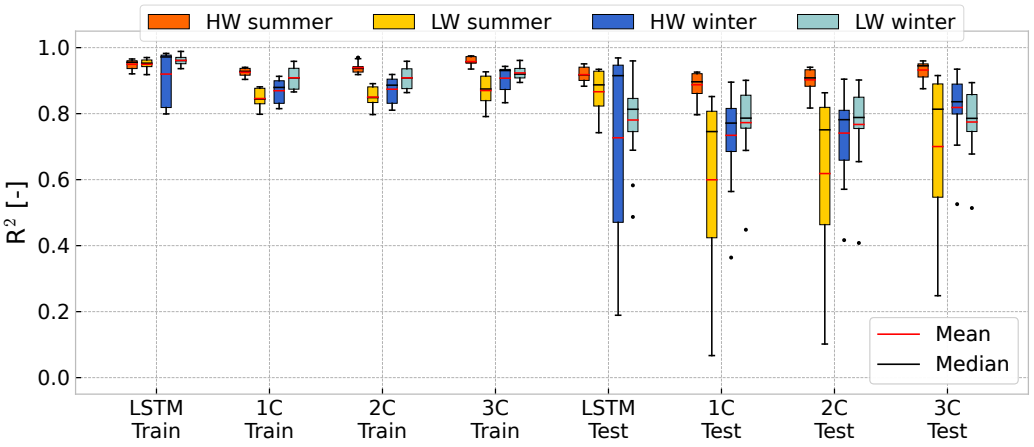
4.1. Model calibration using simulation output

Fig. 6 allows comparing the prediction performance of neural networks and grey-box models. The boxplot groups KPIs by model, building structure (lightweight versus heavyweight constructions), and season (winter versus summer). Left and right side of the plots compare the training and testing phases of all models. Both $RMSE$ and R^2 result in worse performances when testing is considered. In particular, the $RMSE$ in the case of the LSTM neural network during the training ranges between 0.2 and 0.4 °C. Grey-box models show a higher dispersion of the $RMSE$, reaching up to 0.7 °C (lightweight buildings in the heating season, light-blue column). As far as the R^2 is concerned, the neural network is always above 0.90, apart from heavy-weight buildings during the heating season (dark blue). Grey-box models perform similarly, resulting in R^2 values that are always higher than 0.8 during the training, even though a slightly worse performance can be noticed for lightweight buildings during the cooling season (yellow). These trends are amplified in the testing dataset. In fact, R^2 has a remarkably lower average and a higher dispersion for heavyweight buildings in the winter season for the LSTM (dark blue) and lightweight buildings in the summer season for grey-boxes (yellow). This behavior is not as evident when looking at $RMSE$ indicator. Indeed in testing, the LSTM obtain the worst performance in case of a lightweight building in the winter season (light blue), while for grey-box models, the only evidence is that the performance on heavyweight structures is better than that on lightweight structures in both seasons. The three thermal capacitances model, 3C, seems to outperform grey-box models of lower order, although such improvement does not always occur – as in the case of lightweight building structures in the heating season (light blue).

Two general considerations emerged from the comments above. First, model performance depends on the building type. Second, the chosen performance indicators do not provide the same information. For these reasons, Fig. 7 shows the $RMSE - R^2$ scatter plot considering different insulation levels and materials configurations. This plot is presented as the RMSE measures the prediction error, while R^2 provides an indication of the extent to which the variance in the predicted variable (average indoor air temperature) can be explained by the model inputs signals – in this case signals such as the outdoor dry-bulb air temperature, the global horizontal irradiance, the internal heat gains, etc. In other words, low R^2 values on the testing sets might reflect an overestimated impact of some input data rather than indicating a poor model performance. This is the case of LSTM in highly insulated heavyweight structures, represented with dark blue points in Fig. 7a. Although having quite good $RMSE$ values, lower than 0.5 °C, these models have very low R^2, below 0.6. A similar consideration holds true for 3C grey-box model in lightweight, uninsulated structures, represented with yellow circles in Fig. 7b. In this case, low R^2 values could be linked to the simplified modeling of solar radiation, which hinders the performance of lumped-capacitance models. Indeed, a wrong estimation of solar heat gains turns, in some cases, into a bad overall model performance that is reflected by the high $RMSE$ values, above 0.5 °C. Nevertheless, it is worth noticing that this type of structure, i.e, lightweight buildings with poor thermal insulation, are quite unusual constructions, rarely encountered in practical application. Looking again at the LSTM results in Fig. 7a, the best results are the points located in the upper left corner, with high R^2 and low $RMSE$. In particular, the building with the best performance is the (old) low-insulated (orange/yellow points), probably due to a clear repetitive day-night indoor air temperature pattern. For (new) insulated constructions the LSTM performance decreases; the worst results are those for lightweight constructions (light blue), for which the $RMSE$ reaches 0.9-1



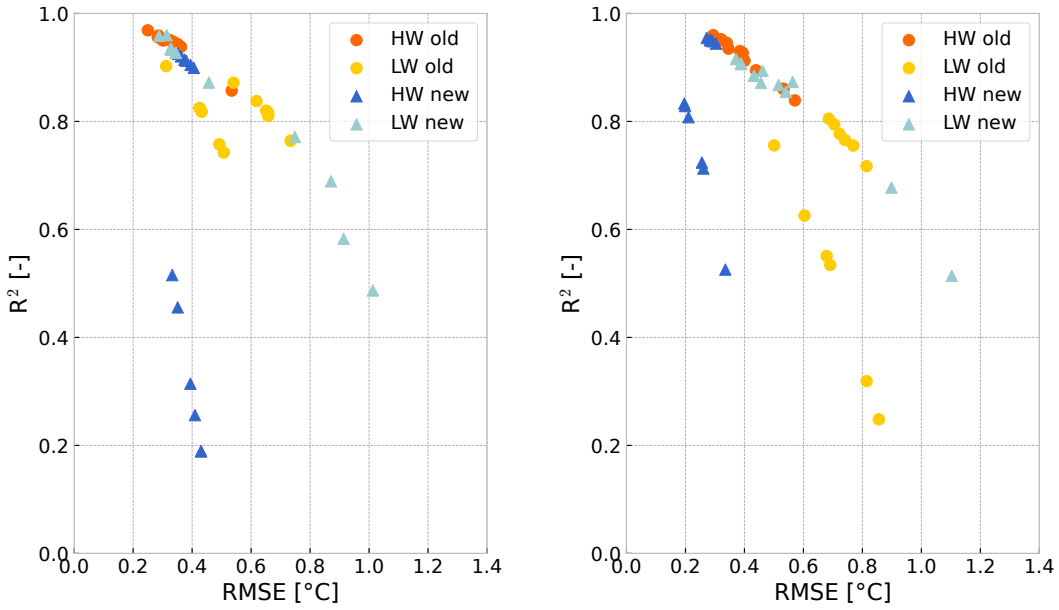
(a) $RMSE$.



(b) R^2.

Figure 6. Effect of model order, season and building structure.

°C in winter simulations (R^2 higher than 0.5). The best results of the 3C model are instead those corresponding to heavy structures, as shown by the previous Figure.



(a) LSTM black-box.

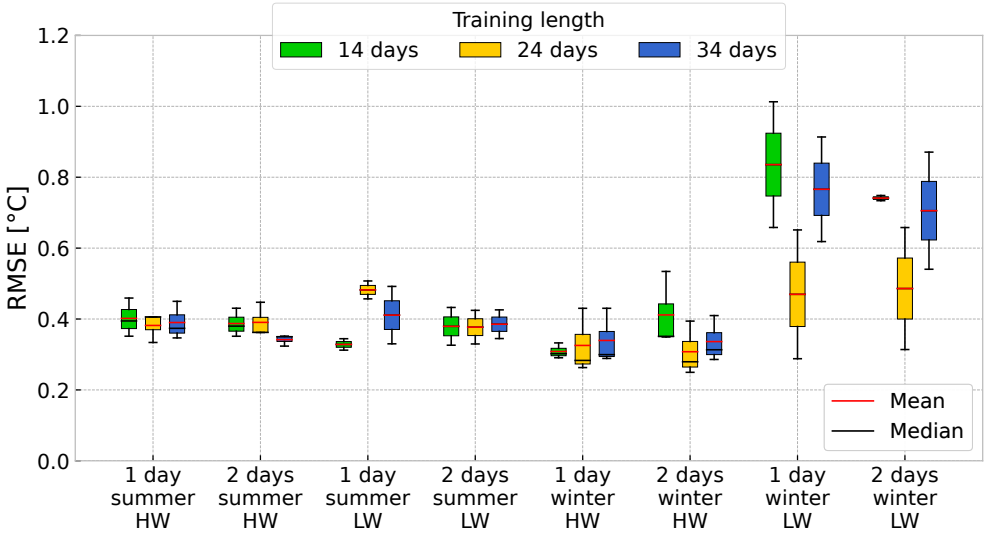
(b) 3C grey-box.

Figure 7. Testing $RMSE$ and R^2 scatter plot with insulation and building construction effect.

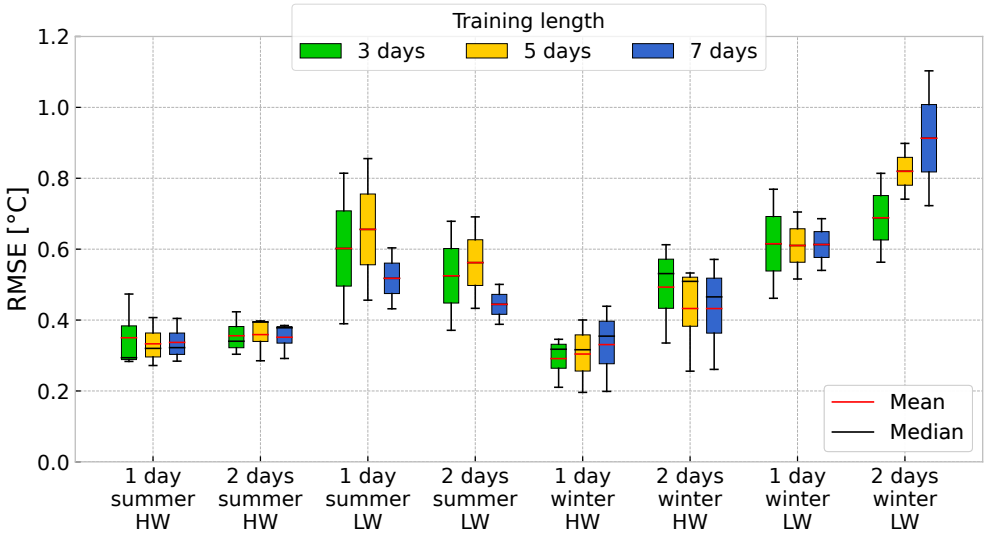
An additional analysis involves the length of training and testing periods. Fig. 8 compares the $RMSE$ between different training and testing days, for the LSTM, Fig. 8a and the 3C model, Fig. 8b . Concerning the LSTM, the plot shows that different training lengths don't affect the prediction performance significantly, except from the light constructions during the winter season (right side of the plot). In this case, the prediction performance differs significantly among training periods. In fact, a training period of 24 days (yellow column) leads to an average $RMSE$ of 0.49 °C, compared to 0.75 °C and 0.70 °C obtained for 14 and 34 days (green and blue column). This result represents an additional proof that Lightweight building constructions in the heating season are critical for the LSTM model, as already clear from previous Fig. 6.

A similar, yet less pronounced behavior emerges for the grey-box model, Fig. 8b. In this case, the influence of the training length is mostly evident for lightweight buildings (LW). Considering the cooling season, a training period of 7 days (blue column) makes it possible to reduce the $RMSE$ values compared to 3 and 5 days (green and yellow). During the winter instead, the behavior is opposite, as the $RMSE$ increases with longer training periods. It is worth saying that this happens only with a two-day testing, i.e. in the most unfavorable conditions. This might suggest that for both black-box and grey-box models, a low prediction performance could be partially compensated by an accurate choice of the training dataset. Moreover, when the description of the building physics is accurate enough, such as in all tests based on heavyweight building structures, grey-box models need a significantly lower amount of data than neural networks.

The last test, presented in Fig. 9, shows the model results considering the modified temperature setpoint schedule, already described in Section 3.2. The left side shows results without setpoint change between training and testing, while the right side



(a) LSTM black-box.



(b) 3C grey-box.

Figure 8. Effect of training and testing data size on the $RMSE$.

shows the performance in case the setpoint is changed during the test period. This test was performed to evaluate the ability of the models to adapt to conditions that may vary daily, as it happens in practice due to changed user needs or due to a different control policy pursued by the Building Energy Management System. Both performance indicators ($RMSE$ and R^2) worsen when the setpoint schedule is changed after the training period (right side). However, results highlight that the black-box model is much more sensitive than the grey-box models in this regard. For instance, in the case of heavyweight buildings in the heating season (dark blue), the average $RMSE$ increases from 0.34 °C to 1.34 °C. For the other cases, the average $RMSE$ increases to 0.9-1.1 °C. On the contrary, grey-box models provide good performances even with the modified setpoint. In the same scenario, for the 3C lumped-capacitance model, the $RMSE$ increases from 0.38 °C to 0.48 °C, and similar behavior occurs for 1C and 2C models. Unexpectedly, in the most critical scenarios (lightweight structures in the cooling season, yellow columns), the grey-box models significantly improve the R^2 performance when compared to tests with the same setpoint schedule, as it can be clearly seen by the narrower dispersion comparing the yellow boxplots in Fig. 9b. At the same time, the changed user setpoint during the testing period affects the R^2 performance on the heavyweight structure in the heating season (dark blue). However, the same deterioration was not recorded as far as the $RMSE$ was concerned.

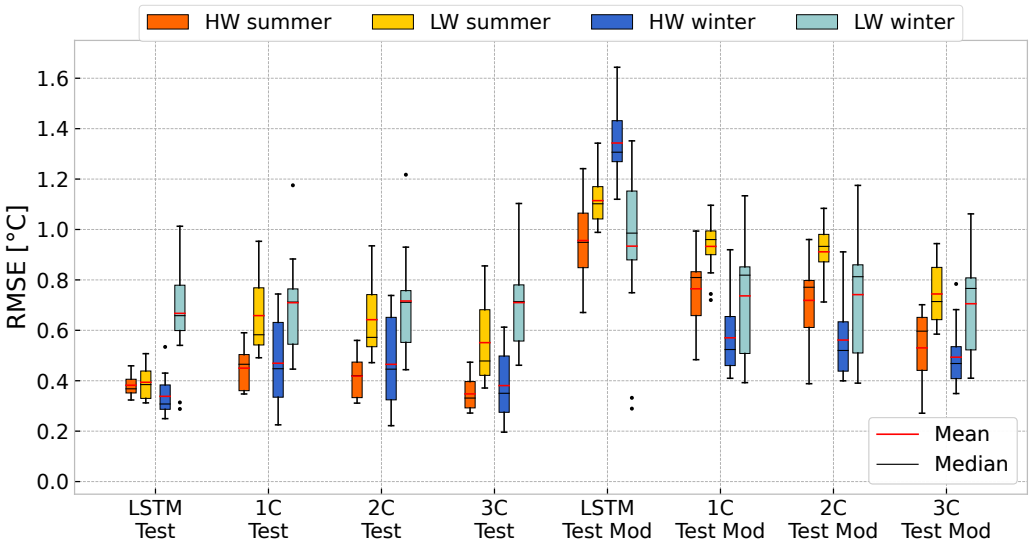
Fig. 10 shows the predicted indoor air temperature profiles for building B90 (slightly insulated heavy structure) during the heating period. The grey-box model (blue line) can reproduce the indoor air temperature fluctuations due to the on-off of the fancoils, simulated in EnergyPlus through a thermostat control. The LSTM neural network (red line) does not reproduce this high-frequency thermal response but it is more precise when a temperature drop occurs in the evenings - as the red line is almost coincident with the black dashed line during these hours. On the other hand, Fig. 10b shows that a change in setpoint during the testing period affects the accuracy of the LSTM model (red line) much more than it does on the grey-box model (blue line).

4.2. Model calibration with monitored data

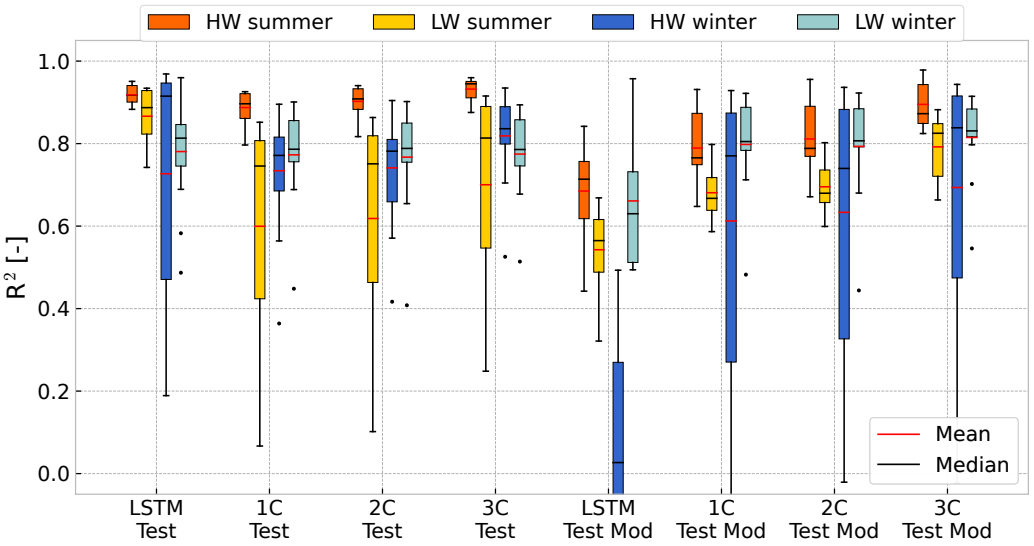
This section shows the results of similar tests performed with the same models on measured datasets from the case study building - see Sections 3.1 and 3.4. Fig. 11 shows that 1C and 2C grey-box models have the highest $RMSE$ and lowest R^2 scores, while the 3C model has better performance concerning both indicators. The average $RMSE$ obtained with the 3C model is 0.38 °C in training and 0.50 °C in testing, while the average R^2 during testing is 0.73. The LSTM neural network provides equivalent results: the average $RMSE$ scores are 0.42 °C and 0.51 °C in training and testing periods, respectively. The average R^2 in testing is 0.72.

All testing results are displayed in the scatter plots of Fig. 12, where they are split between cooling (left) and heating season (right). LSTM, orange stars, achieves $RMSE$ values between 0.3 °C and 0.7 °C in both conditions, while the R^2 is always higher than 0.50. On the contrary, 3C results show good performances for the summer season, but worse results in the winter season, where the minimum $RMSE$ reaches 0.45 °C.

To further analyze the models, the indoor air temperature profiles predicted by both LSTM and 3C models are compared to the measured one in Fig. 13. During the winter season, measured data show a fluctuating indoor temperature, that is properly followed both by LSTM and the grey-box thermal network. The 3C model seems to

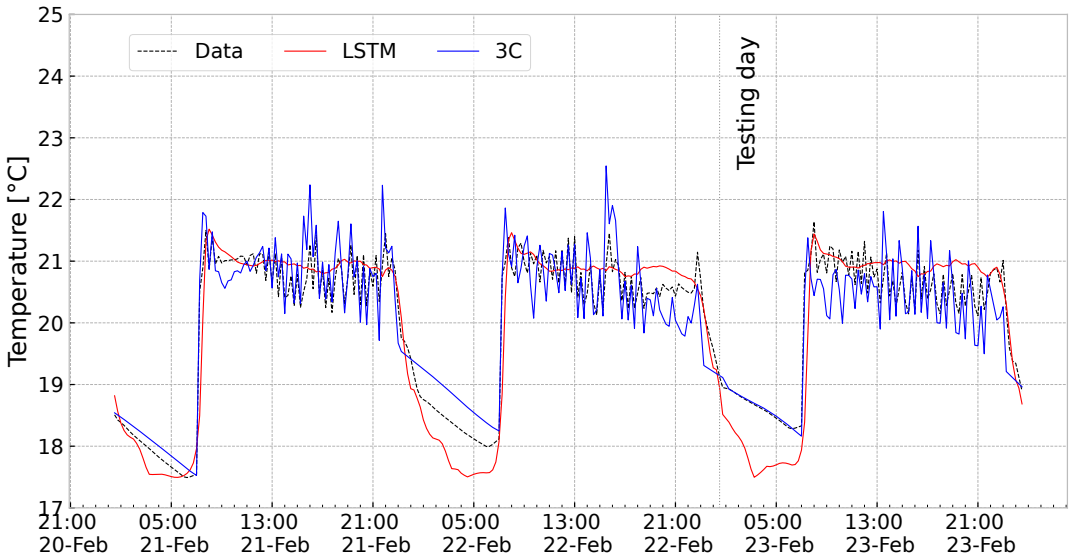


(a) $RMSE$.

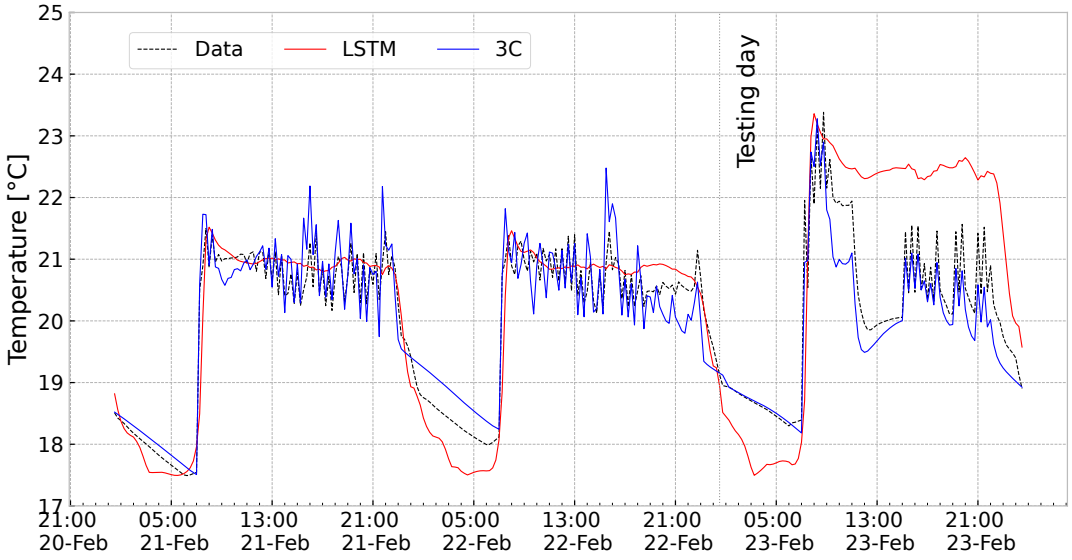


(b) R^2.

Figure 9. Effect of a modified setpoint strategy on the testing results.

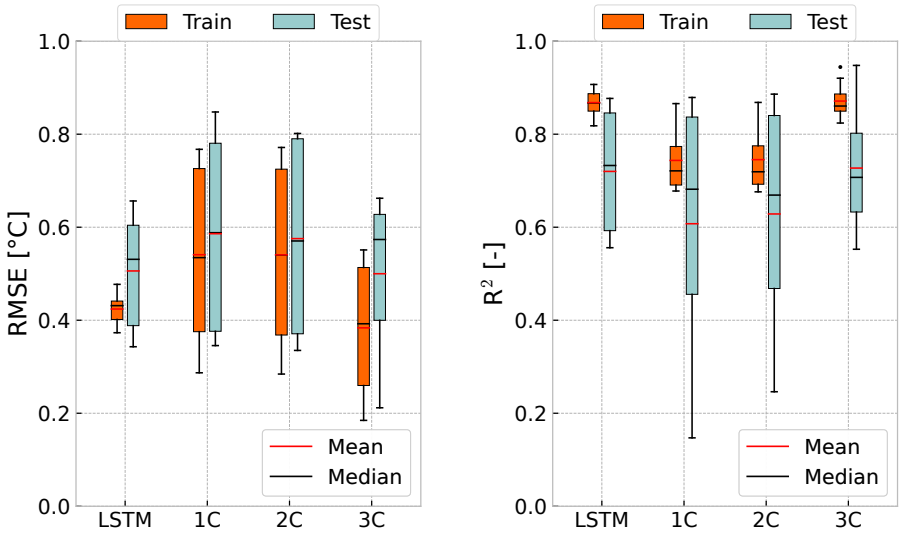


(a) Standard setpoint.



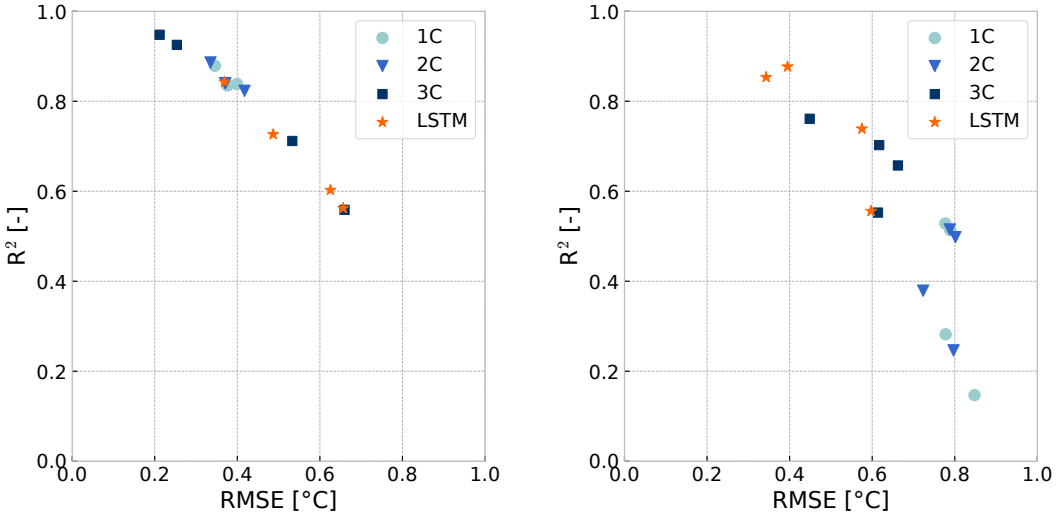
(b) Modified setpoint.

Figure 10. Temperature prediction for the LSTM neural network and grey-box 3C model: building B90 in the heating season.



(a) *RMSE*. (b) R^2.

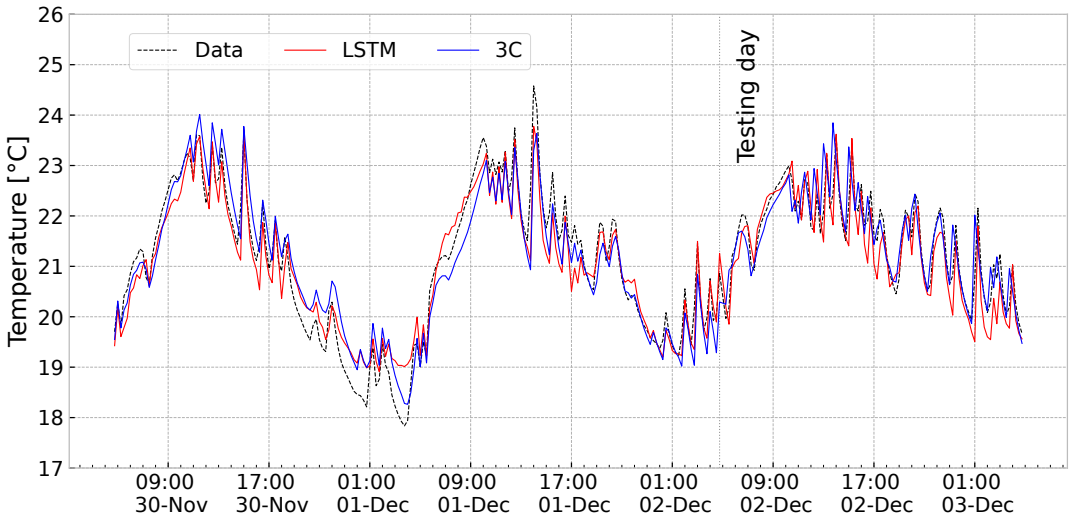
Figure 11. Monitored data results: average KPIs among different training and testing periods.



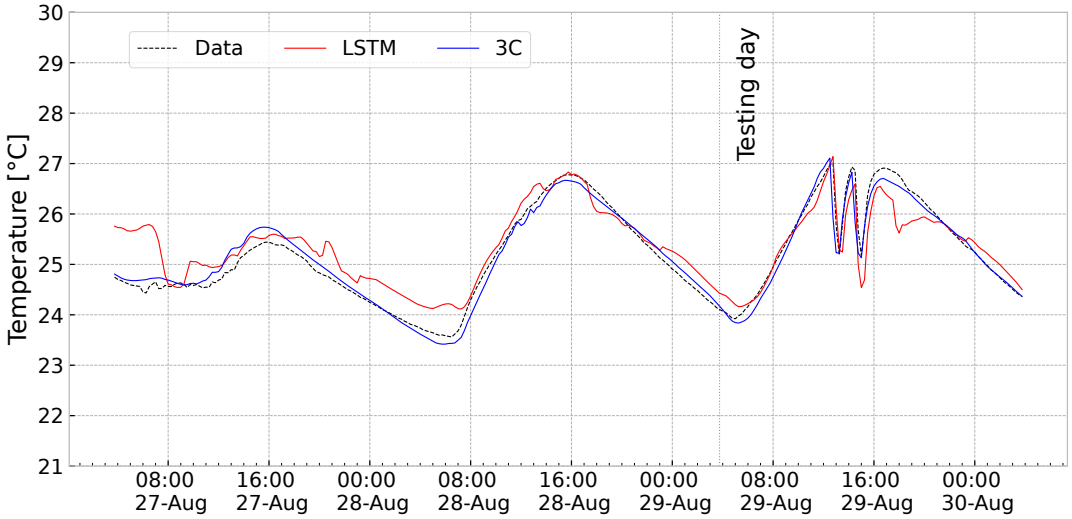
(a) Summer. (b) Winter.

Figure 12. Monitored data results, season comparison.

better reproduce the transient thermal behavior of the building in some circumstances, such as during the night setbacks and also in the late morning. In both cases, the LSTM slightly overestimates the average indoor air temperature. It is worth noting that the KPIs obtained in the previous Section on the synthetic datasets assumed perfect knowledge of the internal heat gains, while this is not the case in the tests based on monitored data.



(a) Heating season.



(b) Cooling season.

Figure 13. Temperature prediction for the LSTM neural network and grey-box 3C model: measured data.

4.3. Discussion

The objective of this research work was to fairly compare the performance of grey-box models and black-box models in predicting the thermal response of buildings.

To this end, both synthetic data obtained from EnergyPlus simulations and measured data from a building were used to support the study with a wide spectrum of scenarios in terms of building structures (heavyweight and lightweight with different thermal insulation levels) for both space heating and space cooling operation. The analysis considered training and testing periods of variable length. LSTM neural networks were chosen as one of the most representative black-box models used in this field, whereas lumped-capacitance models up to the third order were built to represent grey-box models. The first result is that the accuracy of indoor temperature prediction is sensitive to the building type for both black-box and grey-box models. For grey-box models, predictions are clearly better on heavyweight structures than on lightweight ones, which might be partially due to the oversimplified solar radiation modeling. In most cases, third-order models outperform models of lower order and behave, on average, similarly to LSTM neural networks. Despite a similar behavior on average, LSTM neural networks show a lower variability of the prediction accuracy. This holds true as long as the boundary conditions remain unaltered between the training and testing phases. In tests with a changed indoor air temperature setpoint schedule, grey-box models maintain an equivalent performance, while LSTM models deviate significantly from the simulated thermal behavior. The data points used by the grey-box models with the best prediction performance (3-5 days) are significantly less than those of the LSTM, which needs 14-24 days of training data. However, the size of the training dataset seems to impact the results significantly only in the worst scenarios, i.e. for lightweight buildings in the heating season for LSTM models and in the cooling season for grey-box models. The analysis does not consider the computational effort of the two approaches. However, it is worth mentioning that even though the training time is comparable on similar machines, NN requires an initial hyper-parameters optimization, which increases the amount of data needed and the computational effort compared to grey-box models. The influence of the season on the model performance is not confirmed when the tests are repeated using measured data from the lab. This could be due to the different ways in which internal heat gains were modeled in the two cases: known a-priori in the predictions based on synthetic EnergyPlus data, estimated from the calibration as an average value in the predictions with measured data. The other trends are instead confirmed: the 3C model seems to perform like the LSTM on average, yet with a higher variance of prediction accuracy. Finally, it is research has been trying to hybridise black-box models through physics-inspired constraints. A comprehensive comparison would also help to assess the performance of emerging modelling approaches.

5. Conclusion

The present study has compared grey-box models and neural networks for reproducing the thermal behavior of buildings, which is particularly interesting for Model Predictive Control applications in the building sector. The study was based on both simulated and measured data. Several analyses were proposed to test the models under different conditions, including seasons, building envelope types, and training and testing days.

Results show that the 3C lumped grey-box model performs slightly worse than the LSTM. The average $RMSE$ in both models is between 0.3 and 0.75 °C, with 1 °C peaks for particularly unfavorable cases. Models' performance is significantly influenced by both season and building structure. Grey-box models behave better on heavyweight structures than on lightweight ones, whereas LSTM seems to behave better in buildings

with lower thermal insulation. Regarding the training data size, 14-24 days seem to be sufficient for LSTM and, in most cases, 3-5 days for the grey-box model with three states, although in some tests increasing the training data to 7 days helps improve their predictions. In most cases, grey-box models with first or second-order do not perform comparably, therefore it seems always reasonable to use a third-order model whenever a grey-box approach is chosen. Although LSTM models need more data points for training, their performance shows lower variability than grey-box models as long as the boundary conditions over the prediction horizon do not change compared to those of the training period. In the latter case, grey-box model performance is less affected thanks to the underlying physical equations, whereas LSTM predictions degrade significantly concerning both R^2 and $RMSE$. In summary, LSTM has a better performance on average, however, grey-box models can be a better solution when:

- setpoint schedule or other control parameters change frequently, which is the case when the user interacts with the system -such as in residential buildings;
- lower computational resources are available (such as when computation is on local hardware) because in grey-box models hyperparameter optimization is not necessary, models are linear, and training data is limited.

Future research will try to reduce the variability of grey-box predictions by the number of calibration parameters and/or by using alternative system identification methods. Besides predictability, testing these models in a real control environment will be crucial for future analyses.

Acknowledgments

Omitted for double-blind review.

References

[1] Moncef Krarti. *Energy Audit of Building Systems: An Engineering Approach*. CRC Press, 3 edition, December 2020.

[2] Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, Juraj Oravec, Michael Wetter, Draguna L. Vrabie, and Lieve Helsen. All you need to know about model predictive control for buildings. *Annual Reviews in Control*, 50:190–232, 2020.

[3] Moncef Krarti. Control Strategies for Building Energy Systems. In *Optimal Design and Retrofit of Energy Efficient Buildings, Communities, and Urban Centers*, pages 117–187. Elsevier, 2018.

[4] Abdul Afram and Farrokh Janabi-Sharifi. Black-box modeling of residential HVAC system and comparison of gray-box and black-box modeling methods. *Energy and Buildings*, 94:121–149, May 2015.

[5] Gianluca Serale, Massimo Fiorentini, Alfonso Capozzoli, Daniele Bernardini, and Alberto Bemporad. Model Predictive Control (MPC) for Enhancing Building and HVAC System Energy Efficiency: Problem Formulation, Applications and Opportunities. *Energies*, 11(3):631, March 2018.

[6] Karl Mason and Santiago Grijalva. A review of reinforcement learning for autonomous building energy management. *Computers & Electrical Engineering*, 78:300–312, September 2019.

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

[8] Klein, S.A. et al. TRNSYS 17: A Transient System Simulation Program, 2017.

[9] Qiong Li, Qinglin Meng, Jiejin Cai, Hiroshi Yoshino, and Akashi Mochida. Applying support vector machine to predict hourly cooling load in the building. *Applied Energy*, 86(10):2249–2256, October 2009.

[10] Debaditya Chakraborty and Hazem Elzarka. Advanced machine learning techniques for building performance simulation: a comparative analysis. *Journal of Building Performance Simulation*, 12(2):193–207, March 2019.

[11] Fatma Mtibaa, Kim-Khoa Nguyen, Muhammad Azam, Anastasios Papachristou, Jean-Simon Venne, and Mohamed Cheriet. LSTM-based indoor air temperature prediction framework for HVAC systems in smart buildings. *Neural Computing and Applications*, 32(23):17569–17585, December 2020.

[12] X.J. Luo, Lukumon O. Oyedele, Anuoluwapo O. Ajayi, and Olugbenga O. Akinade. Comparative study of machine learning-based multi-objective prediction framework for multiple building energy loads. *Sustainable Cities and Society*, 61:102283, October 2020.

[13] Mohammad Navid Fekri, Harsh Patel, Katarina Grolinger, and Vinay Sharma. Deep learning for load forecasting with smart meter data: Online Adaptive Recurrent Neural Network. *Applied Energy*, 282:116177, January 2021.

[14] Baptiste Schubnel, Rafael E. Carrillo, Paolo Taddeo, Lluc Canal Casals, Jaume Salom, Yves Stauffer, and Pierre-Jean Alet. State-space models for building control: how deep should you go? *Journal of Building Performance Simulation*, 13(6):707–719, November 2020.

[15] Byung-Ki Jeon and Eui-Jong Kim. LSTM-Based Model Predictive Control for Optimal Temperature Set-Point Planning. *Sustainability*, 13(2):894, January 2021.

[16] Chujie Lu, Sihui Li, and Zhengjun Lu. Building energy prediction using artificial neural networks: A literature survey. *Energy and Buildings*, 262:111718, 2022.

[17] Michael Dahl Knudsen, Laurent Georges, Kristian Stenerud Skeie, and Steffen Petersen. Experimental test of a black-box economic model predictive control for residential space heating. *Applied Energy*, 298:117227, September 2021.

[18] Felix Bünning, Benjamin Huber, Adrian Schalbetter, Ahmed Aboudonia, Mathias Hudoba de Badyn, Philipp Heer, Roy S. Smith, and John Lygeros. Physics-informed linear regression is competitive with two Machine Learning methods in residential building MPC. *Applied Energy*, 310:118491, March 2022.

[19] Lorenz F. and Masy G. Mèthode d'évaluation de l'économie d'énergie apportée par l'intermittence de chauffage dans les bâtiments. Traîtement par differences finies d'un model a deux constantes de temps. Technical Report GM820130-01, Faculte des Sciences Appliquees, University de Liege, Liege, Belgium., 1982.

[20] Crabb J.A., Murdoch N., and Penman J.M. A simplified thermal response model. *Building Serv. Eng. Res. Technology*, 8:13–19, 1987.

[21] A. Tindale. Third-order lumped-parameter simulation method. *Building Services Engineering Research and Technology*, 14(3):87–97, August 1993.

[22] Klaus Kaae Andersen, Henrik Madsen, and Lars H. Hansen. Modelling the heat dynamics of a building using stochastic differential equations. *Energy and Buildings*, 31(1):13–24, January 2000.

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

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

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

[26] Hassan Harb, Neven Boyanov, Luis Hernandez, Rita Streblow, and Dirk Müller. Development and validation of grey-box models for forecasting the thermal response of occupied buildings. *Energy and Buildings*, 117:199–207, April 2016.

[27] Damien Picard, Ján Drgoňa, Michal Kvasnica, and Lieve Helsen. Impact of the controller model complexity on model predictive control performance for buildings. *Energy and Buildings*, 152:739–751, October 2017.

[28] M.J. Jiménez, H. Madsen, and K.K. Andersen. Identification of the main thermal characteristics of building components using MATLAB. *Building and Environment*, 43(2):170–180, February 2008.

[29] Roel De Coninck, Fredrik Magnusson, Johan Åkesson, and Lieve Helsen. Toolbox for development and validation of grey-box building models for forecasting and control. *Journal of Building Performance Simulation*, 9(3):288–303, May 2016.

[30] O.M. Brastein, D.W.U. Perera, C. Pfeifer, and N.-O. Skeie. Parameter estimation for grey-box models of building thermal behaviour. *Energy and Buildings*, 169:58–68, June 2018.

[31] D.H. Blum, K. Arendt, L. Rivalin, M.A. Piette, M. Wetter, and C.T. Veje. Practical factors of envelope model setup and their effects on the performance of model predictive control for building heating, ventilating, and air conditioning systems. *Applied Energy*, 236:410–425, February 2019.

[32] Eva Žáčeková, Zdeněk Váňa, and Jiří Cigler. Towards the real-life implementation of MPC for an office building: Identification issues. *Applied Energy*, 135:53–62, December 2014.

[33] David Sturzenegger, Dimitrios Gyalistras, Manfred Morari, and Roy S. Smith. Model Predictive Climate Control of a Swiss Office Building: Implementation, Results, and Cost–Benefit Analysis. *IEEE Trans. Contr. Syst. Technol.*, 24(1):1–12, January 2016.

[34] Massimo Fiorentini, Gianluca Serale, Georgios Kokogiannakis, Alfonso Capozzoli, and Paul Cooper. Development and evaluation of a comfort-oriented control strategy for thermal management of mixed-mode ventilated buildings. *Energy and Buildings*, 202:109347, November 2019.

[35] Juan Hou, Haoran Li, Natasa Nord, and Gongsheng Huang. Model predictive control under weather forecast uncertainty for HVAC systems in university buildings. *Energy and Buildings*, 257:111793, February 2022.

[36] Jacopo Vivian, Lorenzo Croci, and Angelo Zarrella. Experimental tests on the performance of an economic model predictive control system in a lightweight building. *Applied Thermal Engineering*, 213:118693, August 2022.

[37] Brandon Amos, Lei Xu, and J. Zico Kolter. Input Convex Neural Networks. *arXiv*, 2016. Publisher: arXiv Version Number: 3.

[38] Felix Bünning, Adrian Schalbetter, Ahmed Aboudonia, Mathias Hudoba de Badyn, Philipp Heer, and John Lygeros. Input Convex Neural Networks for Building MPC. *arXiv*, 2020. Publisher: arXiv Version Number: 1.

[39] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707,

February 2019.

[40] Loris Di Natale, Bratislav Svetozarevic, Philipp Heer, and Colin N. Jones. Physically Consistent Neural Networks for building thermal modeling: theory and analysis. *Applied Energy*, 325:119806, November 2022.

[41] Gargya Gokhale, Bert Claessens, and Chris Develder. Physics informed neural networks for control oriented thermal modeling of buildings. *Applied Energy*, 314:118852, May 2022.

[42] Javier Arroyo, Carlo Manna, Fred Spiessens, and Lieve Helsen. Reinforced model predictive control (RL-MPC) for building energy management. *Applied Energy*, 309:118346, March 2022.

[43] Krzysztof Arendt, Muhyiddine Jradi, Hamid Reza Shaker, and Christian T. Veje. Comparative Analysis of White-, Gray- and Black-box Models for Thermal Simulation of Indoor Environment: Teaching Building Case Study. In *2018 Building Performance Modeling Conference and SimBuild*, 2018.

[44] International Standard Organisation - ISO. ISO 13790:2008 Energy performance of buildings — Calculation of energy use for space heating and cooling. Technical report, International Organization for Standardization, 2008.

[45] Jacopo Vivian, Angelo Zarrella, Giuseppe Emmi, and Michele De Carli. An evaluation of the suitability of lumped-capacitance models in calculating energy needs and thermal behaviour of buildings. *Energy and Buildings*, 150:447–465, 2017.

[46] Weather Data — EnergyPlus.

[47] Tobias Loga, Britta Stein, and Nikolaus Diefenbach. TABULA building typologies in 20 European countries—Making energy-related features of residential building stocks comparable. *Energy and Buildings*, 132:4–12, November 2016.

[48] Philip Santosh. eppy Documentation, February 2021.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[50] scikit-learn: machine learning in python — scikit-learn 1.0.2 documentation.

[51] Mary Ann Branch, Thomas F. Coleman, and Yuying Li. A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems. *SIAM J. Sci. Comput.*, 21(1):1–23, January 1999.

[52] William F. Holmgren, Clifford W. Hansen, and Mark A. Mikofski. pvlib python: a python package for modeling solar energy systems. *Journal of Open Source Software*, 3(29):884, September 2018.

[53] Debaditya Chakraborty and Hazem Elzarka. Performance testing of energy models: are we using the right statistical metrics? *Journal of Building Performance Simulation*, 10 2017.

Appendix A. Parameters analysis

Table A1 shows the value of the nominal parameters for each building envelope, calculated as specified in Section 2.1.

Table A1. Nominal parameters of the grey-box models.

Unit	B70	B90	BN	B	Bni
C_m	6.16E+07	6.16E+07	6.16E+07	2.60E+07	2.60E+07
$H_{tr,em}$	147.237	101.737	50.2769	56.8163	168.087
$H_{tr,is}$	816.96	816.96	816.96	816.96	816.96
$H_{tr,ms}$	6464.64	6464.64	6464.64	5387.2	5387.2
$H_{tr,w}$	13.44	13.44	13.44	13.44	13.44
H_{ve}	17.76	17.76	17.76	17.76	17.76
k_{conv}	0.95	0.95	0.95	0.95	0.95
$k_{s,gla}$	0.6	0.6	0.6	0.6	0.6
$k_{s,opa}$	0.024	0.024	0.024	0.024	0.024
ϕ_0	400	400	400	400	400
k_a	3	3	3	2.5	2.5
k_s	0.2	0.2	0.2	0.2	0.2
C_s	3.08E+06	3.08E+06	3.08E+06	1.30E+06	1.30E+06
C_i	213120	213120	213120	213120	213120

Figure A1 shows the ratio between calibrated and nominal thermal capacitances in the 3C grey-box model as a function of building type and season. In the heating season (blue dots), the calculation of the nominal thermal capacitances seem to overestimate C_m, and to underestimate C_s and C_i. As far as the cooling season is concerned (orange dots), the calibrated values of the thermal capacitance C_m are higher than the nominal values for heavyweight building envelopes. The ratio between the calibrated and nominal value decreases with increasing thermal insulation. Similar trends occur for C_s and C_i.

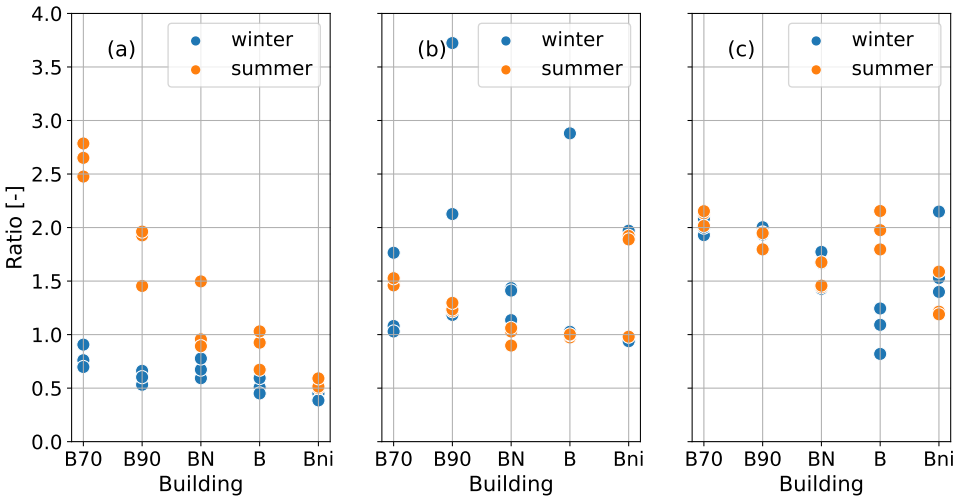


Figure A1. Ratio between calibrated and nominal thermal capacitances in the 3C grey-box model as a function of building type and season: (a) C_m, (b) C_s and (c) C_i.