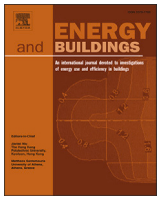
Energy & Buildings 294 (2023) 113217

Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enbuild



Assessing the impact of employing machine learning-based baseline load prediction pipelines with sliding-window training scheme on offered flexibility estimation for different building categories

Italo Aldo Campodonico Avendano [a,*,1], Farzad Dadras Javan [b,1], Behzad Najafi [b], Amin Moazami [c,a], Fabio Rinaldi [b]

[a] Department of Ocean Operations and Civil Engineering, Faculty of Engineering, NTNU, 6009 Ålesund, Norway
[b] Dipartimento di Energia, Politecnico di Milano, Via Lambruschini 4, 20156 Milano, Italy
[c] Department of Architectural Engineering, SINTEF Community, SINTEF AS, Børrestuveien 3, 0373 Oslo, Norway

ARTICLE INFO

Keywords:
Baseline prediction
Demand flexibility
Commercial buildings
Machine learning
Smart meters
Sliding-window training
Pipeline optimization

ABSTRACT

The present study is focused on assessing the impact of the performance of baseline load prediction pipelines on the estimation (by the grid operator) accuracy of the flexibility offered by different categories of buildings. Accordingly, the corresponding impact of employing different machine learning (ML) algorithms, with sliding-window and offline training schemes, for hour-ahead baseline load prediction has been investigated and compared. Using a smart meter measurements dataset, training window sizes and the most promising pipeline for each building category are first identified. Next, the consumption profiles of five buildings (belonging to each category), with the regular operation (baseline load) and while offering flexibility, are physically simulated. Finally, the identified pipelines are used for predicting the baseline loads, and the resulting error in estimating the provided flexibility is determined. Obtained results demonstrate that the identified most promising prediction pipeline (extra trees algorithm with a sliding window of 5 weeks) offers a notably superior performance compared to that of offline training (average R^2 score of 0.91 vs. 0.87). Employing these pipelines permits estimating the provided flexibility with acceptable accuracy (flexibility index's mean relative error between -2.45% to +2.79%), permitting the grid operator to guarantee fair compensation for buildings' offered flexibility.

1. Introduction

As one of the key measures to reduce fossil fuel consumption and the corresponding consequent emissions [1], global efforts have been specifically dedicated to expanding the scale of distributed renewable generation [2,3]. Introducing distributed renewable energy sources, particularly at the individual building level, given their intermittent generation due to their variable and unpredictable nature, imposes a challenge of balancing energy supply and demand in real-time (to ensure a consistent and steady energy supply) [4,5]. Smart grids are accordingly introduced to balance the load by integrating the data from end-users, producers, and prosumers while improving the efficiency of operations and controlling the distribution systems [6]. To avoid imbalances in the electrical grid, the Transmission System Operator (TSO) needs to acquire large quantities of active power reserves and ancillary services to control the system's frequency and ensure the reliability and resilience of the grid [7]. The electrical grid imbalances can be handled using flexibility on the supply side (for instance, throttling the power generation rate) or from the demand side [8]. One of the promising alternatives in this context is utilizing the demand flexibility that can be offered by buildings.

1.1. Buildings energy flexibility as a solution to handle grid imbalance

Consumption of the building sector is responsible for approximately 30% of global energy consumption by 2017. Furthermore, a jump from 33% to 55% is expected in the share of electrical consumption of buildings in the corresponding overall demand by the year 2050 [9]. In the same context, the demand for heating, ventilation, and air-conditioning (HVAC) systems in buildings accounts for almost 38% of the energy

* Corresponding author.
E-mail address: italo.a.c.avendano@ntnu.no (I.A. Campodonico Avendano).
[1] I.A. Campodonico Avendano and F. Dadras Javan equally contributed to this work.

https://doi.org/10.1016/j.enbuild.2023.113217
Received 23 March 2023; Received in revised form 16 May 2023; Accepted 31 May 2023
Available online 2 June 2023

Nomenclature

airTemperature-nh	Outdoor air temperature n hours ahead
ANN	Artificial neural network
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
CMY	Current meteorological year
CNN	Convolutional neural network
CV(RMSE)	Coefficient of Variation of Root Mean Square Error
day_of_week_sin	Day of the week encoded with sine function
dayTime_cos	Time of the day encoded with cosine function
DR	Demand response
electricity-nh	Electrical consumption n hours ahead
ETR	Extra trees regressor
FF	Flexibility function
FI	Flexibility index
GBE	Grid-interactive efficient buildings
HVAC	Heating, ventilation and air-conditioning
IPMVP	International Performance Measurement and Verification Protocol
KNN	K nearest neighbor regressor
LR	Linear regression
LSTM	Long short-term memory
MAPE	Mean absolute percentage error
MI	Mutual information
ML	Machine learning
MLR	Multi-linear regression
MRD	Mean relative deviation
NMBE	Normalized mean bias error
PACU	Packaged air conditioning unit
PTR	Peak time rebate
R^2	Coefficient of determination
RE	Relative error
RNN	Recurrent neural network
RFR	Random forest regressor
SO	System operator
solar-nh	Direct solar radiation n hours ahead
SVM	Support vector machine
SVR	Support vector regressor
TCL	Thermostatically controlled load
TMY	Typical meteorological year
TSO	Transmission system operator
UTC	Coordinated universal time
XGBR	XGBoost regressor
VAV	Variable air volume

consumption in buildings constituting around 12% of the global energy demand [10].

Apart from the notable share of the building sector in the global demand (along with the corresponding expected rise), the role of this sector is pivotal in decarbonization scenarios, owing to the type of loads in buildings and specifically the HVAC systems (that are among the thermostatically controlled loads (TCLs) [11]), which makes them a promising choice for providing demand flexibility to the grid. In this context, the *energy flexibility* of a building is the capacity for altering demand and generation upon different climatic conditions, grid, or end-user needs and will authorize demand side management [12]. The U.S. Department of Energy defines five demand-side management (DSM) strategies for grid-interactive efficient buildings (GEBs): efficiency, load shedding, shifting, modulating, and on-site electricity generation [4]. In this context, energy-flexible buildings are those that are capable of managing their energy demand and generation based on different requirements such as grid congestion, environmental conditions, and user needs [12].

By providing the buildings with flexibility in operation, DSM allows adjusting the power consumption based on the grid supplies [4,13]. Therefore, the flexibility obtained from building' DSM permits the incorporation of renewable energy sources into the grid with high efficiency. Additionally, DSM cuts off operational expenses and is much more cost-efficient than investing in increasing generation capacity, new standby power plants, or reinforcing the grid. [14–17]. DSM is also facilitated due to the penetration of smart metering systems providing energy measurements with hourly or sub-hourly frequency [18].

Residential, industrial, and non-residential buildings (e.g., offices) can be utilized to provide flexibility to the grid in a demand response program where utility providers try to balance demand and supply by designing incentive/price-based programs to encourage energy users to change their consumption behaviors for balancing the supply and demand [4,14]. Considering the application of demand-side strategies, the demand response can be applied based on direct or indirect control of the system operator (SO) [19]. In direct control, the consumer (or smart HVAC schedule management system [20]) can decide to shift/reduce the energy consumption of the building to reduce the bills based on incentive or penalty-aware signals, which can include CO_2 emissions, energy efficiency, and energy price [21,22]. In an indirect control instead, the system operator can decide in a group of buildings or in an urban area if it is possible and worth applying flexibility measures considering the welfare of each building [23].

In order to achieve demand-side flexibility, it is crucial to determine the horizon of action that the flexibility measure will have on the demand of the building, where time scales of seconds allow acting on balancing the grid under unpredictable fast changes. At the same time, more extended periods from minutes to hours permit modifying the consumption in response to changes in forecasted demand (balancing forecast error between load and generation) [24]. The baseline load refers to the load the building would have consumed if no demand response (DR) measure was in place. Thus, the load reduction refers to the difference between the baseline load and the metered reduced load during the DR event [25]. In this context, the dynamic behavior of the energy demand of an energy-flexible building under a penalty signal can be characterized by the flexibility function (FF) proposed by Junker et al. [22], in which the energy produced at time t based on signal response is described as:

$$E_{flexibility,t} = \sum_{k=0}^{\infty} h_k(\theta)\lambda_{t-k} + R_t \tag{1}$$

where R_t is the non-responsive load, $h_k(\theta)$ corresponds to the impulse response function triggered by the penalty signal λ_{t-k}, k time -steps before time t. The response function depends on external variables θ, such as weather conditions and occupancy, making the flexibility function non-linear and time-varying.

Accordingly, Fig. 1 shows the expected behavior of the demand response when a step-shaped penalty signal is imposed. The system considered a reaction time before responding to the penalty signal, to induce later a reduction (or increase) in demand that results in the flexibility available during the investigated period.

Considering the energy load based on a signal response, it is possible to define the flexibility function considered in this study as the energy reduction [2,22,26]:

$$FF = \sum_{t=0}^{T} E_{baseline,t} - E_{flexibility,t} \tag{2}$$

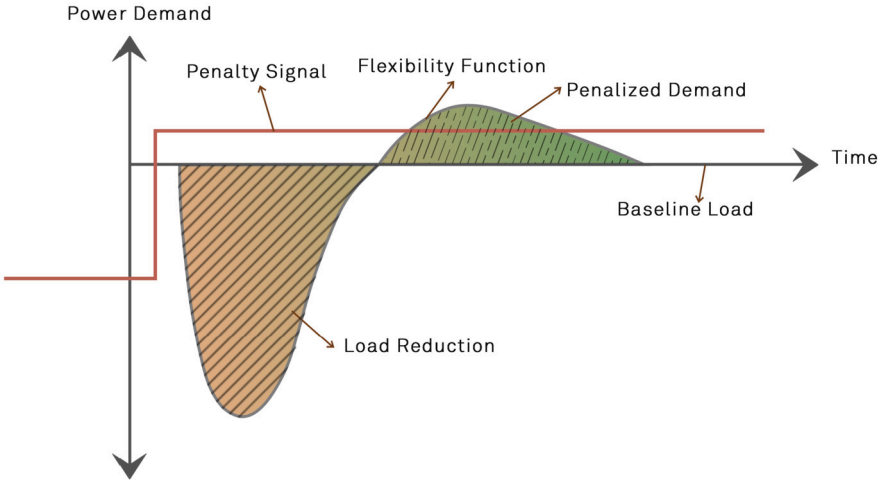


Fig. 1. The expected load behavior of buildings when flexibility measures are implemented in response to grid signals is represented, as proposed by Junker et al. [22].

where T is the duration of the demand response and $E_{baseline,t}$ is the baseline energy demand or penalty-unaware demand considered in the event that no flexibility measure has been applied to the building.

Furthermore, the Flexibility Index (FI) [22] (that is a value between 0 and 1) is used to quantify the cost related to reducing energy consumption in the peak shaving period when the grid manager is aware of the penalty signal. The Flexibility Index is determined in its discretized form using the following formulation:

$$FI = 1 - \frac{\sum_{t=0}^{T} \lambda_t E_{flexibility,t}}{\sum_{t=0}^{T} \lambda_t E_{baseline,t}} \tag{3}$$

where λ_t corresponds to the normalized penalty signal at time t, and for this case, is considered as $\lambda_t = \lambda = 1[\frac{J}{J}]$ to obtain the reduced energy consumption in the peak shaving period.

1.2. Literature review on baseline prediction during demand flexibility in buildings

In order to accurately estimate the energy flexibility offered by a building to guarantee the corresponding fair compensation, the grid operator needs a precise prediction of the baseline load consumption [27], as the error in the corresponding prediction results in unjustified financial penalization or over-compensation of DR participants [28]. Therefore, achieving an accurate and unbiased estimation of the participants' expected baseline load is crucial; thus, it is extensively studied in the literature. The methodologies employed in this area can be broadly categorized into three main approaches: Averaging, Control group, and Regression methods [25]. Averaging methods calculate the average of the historical load values on normal days (without DR events) [29–31]. In the control group method, the baseline load for DR participants is assumed to be the concurrent load data of control customers who do not participate in DR [32]. In the context of regression methods, the relationship between the input variables (features) and the target baseline load is found by fitting a linear/non-linear function [33]. Multiple Linear Regression (MLR) [29] and Support Vector Regression (SVR) are among the most commonly investigated algorithms. In this context, Chen et al. [34] used SVR to predict the demand response baseline for office buildings. The authors concluded that their model that is provided with outside temperature two hours before the DR event as an input performed better than the other DR baseline forecasting methods. To predict the energy consumption at the level of a single household to appropriately apply demand response, Estebsariand Rajabi [35] developed a hybrid model based on time-series image encoding and Convolutional Neural Network (CNN). They compared the results with other forecasting methods, such as SVM, Artificial Neural Network (ANN), and CNN, and achieved a mean absolute percentage error of around 12%. In a research conducted by Sha et al. [27], a toolkit for demand response baseline calculation was proposed. Data were collected from 20 commercial buildings. The mean coefficient of variation of the root mean square error (CV-RMSE) of all the buildings aggregated was shown to be less than 8%.

In the general area of load forecasting, Khalid et al. [36] deployed a recurrent neural network (RNN) and long short-term memory (LSTM) to predict the electricity demand and price values for one week, one month, and three months. They used multiple variables as inputs for LSTM, which obtained more accurate results compared to the conventional univariate LSTM. Fan et al. [37] proposed an approach based on integrating LSTM with human behavior pattern recognition to predict 24-hour demand. The dataset consisted of 8-week electricity consumptions from 2337 residential customers. The model was additionally improved with a multi-layer neural network. The obtained results demonstrated that in 24 hours, 94% of customers would receive hourly load prediction with a MAPE of less than 20%. However, deep learning models require a significant amount of data and would not perform on single building levels. Moreover, results obtained by SVR and Extreme Gradient Boosting (XGBoost) were shown to be outperforming LSTM in a benchmark analysis performed by Huang et al. [38]. They also concluded that the selection of the machine-learning model should be based on building energy data's natural characteristics, and the hyperparameter tuning or mathematical modification within an algorithm would not be sufficient. In this research CV of 14.25% was obtained in the best-case results. Moreover, Cerquitelli et al. [39] collected smart data from 12 residential buildings with district heating in Italy over the winter and the corresponding weather data to perform predictions on the buildings' consumption using the sliding window method. The study was limited to using data between 5–10 p.m., and accuracy of MAPE between 6–19% was achieved for an hour ahead prediction. However, the results deteriorated when the procedure was expanded to the whole day [39]. Comparing the performance of five different state-of-the-art machine learning (ML) algorithms in predicting the heating energy of a Chinese residential district, Wang et al. [40] discovered that *random forest* outperforms the other algorithms owing to its precision, robustness, and interpretability. A systematic literature review by Al-Shargabi et al. [41] demonstrated that most studies focus on load forecasting for residential buildings.

However, no previous comprehensive study has been dedicated to assessing the impact of the performance of load prediction pipelines on the estimation (by the grid operator) accuracy of the flexibility offered by different categories of buildings. Specifically, the baseline load

prediction performance of tree-based machine learning algorithms, particularly while employing the sliding window training scheme, and the resulting impact on offered flexibility estimation should still be assessed. Furthermore, given the particular characteristics of the load profiles of buildings belonging to specific use type categories, the aggregate benchmarking results (for all buildings with different categories) reported in the literature do not necessarily provide useful insights for specific building categories. Moreover, to the best of the authors' knowledge, non of these studies have tested the performance of the benchmarked pipelines through physics-based simulations to predict the baseline load during a demand response event to quantify the extent to which the baseline load prediction error is propagated in the estimation of the demand flexibility offered by a building.

1.3. Contributions of the present study

Motivated by the above-mentioned research gap, the present work is focused on investigating the accuracy of the estimation (by the grid operator) of offered flexibility (by buildings), which is achieved by employing ML-based baseline load forecasting pipelines with sliding window and offline training schemes. Accordingly, the first part of the study is focused on implementing machine learning (ML)-based pipelines using benchmark (state-of-the-art) ML algorithms aimed at one-hour ahead load prediction for a large set of buildings belonging to different categories of use. Offline and sliding window training schemes are implemented and, for each building category, the most promising algorithm as well as the optimal size of the training window (for the second scheme), which leads to the highest accuracy, are identified. It is noteworthy that the prediction horizon is chosen to be one hour as it is one of the most commonly utilized sampling rates of smart meters, and it is coherent with the flexibility time horizon considered for commercial and industrial buildings [42]. This time horizon is already in use in some implemented policies, such as the emergency load reduction program of the California State [43], which targets voluntary energy flexibility in a specific time schedule for a time horizon between 1 and 6 hours.

In the second part, the physical modeling of five example buildings belonging to the investigated service categories has been performed to simulate the corresponding consumption profile with the regular operation schedule (baseline load) and while undergoing a flexibility measure (increase in the cooling setpoint in the context of a demand response program). By comparing these two profiles, the real flexibility offered by the building (represented in terms of the Flexibility Index) is determined. Next, in order to determine the estimated offered flexibility (by the grid operator or the demand side management program's authority), pipelines proposed in the first phase of the work are utilized for predicting the baseline load. By comparing the simulated and estimated flexibility index, the corresponding achieved accuracy for different building categories is investigated. Hence, The contributions of this paper can be summarized below:

- Evaluating the performance of different regression algorithms, including ensemble models that use decision trees, for predicting the baseline load during demand response events.
- Investigating the performance of ML-based baseline load forecasting pipelines with offline and sliding window training schemes and identifying the most promising algorithm and training window size for each building category
- Physics-based simulations of regular operation (baseload) and demand response scenario for modeled sample buildings (belonging to each of the categories under investigation).
- Assessing the performance of the benchmarked pipelines in predicting the baseline load, while undergoing the DR scenarios, and quantifying the impact of the corresponding prediction error on the estimation accuracy of the offered demand flexibility.

2. Case study

This section first describes the dataset used to benchmark the ML-based pipelines for baseline load forecasting. Next, the physical models that are utilized to simulate the flexibility scenario and assess the offered flexibility (deploying the predictive pipelines to estimate the baseline load and assess the load reduction) are presented.

2.1. Utilized dataset

'Building Data Genome Project 2' [44] a publicly available dataset, which comprises two years (2016 and 2017) of hourly measurements of electrical, heating, chilled water, and steam consumption from sites scattered in North America and Europe is employed. The database of each building comes with additional metadata, which includes the category (and sub-category) of the building, along with the corresponding specific energy consumption. Lastly, the weather data per location are included in the same database. In the present work, the incident solar radiation obtained from *Prediction of Worldwide Energy Resource (POWER) [45]* is also added to the available weather data, and some corrections regarding the time zones and units are applied.

2.2. Characteristics of the physical models

The modeling procedure was performed using *EnergyPlus V9.4* [46] software and its *Python API* [47]. The models used in this study consist of an office, a primary school, a mid-size residential building, a swimming pool for the *entertainment* category, and a library for the *public services* category, as shown in Fig. 2. The first three buildings were developed under the *ANSI/ASHRAE/IES Standard 90.1* [48] and have been provided by a study conducted by Deru et al. [49]. The library building model instead corresponds to a small office model from the same study, where the internal gains, occupancy, and lighting were adapted following the *ASHRAE* criteria. The swimming pool building simulated in this study is present in the example buildings provided by *EnergyPlus*. The simulations are conducted assuming that the models are situated in Los Angeles, California (Climatic zone 3B), one of the locations included in the validation approach's data set. Moreover, the weather file corresponding to the exact location was first converted to the current meteorological year (CMY) for 2016 and 2017 starting from the typical meteorological year (TMY) weather file using *diyepw* tool [50], and then were added and deployed in the simulations. The detailed specifications of the buildings are presented in Table 1.

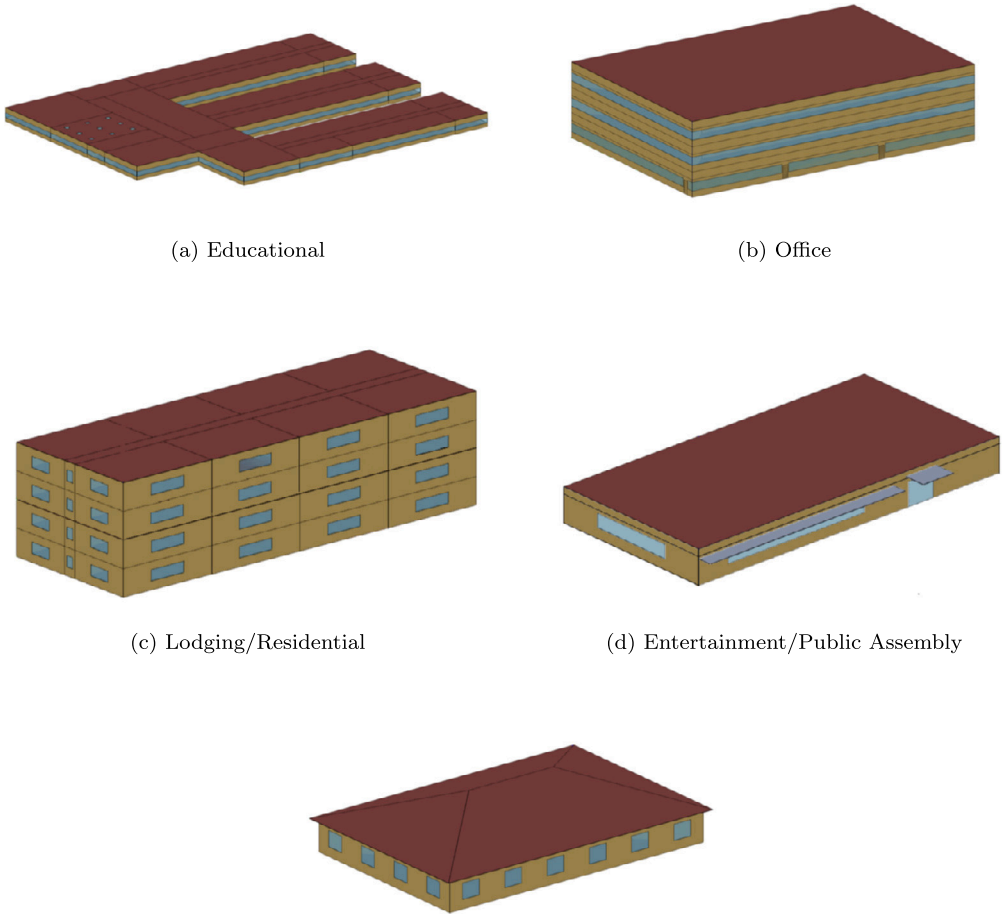
3. Methodology

The first step of the present work is dedicated to implementing and benchmarking ML-based pipelines for baseline load forecasting (an hour ahead) of different categories of commercial buildings. The second step is instead focused on simulating the offered flexibility employing the physical models of 5 different categories of buildings. Finally, the predictions provided by the most suitable pipelines (identified in the first step) are utilized together with the results of the flexibility simulations (second step) to assess the accuracy of the estimation of offered flexibility (represented as the flexibility index) that is performed by the grid operator.

A schematic summary of the implemented methodology is presented in Fig. 3.

3.1. Benchmarking ML-based pipelines for baseline load forecasting

The first part of the work is focused on benchmarking ML-based pipelines with a sliding window training scheme for one-hour-ahead baseline load prediction in large buildings. In this context, the utilized dataset is first described, and the methodology employed for implementing the ML-based pipelines is then presented.



(a) Educational (b) Office

(c) Lodging/Residential (d) Entertainment/Public Assembly

(e) Public Service

Fig. 2. Sample buildings utilized in physical phenomena-based simulations in *EnergyPlus*.

Table 1
Description of the buildings used in the physical-based simulations in *EnergyPlus*.

Type	Education	Office	Lodging / Residential	Entertainment / Public Assembly	Public Services
Total Floor Area	6871 [m^2]	4979.6 [m^2]	3135 [m^2]	463.6 [m^2]	511 [m^2]
Orientation	0° NE	0° NE	0° NE	30° NE	0° NE
Thermal Zones	25	18	27	5	5
Window Fraction	35%	33%	15%	29%	21%
Heating type	Gas water boiler/Furnace	Gas water boiler	Gas furnace	Gas water boiler	Gas furnace
Cooling type	Packaged air conditioning unit (PACU)	PACU	Split system	Compression chiller with air cooled condenser	Unitary split
Distribution	Centralized air-handling unit - PACU	Variable air volume (VAV) system	Centralized air-handling unit - PACU	VAV system	Centralized air-handling unit - Split

3.1.1. Data cleaning and filtering

Based on the scope of this work and considering the information provided that was provided about the utilized dataset (section 2.1), only the buildings in which the smart meter data includes the consumption of an electrically fed cooling system should be considered. Therefore, only the buildings with chilled water meters (along with the electrical smart meter) are included. Then, a correlation analysis was performed since it was observed that there is a low correlation between the 'chilled water and electrical meter' in some buildings. For that, Pearson's correlation and normalized Mutual Information have been considered to find the correlation between electricity and chilled water consumption (along with the corresponding lagged value). A moderate correlation of 0.4 [51] has been defined and utilized as a threshold to filter the buildings for which a correlation is observed.

The next step consists of filtering out the buildings without a category or a specified service, followed by the outlier detection based on the work shown by Kissell [52], which consists of removing the buildings outside of the range $\mu \pm 2\sigma$ of specific energy consumption (where μ represents the average, and σ is the standard deviation of the feature used for the filtering).

Considering the data processing and the filtering of the building, the final dataset used for this study consists of 99 buildings in which chilled water demand is correlated to electrical consumption. These buildings belong to five service categories: *education*, *office*, *lodging/residential*, *entertainment/public assembly*, and *public services*; and are positioned in seven locations in North America, as observed in Fig. 4.

The buildings of the various categories are widely spread in different areas (see Fig. 5) where the *educational* and *office* building categories,

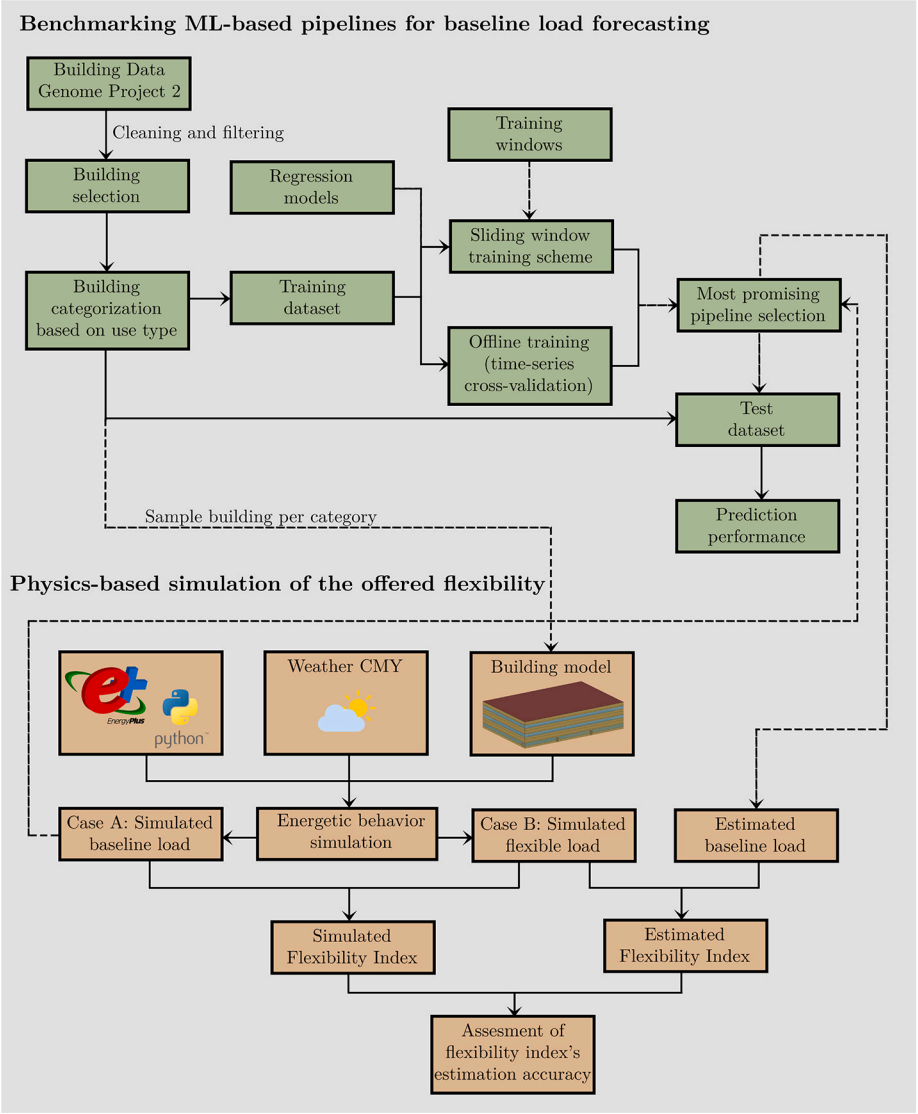


Fig. 3. Summary of the implemented methodology in this work.

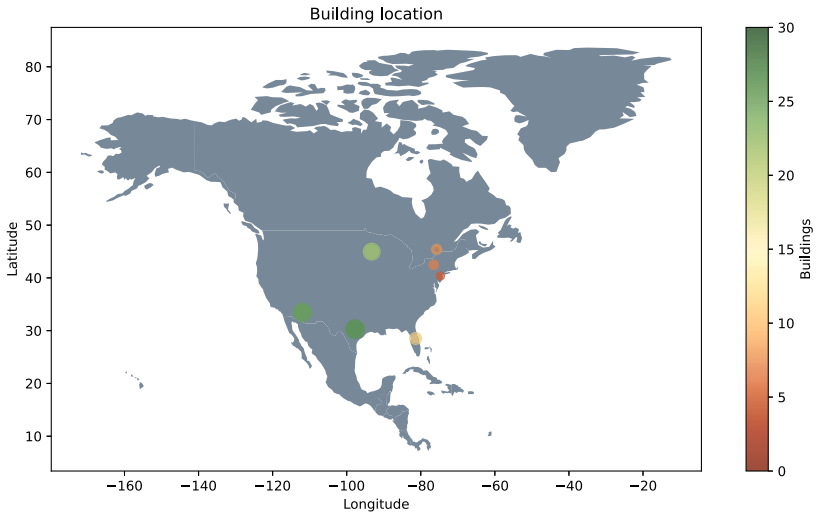


Fig. 4. Geographic distribution of the sample buildings used in the benchmarking ML-based pipelines for baseline load forecasting. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

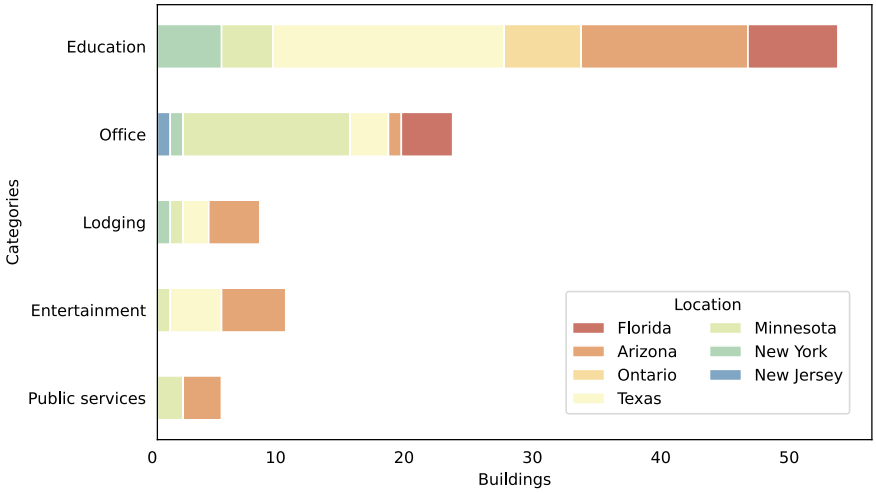


Fig. 5. Buildings' distribution in the different locations, aggregated by category.

with 53 and 23 each, incorporate most of the buildings of this study, as shown in Fig. 5.

3.1.2. Implementation of machine learning-based pipelines

Five different machine learning regressor algorithms, such as: *'random forest'*, *'extra trees'*, *'XGBoost'*, *'support vector machine'* and *'k nearest neighbor'*, and *linear regression* are used for the hour ahead prediction of each building's consumption. The outdoor conditions (ambient temperature and direct solar radiation), the encoded day of the week, the time of the day, and the lagged features of these variables, along with the lagged values of the electricity consumption, are employed as input features. The features regarding the chilled water meter consumption are deployed in the filtering stage to ensure the correlation of the electrical consumption of the building with chilled water usage and are not used in the modeling procedure. Various features are utilized to train the regression models for hour-ahead prediction of the baseline electrical load, including the outdoor temperature, direct solar radiation, and electrical consumption, considering 1, 2, 3, and 4 hours ahead of the prediction. Furthermore, the day of the week (encoded in a sine function) and the hour of the day (encoded in a cosine function) serve as additional training features. Although the flexibility scenarios in the physics-based simulation are triggered by setpoint modifications, in order to develop generalizable models (only employing hourly smart meter data) that are not bound to limitations in the dataset (as in the case of the *'Building Genome Data Project 2'*, which only provides smart meter data and does not include any information about indoor temperature setpoint), this feature is not used in the training procedure of ML-based algorithms.

Two different training approaches are next implemented while utilizing the summer period (from July to September) of the first year (2016) for the training/validation procedure. In the first approach, offline (batch) learning is implemented and assessed using time-series cross-validation. In the second approach instead, the sliding window training scheme [53] (the description of which is provided in the section 4.3) is implemented. In order to determine the most suitable size of the training window, this training and validation scheme is implemented with six window sizes of one week (168 hours), two weeks (336 hours), three weeks (504 hours), four weeks (672 hours), five weeks (840 hours), and six weeks (1008 hours). It is noteworthy that in order to assess the general performance of machine learning models without being optimized for a specific building, the corresponding hyperparameters are not fine-tuned. Furthermore, the implementation of around 172800 pipelines and the utilization of a sliding window training scheme in this research makes optimizing the algorithms' hyperparameters impractical, while it also impedes the above-mentioned aim of reporting models' generalizable performance. *Mean absolute percentage error (MAPE)* and *coefficient of determination (R^2 score)*, are utilized to assess the performance of the prediction pipelines, while the latter one is used as the primary metric. The resulting determined average performance for each category of buildings is then calculated. The most promising pipelines that result in the highest R^2 score (using the offline and sliding window approaches, respectively) for each category of buildings are next determined. Finally, in order to assess the performance of identified most promising pipelines, the corresponding performance over the test subset (the same period in the next year) is investigated.

3.2. Physical simulation of the offered flexibility

This subsection presents the second stage of this work, which involves simulating the baseline load profile and the offered flexibility employing physics-based energetic behavior simulation models.

3.2.1. Simulation methodology

Two scenarios are considered to simulate the offered demand flexibility. The first case (Case A), referred to as the "simulated baseline consumption", includes the regular working schedule of the buildings involving predefined setpoint temperatures for the different zones during working hours. The second scenario (case B), referred to as "flexible consumption", simulates the implementation of a load-shifting scheme for a demand response program, such as *Peak Time Rebates (PTR)*, where pre-established peak periods are considered, during which the customers receive an incentive upon the reduction of their demand [4]. A case of this demand response program has already been deployed by the State of California [43], in which buildings can reduce congestion on the electrical grid during peak hours between June and October each year. In this simulation, the cooling setpoint is raised for an interval of one hour once a signal (for reducing consumption) is received at 4 p.m. (the maximum rise in zones' temperature is kept below 2 °C). By comparing the two simulated consumption profiles, the corresponding real offered flexibility (represented in terms of flexibility index) is determined. The simulations are conducted considering two CMY weather files (2016 and 2017) for Los Angeles, CA, to provide a similar comparison with the ML-based pipeline benchmarking procedure. The results obtained from the simulation of the year 2016 are employed as the training data with the offline approach, while both sliding window and offline training methods are tested on the simulations performed for the year 2017. Details of these two scenarios are provided in the following subsections:

Case A The first case consists of baseline consumption simulation under the standard operation schedule of the building. Therefore, a yearly simulation with a frequency of 10 minutes is conducted considering a standard setpoint temperature of 24 °C for cooling. During off-schedule hours, the cooling setpoint temperature was increased to 30 °C to avoid the operation of HVAC equipment. The working hours are defined from Monday to Friday (except for holidays) from 7 a.m. to 6 p.m., while the off-schedule hours include all the other intervals, weekends, and holidays.

Case B In Case B, the flexibility measures proposed for the California State for grid decongestion based on incentives are applied under the same conditions of Case A. Therefore, in the period between June 1st and October 31st and at 4 p.m. and for a maximum of one hour, a simulation is performed, considering a reduction in the electricity consumed based on setpoint modifications for cooling demand, establishing a maximum increase of 2 °C (till 26 °C) in this period. The process is terminated if any of the zones of the building independently exceed their corresponding setpoint temperature and the building is restored to the standard conditions. Since this strategy is based on a signal, the modifications in the setpoints are produced randomly (50% of possibilities to occur) on each day of the working week. The setpoint management and control have been implemented by the package *EMSPY*, developed by [54], which provides an interface for the interaction between *EnergyPlus* and its *Python API*.

3.2.2. Offered flexibility estimation and the assessment of the corresponding accuracy

In order to estimate the real offered flexibility while a building is undergoing a flexibility measure in the context of a demand response program, the modified consumption of the building should be compared with the consumption profile that the building would have had if flexibility measures were not taken (baseline load), which is clearly not available. In this context, an alternative would be employing prediction pipelines that are trained using historical data of buildings' regular operation. Accordingly, the most promising offline and the sliding window (with optimal training window size) training schemes pipelines proposed in the first phase are next utilized to predict the (simulated) baseline consumption during the periods where the flexibility measures are implemented. Therefore, simulations are conducted for each building using the 2016 weather data, and the corresponding electrical baseline consumption is found, which allows for the training of the offline pipelines prior to their deployment for predicting the baseline of 2017. Next, the real offered flexibility index is determined using the simulated baseline load consumption with the weather data of 2017 and the demand profile while undergoing flexibility measure. The estimated flexibility is instead calculated while employing the predicted baseline load instead of the simulated one. Finally, the average (global) relative error in the estimation of the offered flexibility (represented in terms of the Flexibility Index) for each category of buildings is calculated.

4. Machine learning-based pipeline implementation: description of correlation indexes, algorithms, sliding window training scheme, and utilized accuracy metrics

In the present section, the description of the correlation index, algorithms, sliding window training schemes, and the accuracy metrics that have been used in machine-learning-based pipeline implementation are provided.

4.1. Importance and relations between features

This present sub-section explains a theoretical description of two different correlation methods used to filter the data.

4.1.1. Pearson's correlation

Pearson's correlation coefficient is a measure of the linear relationship between two variables [55]. It is calculated as the covariance ratio of the two variables to the product of their standard deviations, as shown in Eq. (4).

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{4}$$

where r is the *Pearson's correlation* coefficient, and $\mathrm{cov}(X,Y)$ is the covariance of X and Y, with σ_X and σ_Y being the standard deviations of X and Y, respectively. This coefficient ranges from -1 to 1, where -1 indicates a strong negative relationship, 0 indicates no relationship, and 1 indicates a strong positive relationship.

For comparison purposes, the results of Pearson's correlation are presented as absolute values between 0 and 1.

4.1.2. Mutual information

Mutual information (MI) is a metric that quantifies the non-linear dependence between two random variables. It is a non-negative value representing the level of information that a variable can provide about the other, capturing the overall relationship between the variables, and can be used to assess the strength of that relationship. In the following equation, based on the results of two random variables X and Y, Mutual Information can be calculated in its discrete expression [56,57]:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{5}$$

where $I(X;Y)$ is the mutual information of random variables X and Y, and $p(x,y)$, $p(x)$, and $p(y)$ are the joint probability, the marginal probability of X, and the marginal probability of Y, respectively.

In the development of this work, a python package named *Ennemi* [58] has been used to calculate the *Mutual information* of different variables using the k-nearest neighbor search (by default, three neighbors). Additionally, the results are normalized in a range between 0 and 1 using Eq. (6), where 0 indicates a strong negative relationship, and 1 indicates a strong positive relationship. In the case of total linearity between the two different variables and if (X,Y) is normally distributed, *Pearson's correlation* coefficient will have the same absolute value as the normalized *mutual information* proposed by Laarne et al. [58].

$$\rho_{I(X;Y)} = \sqrt{1 - exp(-2\ I_{X;Y}))} \tag{6}$$

4.2. Regression algorithms

In the present sub-section, five different machine learning regression models and *linear regression* that are utilized in the presented work and are sourced from the Scikit-learn [59] and XGBoost [60] python libraries are briefly presented.

4.2.1. Linear regression (LR)

Linear regression is a statistical method that is used to determine the linear relationship between a dependent variable and one or more independent variables. The goal of *linear regression* is to minimize the residual sum of squares between the observed data and the predicted values by using a linear approximation [61]. In the case of multiple independent variables, *linear regression* can be mathematically represented as follows:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n \tag{7}$$

where the dependent variable is represented by y, the independent variables are represented by x_1, x_2, ..., x_n, and the coefficients are represented by b_0, b_1, b_2, ..., b_n.

4.2.2. Random forest regressor (RFR)

Random forest [62,63] is a machine learning model based on an ensemble of decision trees to make predictions. Each tree in the forest is

trained on a random sample of the data, and the final prediction is made by combining the outputs of all the individual trees. This approach may be used to reduce overfitting and improve the model's overall accuracy by creating a more generalized model that is less sensitive to the specific details of the training data. Given the scope of this work, *random forest* is used for regression tasks.

4.2.3. Extra tress regressor (ETR)

Extra trees regressor [64] is a machine learning algorithm based on the principles of random forest, which similarly employs an ensemble of decision trees to make predictions. However, the trees in an *extra trees* model are trained using a more randomized approach that helps to improve their ability to generalize and make accurate predictions on new data.

4.2.4. XGBoost regressor (XGBR)

XGBoost (extreme gradient boosting) [60] is a machine learning library that uses gradient boosting to make predictions, which in this work's context is the regression tasks. It trains weak decision trees, then combines them to create a strong learner that can accurately predict outcomes. At each iteration, the algorithm uses the errors from the previous iteration to train the next tree, with the goal of minimizing the overall error of the model. This process is repeated until the desired number of trees has been trained, at which point the model is ready to make predictions on new data.

4.2.5. Support vector regressor (SVR)

A *support vector regressor* [65] is a specialized form of *support vector machine (SVM)* designed to perform regression tasks. At its core, a support vector regressor seeks to identify the hyperplane that best divides the data, maximizing the margin between the hyperplane and the data points. By maximizing this margin, the *support vector regressor* can improve the generalizability and robustness of the model. Additionally, the *support vector regressor* has several advantages compared to other regression methods, such as the ability to handle high-dimensional data and the use of regularization to prevent overfitting.

4.2.6. k-Neighbors regressor (KNR)

k neighbors regressor [66] is an instance-based learning algorithm that can be used for regression tasks. Here, the value of k is a user-defined hyperparameter that determines the number of nearest neighbors to consider when making a prediction. The algorithm calculates the distances between the new data and the training data points. Then, it selects the *k nearest neighbors* based on these distances, and finally, it calculates the mean or median of those neighbors as the predicted value for the new data point. Because *k neighbors regressor* is an instance-based learning algorithm, it does not build a model to make predictions. Nevertheless, it makes predictions based on the similarity of the new data point to the training data, which indicates that the algorithm can be computationally efficient and yet be sensitive to noise in the training data.

The choice of these algorithms is based on their different capabilities. For example, three-based models are capable of modeling complex relationships between variables, but KNN and linear regression are simpler to interpret and scalable since they can perform better when testing data is not present in the range of the training data (models are able to interpolate), property that is not shared with tree-based models.

4.3. Sliding window training

The offline learning method involves training the models utilizing the entire training set (batch training), whereas the sliding window training constantly performs training on a fixed-sized subset of the most recent data. In this approach, a training window with a fixed size is slid over the whole dataset (by continuously adding more recent data and discarding the old ones), providing prediction for each hour, the average accuracy of which is reported.

4.4. Evaluation metrics

The present section illustrates the metrics used to compare the accuracy of the different models and the physics-based simulation.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

$$RE_i = \frac{y_i - \hat{y}_i}{y_i} \tag{9}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|RE_i\right| \tag{10}$$

$$MRD = \frac{100\%}{n}\sum_{i=1}^{n} RE_i \tag{11}$$

The *coefficient of determination* or R^2 (Eq. (8)) is considered the main accuracy metric for the selection of the best pipelines for the different categories of buildings, while the *mean absolute percentage error* or *MAPE* (Eq. (10)) is presented to quantify the mean absolute deviation error in terms of percentage, making it comparable among the different studied building. Nevertheless, the *relative error* (Eq. (9)) and the *mean relative deviation* or *MRD* (Eq. (11)) are utilized to present the associated accuracy when the *flexibility index* is calculated in the physics-based simulations.

Following the ASHRAE Guideline 14 [67,68], the Normalized Mean Bias Error (NMBE) and the Coefficient of Variation of the Root Mean Squared Error (CV(RMSE)) are also utilized for assessing the reliability of the results.

$$NMBE = \frac{100\%}{\bar{y}}\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n} \tag{12}$$

$$CV(RMSE) = \frac{100\%}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1}} \tag{13}$$

The NMBE is a standard performance metric that calculates the mean error of a sample, with negative values indicating over-predictions and positive values indicating under-predictions. However, this metric can be affected by the cancellation of errors leading to underestimation of the actual error.

CV(RMSE) represents the capability of the model to predict the overall shape of the data, and contrary to NMBE, it is not subject to cancellation of errors [67–69].

ASHRAE Guideline 14 considers an hourly prediction model calibrated when the NMBE value is within the range of ±10%, while for CV(RMSE), the maximum acceptable error is 30%. Furthermore, the International Performance Measurement and Verification Protocol (IPMVP) [70] establishes a more stringent criterion, defining the permissible range for NMBE within ±5% and specifying that CV(RMSE) values must fall below 20%.

5. Results and discussion

In the first part of the present section, the results of the developed machine learning-based pipelines for base load prediction are presented. The subsequent part involves the conducted physics-based simulations of base and flexible loads to determine the offered flexibility. Lastly, employing the physics-based simulations of the flexibility provided by different categories of buildings and the corresponding base-load estimations presented in the first part, the accuracy of the resulting estimation of offered flexibility (utilizing the *flexibility index* index as the metric) is investigated and discussed.

5.1. Results and discussions on ML pipeline implementation for base-load forecasting

As was pointed out in section 3.1.2, hour-ahead prediction pipelines, employing five machine learning algorithms and linear regression, uti-

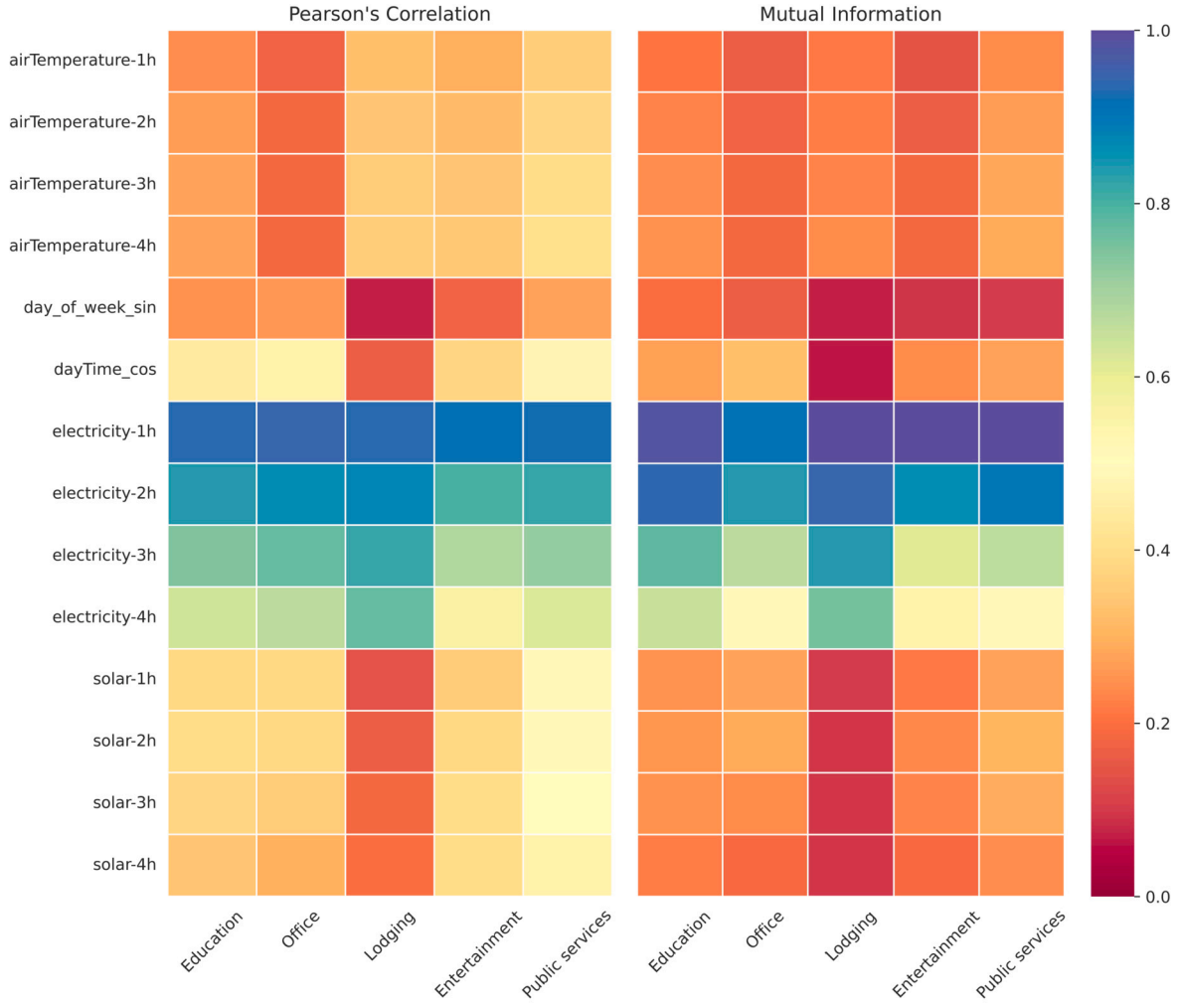


Fig. 6. Pearson's correlation (left) and mutual information (right).

lizing an offline (batch) training approach, are first developed. For each building, the performance of the pipelines is assessed (through time series cross-validation considering *coefficient of determination* as the key evaluation metric) employing the available data from June to September 2016. The average performance achieved by each machine learning algorithm for each category of buildings is subsequently determined. Next, the performance of these pipelines while implementing the sliding window training scheme employing five different training window sizes is assessed (with the same evaluation metric). In the next step, the most promising offline and sliding window pipelines for each building category, which result in the highest accuracy, are determined.

Fig. 6 presents the 14 features that are used to train the ML along with the heatmap of *Pearson's correlation* and *mutual information* between each feature and the predicted consumption for each considered category. In general, there is a strong correlation and mutual information (ranging between 0.6 to 1) between electrical consumption and its lagged values across all building categories. Public services and education buildings exhibit a higher correlation with solar radiation variables and time of day, indicating a significant reliance on seasonal daily patterns. Lodging, entertainment, and public services also show a relatively high correlation with outdoor temperature.

The obtained average performance of the implemented pipelines, for each building category, represented using R^2 score and MAPE, are provided in Table 2. Considering R^2 score as the key metric, it can be observed that in all of the investigated building categories, pipelines with sliding window training schemes offer a notably higher accuracy compared to pipelines with offline training. Furthermore, it is shown that using the extra trees regression algorithm in all of the implemented pipelines with sliding window training (regardless of the training window size) results in the most elevated performance. The optimal size of the training window is also demonstrated to be in the range of 4 to 6 weeks for different categories of buildings. However, increasing the size of the training window beyond two weeks is shown to result in rather marginal improvements in all of the considered categories. For instance, the *extra trees regressor* offers an average R^2 score of 0.935 with six weeks of training for *educational* category, while an only slightly lower R^2 score (0.93) could also be achieved using a window size of two weeks.

In addition, among the developed offline prediction pipelines, *extra trees regressor* has also been demonstrated to be the algorithm with the highest achieved performance for three building categories of *educational*, *entertainment/public assembly*, and *public services*. For *office* and *lodging/residential* categories instead, the linear regression algorithm provides the highest performance. Moreover, it can be observed that *support vector regressor* or *k nearest neighbors* offers a notably lower performance compared to the other state-of-the-art algorithms. Thus, at least without any additional pre-processing step or hyper-parameter tuning, these two algorithms are clearly unsuitable choices for the implemented baseline prediction pipelines.

Taking into account the *mean absolute percentage error (MAPE)* as the performance metric provides similar insights about the most promising pipelines. As can be observed in Table 2, pipelines implemented employing extra trees regressor with sliding window training scheme are shown to result in the lowest *MAPE* score, while the optimal size of the

Table 2
R^2 and *MAPE* scores grouped results of the baseline prediction for the validation dataset.

Model	Education - R^2							Education - MAPE [%]						
	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline
LR	0.88	0.885	0.885	0.888	0.889	0.889	0.88	8.04	7.76	7.59	7.64	7.62	7.61	7.83
RFR	0.908	0.919	0.923	0.925	0.927	0.927	0.893	6.26	5.68	5.5	5.41	5.33	5.35	7.16
XGBR	0.898	0.912	0.917	0.918	0.921	0.923	0.879	6.18	5.67	5.5	5.37	5.32	5.29	7.4
ETR	0.923	0.93	0.932	0.933	0.933	**0.935**	**0.896**	5.57	5.2	5.11	**5.05**	5.19	5.13	**7.11**
SVR	0.612	0.715	0.75	0.764	0.772	0.778	0.732	16.55	14.2	13.36	13.03	12.9	12.86	12.52
KNR	0.68	0.725	0.739	0.745	0.749	0.749	0.684	13.08	11.5	10.97	10.77	10.71	10.41	12.37

Model	Office - R^2							Office - MAPE [%]						
	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline
LR	0.854	0.863	0.864	0.864	0.864	0.863	**0.854**	9.7	9.49	9.42	9.38	9.34	9.32	9.43
RFR	0.877	0.887	0.891	0.891	0.892	0.892	0.791	7.44	6.71	6.43	6.25	6.16	6.16	7.12
XGBR	0.861	0.873	0.877	0.878	0.879	0.878	0.778	7.34	6.73	6.53	6.41	6.33	6.34	7.41
ETR	0.887	0.892	0.894	**0.895**	0.894	**0.895**	0.792	6.6	6.19	6.04	5.93	5.88	**5.85**	**6.99**
SVR	0.476	0.595	0.638	0.657	0.67	0.678	0.657	18.69	15.79	14.76	14.04	13.56	13.15	13.28
KNR	0.601	0.629	0.642	0.646	0.649	0.647	0.475	17.64	15.76	14.74	14.31	13.95	13.77	15.08

Model	Lodging/residential - R^2							Lodging/residential - MAPE [%]						
	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline
LR	0.849	0.857	0.858	0.859	0.859	0.858	**0.839**	4.98	4.86	4.85	4.86	4.88	4.9	**5.38**
RFR	0.857	0.861	0.866	0.867	0.867	0.866	0.674	4.52	4.42	4.36	4.31	4.29	4.31	6.32
XGBR	0.833	0.839	0.848	0.85	0.853	0.855	0.636	4.76	4.62	4.57	4.53	4.5	4.49	6.65
ETR	0.866	0.87	0.875	0.876	**0.877**	0.876	0.671	4.34	4.26	4.19	**4.18**	**4.18**	4.19	6.35
SVR	0.623	0.65	0.662	0.662	0.66	0.662	0.541	8.23	6.83	6.79	6.75	6.72	6.72	8.25
KNR	0.715	0.713	0.695	0.682	0.668	0.654	0.264	7.25	7.08	7.31	7.48	7.63	7.77	11.08

Model	Entertainment/public assembly - R^2							Entertainment/public assembly - MAPE [%]						
	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline
LR	0.857	0.861	0.861	0.861	0.861	0.86	0.853	8.25	8.07	7.99	7.98	7.94	7.92	8.09
RFR	0.886	0.896	0.901	0.904	0.906	0.905	0.86	6.43	6.02	5.8	5.72	5.66	5.69	6.92
XGBR	0.875	0.888	0.892	0.896	0.897	0.9	0.851	6.62	6.2	6.02	5.98	5.94	5.89	7.13
ETR	0.903	0.91	0.912	0.914	**0.915**	0.914	**0.867**	5.94	5.62	5.51	5.44	**5.42**	**5.42**	**6.76**
SVR	0.461	0.547	0.585	0.61	0.627	0.641	0.631	17.27	14.74	13.36	12.57	12.08	11.71	11.65
KNR	0.71	0.742	0.757	0.767	0.771	0.774	0.708	11.03	10.24	9.88	9.69	9.52	9.43	10.6

Model	Public services - R^2							Public services - MAPE [%]						
	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline	1-week	2-weeks	3-weeks	4-weeks	5-weeks	6-weeks	Offline
LR	0.876	0.878	0.879	0.878	0.878	0.877	0.872	9.75	9.57	9.59	9.56	9.57	9.52	9.86
RFR	0.903	0.915	0.919	0.92	0.922	0.923	0.882	7.47	6.83	6.58	6.37	6.29	6.25	7.62
XGBR	0.898	0.91	0.915	0.916	0.919	0.921	0.876	7.49	6.89	6.74	6.56	6.38	6.37	7.82
ETR	0.921	0.931	0.934	0.935	0.936	**0.937**	**0.888**	6.54	6.0	5.78	5.68	5.59	**5.54**	**7.24**
SVR	0.375	0.504	0.553	0.586	0.612	0.631	0.62	28.84	25.87	24.17	22.99	22.09	21.31	20.73
KNR	0.763	0.818	0.84	0.845	0.847	0.85	0.8	12.51	10.49	9.71	9.46	9.27	9.13	10.28

training window is demonstrated to be between 4 and 6 weeks. Furthermore, this algorithm is also shown to be the most suitable choice for pipelines with offline training for all building categories except that of *lodging/residential* buildings (for which linear regression leads to a slightly lower *MAPE* score). Moreover, pipelines with sliding window training scheme evidently result in lower *MAPE* scores compared to those offered by the pipelines trained using the offline approach.

Finally, to assess the performance of the determined most promising prediction pipelines for unseen data, the corresponding accuracy while being applied to the test subset (that includes the same period in the following year) is investigated. Accordingly, for each building category, the pipelines that result in the highest validation R^2 score, with sliding window and offline training schemes, are first chosen. Following the previously discussed results, for pipelines with sliding window training, *extra trees regressor* is employed as the algorithm while a training window of 6 weeks is utilized for *education*, *office* and *public services* categories. A training window of 5 weeks is instead employed for the rest of the categories. In the case of offline pipelines instead, *linear regression* is chosen for the *office* and *lodging/residential* categories while *extra trees regressor* is selected for all the other categories. The *extra trees regressor* algorithm has also shown an outstanding performance in many energy-related applications [71–73]. Through implementing extreme randomization without the need for exhaustive search at each decision tree node, *extra trees regressor* has been demonstrated to deliver superior performance compared to other tree-based algorithms, including *random forest* [74]. The additional level of randomness is introduced by creating decision trees based on random subsets of features and introducing random thresholds. This makes *extra trees regressor* less prone to over-fitting, which is essential when the dataset is noisy or has high-dimensional input features. The superior performance of this algorithm can thus be attributed to the building load data having noise and complex relationships between input features and output.

The result of applying the selected pipelines on the test dataset are presented in Table 3. It can be observed that using the identified optimal pipelines with a sliding window training scheme for the test subset achieves a performance that is in a similar range as the one that was obtained for the validation subset (for the case of the *office* category even an improvement is observed and the R^2 score is enhanced from of

Table 3
Test results obtained using the best pipeline with the online and offline approaches.

Category	Pipeline	R^2	MAPE [%]	CV (RMSE) [%]	NMBE [%]
Sliding window training					
Education	ETR - 6 weeks	0.936 (std = 0.051)	4.35 (std = 3.62)	6.79 (std = 3.60)	0.20 (std = 0.30)
Office	ETR - 6 weeks	0.937 (std = 0.074)	5.80 (std = 4.63)	11.74 (std = 16.13)	0.15 (std = 0.26)
Lodging/residential	ETR - 5 weeks	0.872 (std = 0.075)	4.31 (std = 2.94)	7.19 (std = 4.06)	0.05 (std = 0.40)
Entertainment/public assembly	ETR - 5 weeks	0.913 (std = 0.039)	5.41 (std = 2.58)	8.75 (std = 3.15)	0.16 (std = 0.39)
Public services	ETR - 6 weeks	0.948 (std = 0.021)	5.84 (std = 3.00)	7.88 (std = 3.41)	0.40 (std = 0.19)
Offline training					
Education	ETR	0.796 (std = 0.272)	8.19 (std = 6.08)	11.65 (std = 8.63)	0.37 (std = 6.18)
Office	LR	0.833 (std = 0.249)	12.81 (std = 13.43)	16.97 (std = 16.31)	-2.62 (std = 5.94)
Lodging/residential	LR	0.822 (std = 0.138)	5.64 (std = 2.97)	7.87 (std = 4.22)	-0.28 (std = 0.41)
Entertainment/public assembly	ETR	0.866 (std = 0.063)	7.58 (std = 2.94)	10.69 (std = 3.34)	-0.16 (std = 3.38)
Public services	ETR	0.923 (std = 0.026)	8.13 (std = 4.68)	9.87 (std = 4.59)	-0.17 (std = 0.66)

0.895 to 0.937). Furthermore, the calculated standard deviation demonstrates that the accuracy obtained for different buildings belonging to each category is also in a similar range (and close to the corresponding determined average). Comparing the performance of the offline pipelines over the test subset with the one obtained for the validation set shows a reduction in the R^2 score (except for the *public services* category, where an improvement is observed). Furthermore, the CV(RMSE) values are lower than the threshold specified by ASHRAE (30%) and IPMVP (20%). However, there is a significant performance improvement when using the sliding window training approach compared to the offline method. This result indicates that the proposed pipelines can effectively generalize the shape of the load curve. Similarly, the pipelines exhibit a remarkably low level of uncertainty concerning NMBE, falling within the recommended range of ±10% as suggested by ASHRAE Guideline 14, indicating an accurate and reliable representation of load prediction throughout the entire period, minimizing the overall occurrence of over or under predictions. Moreover, similar to the validation subset, the accuracy offered by the pipelines with the sliding window training scheme is higher than the one that can be achieved using offline training.

It is thus demonstrated that the identified pipelines with sliding window training scheme (that were shown to have superior performance compared to the ones with offline training over the validation set) are not prone to over-fitting. It is also revealed that the re-training procedure (with recent data) that is implemented in these pipelines permits achieving an elevated performance also for the unseen (test) data (while a decrement in performance is commonly observed utilizing offline training). These pipelines are thus employed for hour-ahead prediction of the (simulated) base load for the buildings during flexibility measures, with the results presented in the following subsection.

5.2. Results and discussions on the estimation of offered flexibility

As described in section 3.2, for all of the considered building models, two simulation scenarios were considered: i) the baseline consumption of the building with the regular (default) setpoint schedules ii) the flexible consumption while undergoing a setpoint modification procedure in the context of a demand response program. Fig. 7 illustrates an example of the indoor temperature with two thermal zones: Zone 0 (760 [m^3]) with a west-facing wall and Zone 1 (1260 [m^3]) with north and east-facing walls. Additionally, Fig. 8 displays the total electrical consumption profile of the building while it interacts with the grid. The increase in setpoint at 16:00 is evident in both figures, where evading the conditioning load (which in turn results in a rise in the temperature) leads to a reduction in the total load of the building. The corresponding baseline (default consumption profile without undergoing setpoint modification) is also demonstrated in Fig. 8. The difference between the baseline load and the flexible load in 8 represents the flexibility offered by the building.

However, the grid management authority does not commonly have access to the baseline load in order to quantify the offered flexibility accurately. Accordingly, the prediction pipelines, developed in the previous part for different building categories, are utilized to estimate the building's baseline load (simulated using the physics-based energy behavior models). In this context, the difference between the predicted and the simulated (real) baseline load results in an error in the estimated flexibility (compared to the simulated one), which should be quantified. In other words, the baseline load prediction error is propagated to the estimation accuracy of the offered flexibility (represented by the *flexibility index*), the extent of which should be assessed. The *flexibility index*, in each day, is first measured employing the simulation results (where both flexible and base load are simulated), and the corresponding estimated flexibility index is instead determined using simulated flexible load and the predicted baseline load for the sliding window and offline training. The difference between the simulated and estimated flexibility index is consequently quantified for each day. The resulting average error (presented using R^2, *MAPE*, and *MRD* scores as accuracy metrics) for all of the considered buildings is represented in Table 4. A strong agreement can be observed between the average values of simulated and estimated *FI* for the sliding window scheme, which is also confirmed by the resulting elevated R^2 scores and low (MAPE) values. Overall, the application of the sliding window training scheme has yielded better results, considering all of the used metrics, compared to the offline version. In particular, the Lodging/Residential and Public Services categories show notably lower R^2 and higher MAPE achieved for the offline approach compared to the sliding-window one. These results demonstrate the accuracy and consistency of the prediction pipelines developed in the previous part, which permit estimating the flexibility offered by buildings belonging to different categories. Moreover, the low values acquired for *MRD* scores for sliding-window training also demonstrate that there is no tendency to overestimate or underestimate the *flexibility index*, helping to minimize the unjustified global penalization or overcompensation of the user. As observed in section 5.1, the achieved CV(RMSE) for both sliding-window and offline training approaches remains below the designated threshold. However, the sliding window scheme exhibits comparatively lower values, particularly in Lodging/residential, Entertainment/public assembly, and Public services. The incorporation of NMBE confirms that all categories utilizing the sliding window scheme lie within the error range specified by ASHRAE Guideline 14 and IPMVP. Nonetheless, in the case of offline training, Lodging/residential and Public services fall outside the range of ±5% proposed by IPMVP.

In order to further assess the accuracy of the estimated *flexibility index* in each building model, the distribution of the relative error for the sliding-window training approach (which is due to the propagation of the corresponding baseline load prediction error) on different days is demonstrated in the boxplot represented in Fig. 9. Firstly, the symmetrical distribution of the error can be observed in all cases, with the 50th percentile located near zero with a small box size indicating that

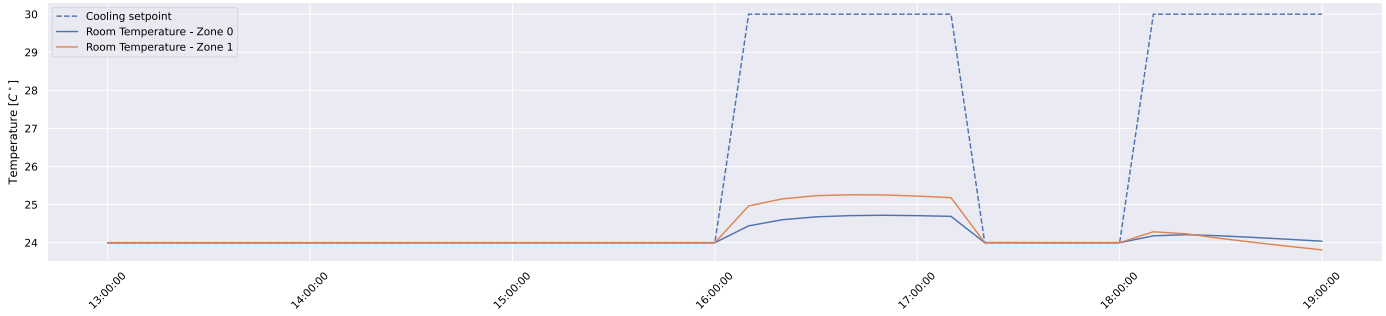


Fig. 7. Change in temperature profile in different zones when flexibility strategies are applied in a physical-based model.

Fig. 8. Change in load when flexibility strategies are applied in a physical-based model. The baseline load and the load while undergoing a flexibility measure are both provided by the energy simulation in EnergyPlus.

Table 4
Average *flexibility index* results and metrics accuracy obtained from the comparison between the simulated case and the predicted case using the best ML pipelines for the baseline prediction in year 2017.

Model Category	Average Flexibility Index [%]		R^2	MAPE [%]	MRD [%]	CV (RMSE) [%]	NMBE [%]
	Simulation	Estimation					
Sliding window training							
Educational	13.79	13.25	0.92	5.54	3.83	7.59	3.90
Office	9.79	9.74	0.97	3.32	1.39	6.22	0.50
Lodging/residential	13.89	14.22	0.91	2.86	−2.44	3.75	−2.40
Entertainment/public assembly	6.86	6.92	0.96	1.75	−0.72	2.21	−0.81
Public services	13.28	12.97	0.98	2.79	1.98	5.66	2.36
Offline training							
Educational	13.79	14.45	0.91	5.64	−4.19	8.00	−4.77
Office	9.79	10.03	0.93	3.45	−1.61	9.47	−2.48
Lodging/residential	13.89	16.02	−0.62	15.40	−15.40	15.97	−15.34
Entertainment/public assembly	6.86	7.01	0.93	2.47	−2.18	3.09	−2.16
Public services	13.28	14.00	0.86	7.52	−4.29	14.12	−5.42

the 25th percentile and 75th percentile of error values are clustered tightly around the median. In addition, the interquartile range of errors is limited to values less than 5% (except for the education category model with a 7.0% relative error), displaying the accuracy of the models. Moreover, whiskers are shown to be not long, which indicates there are no extreme outliers which further confirms the consistency of the predictions.

6. Conclusion

The current work introduced a data-driven approach, implementing machine learning-based pipelines to predict baseline consumption of buildings whilst they are providing flexibility in a summer cooling scenario. A dataset including the consumption profiles of 99 buildings belonging to different categories of services was accordingly utilized. A collection of five machine learning algorithms and linear regression, considering both offline and sliding window training schemes (with five different training window sizes), were assessed to benchmark a machine Learning-based pipeline for each one of the considered buildings categories. The performance of each pipeline was evaluated while keeping the coefficient of determination (R^2) score considered as the primary evaluation metric. Next, the most promising offline and sliding window pipeline (and the corresponding optimal training window) leading to the highest accuracy was identified for each building category.

It was revealed that using the extra trees regression algorithm with sliding window training outperforms other pipelines, and the optimal size of the training window for different categories of buildings was shown to be in the range of 4 to 6 weeks (with slight improvement be-

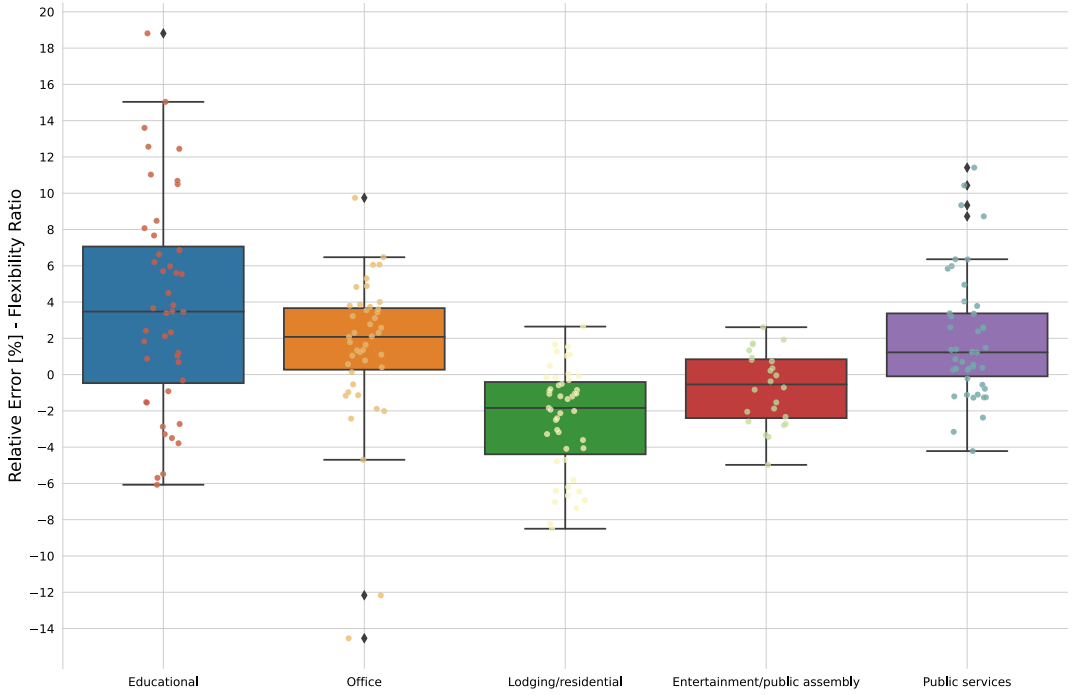


Fig. 9. Relative error distribution for the *Flexibility Index* obtained in calculating the available flexibility with the simulated baseline consumption and the predicted baseline consumption using the selected pipelines for the online case.

yond two weeks of training). Exerting the discovered pipelines with a sliding window training scheme on the test dataset achieved a similar or slightly better performance than that of the validation subset, indicating that the over-fitting issue has been avoided. Additionally, the calculated low values of standard deviation demonstrated that the obtained performance for different buildings belonging to each category is in a similar range.

The absence of a baseline load during the flexibility measures impedes the accurate calculation of the offered flexibility. Therefore, the second part of the study was dedicated to the deployment of the identified ML-based pipelines to predict the baseline load consumption of model buildings belonging to five different categories at the time the building provides flexibility. Therefore physics-based simulations (using the *EnergyPlus* software) were performed to model the baseline load and the flexible consumption during a setpoint modification procedure for a demand response program. The scenario of building offering flexibility was simulated by increasing the cooling setpoints for an interval of 1 hour on different days while ensuring that the temperature did not rise by more than 2 degrees, leading to the deactivation of the HVAC system and a consequent reduction of the load. Next, the proposed pipelines were employed for baseline load prediction, and the estimated flexibility index was determined using the predicted baseline and the simulated consumption while undergoing the flexibility measure. By comparing the determined estimated FI and the corresponding simulated values, an elevated agreement with a low MRD score (values between -2.45% to +2.79%), was observed, indicating the accuracy of the prediction that permit estimating the offered FI with acceptable accuracy. Finally, for all of the considered building models, the symmetrical distribution of the FI estimation's relative error on different days, with the 50th percentile close to zero, further confirmed the fact that using the proposed pipelines evades unjustified global penalization or overcompensation of the user. Consequently, the proposed set of pipelines was demonstrated to be an effective tool for predicting the baseline of different categories of buildings in order to achieve an accurate estimation of the offered flexibility index.

CRediT authorship contribution statement

I.A. Campodonico Avendano and F. Dadras Javan equally contributed to this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The utilized data is already open access.

References

[1] T. Kåberger, Progress of renewable electricity replacing fossil fuels, Global Energy Interconnect. 1 (1) (2018) 48–52, https://doi.org/10.14171/j.2096-5117.gei.2018.01.006, https://www.sciencedirect.com/science/article/pii/S2096511718300069.

[2] G.M. Tina, S. Aneli, A. Gagliano, Technical and economic analysis of the provision of ancillary services through the flexibility of hvac system in shopping centers, Energy 258 (2022) 124860, https://doi.org/10.1016/j.energy.2022.124860, https://www.sciencedirect.com/science/article/pii/S0360544222017637.

[3] IRENA, IRENA renewable energy capacity statistics 2015, http://www.irena.org/DocumentDownloads/Publications/IRENA-RE-Capacity-Statistics-2015.pdf, 2015.

[4] H. Li, Z. Wang, T. Hong, M.A. Piette, Energy flexibility of residential buildings: a systematic review of characterization and quantification methods and applications, Adv. Appl. Energy 3 (2021) 100054.

[5] F. Pallonetto, M. De Rosa, F. Milano, D.P. Finn, Demand response algorithms for smart-grid ready residential buildings using machine learning models, Appl. Energy 239 (2019) 1265–1282.

[6] A. Andreotti, G. Carpinelli, D. Lauria, V. Calderaro, V. Galdi, A. Piccolo, A tool for smart grid representation in presence of res: an application to state estimation problem, in: 2016 International Symposium on Power Electronics, Electrical Drives, Automation and Motion, SPEEDAM, IEEE, 2016, pp. 276–281.

[7] Ten questions concerning energy flexibility in buildings, Build. Environ. 223 (2022) 109461, https://doi.org/10.1016/j.buildenv.2022.109461, https://www.sciencedirect.com/science/article/pii/S0360132322006928.

[8] P.D. Lund, J. Lindgren, J. Mikkola, J. Salpakari, Review of energy system flexibility measures to enable high levels of variable renewable electricity, Renew. Sustain. Energy Rev. 45 (2015) 785–807.

[9] IEA, The critical role of buildings, https://www.iea.org/reports/the-critical-role-of-buildings, 2019.
[10] M. González-Torres, L. Pérez-Lombard, J.F. Coronel, I.R. Maestre, D. Yan, A review on buildings energy information: trends, end-uses, fuels and drivers, Energy Rep. 8 (2022) 626–637, https://doi.org/10.1016/j.egyr.2021.11.280, https://www.sciencedirect.com/science/article/pii/S235248472101427X.
[11] Load Management: Model-Based Control of Aggregate Power for Populations of Thermostatically Controlled Loads, Energy Convers. Manag. 55 (2012) 36–48.
[12] Jensen S. Østergaard, A. Marszal-Pomianowska, R. Lollini, W. Pasut, A. Knotzer, P. Engelmann, et al., Iea ebc annex 67 energy flexible buildings, Energy Build. 155 (2017) 25–34, https://doi.org/10.1016/j.enbuild.2017.08.044, https://www.sciencedirect.com/science/article/pii/S0378778817317024.
[13] M. Neukomm, V. Nubbe, R. Fares, Grid-Interactive Efficient Buildings, US Dept. of Energy (USDOE)/Navigant Consulting, Inc., Washington, DC (United States)/Chicago, IL (United States), 2019.
[14] K. Aduda, T. Labeodan, W. Zeiler, G. Boxem, Demand side flexibility coordination in office buildings: a framework and case study application, Sustain. Cities Soc. 29 (2017) 139–158.
[15] S. Mohagheghi, J. Stoupis, Z. Wang, Z. Li, H. Kazemzadeh, Demand response architecture: integration into the distribution management system, in: 2010 First IEEE International Conference on Smart Grid Communications, IEEE, 2010, pp. 501–506.
[16] X. Xue, S. Wang, C. Yan, B. Cui, A fast chiller power demand response control strategy for buildings connected to smart grid, Appl. Energy 137 (2015) 77–87.
[17] P. Pinson, H. Madsen, et al., Benefits and challenges of electrical demand response: a critical review, Renew. Sustain. Energy Rev. 39 (2014) 686–699.
[18] E. Commission, D.G. for Energy, C. Alaton, F. Tounquet, Benchmarking smart metering deployment in the EU-28: final report, Publications Office, 2020.
[19] F. Tahersima, P.P. Madsen, P. Andersen, An intuitive definition of demand flexibility in direct load control, in: 2013 IEEE International Conference on Control Applications, CCA, 2013, pp. 521–526.
[20] F. Dadras Javan, H. Khatam Bolouri Sangjoeei, B. Najafi, A. Haghighat Mamaghani, F. Rinaldi, Application of Machine Learning in Occupant and Indoor Environment Behavior Modeling: Sensors, Methods, and Algorithms, Springer, ISBN 978-3-030-72322-4, 2022.
[21] H.T. Haider, O.H. See, W. Elmenreich, A review of residential demand response of smart grid, Renew. Sustain. Energy Rev. 59 (2016) 166–178, https://doi.org/10.1016/j.rser.2016.01.016, https://www.sciencedirect.com/science/article/pii/S1364032116000447.
[22] R.G. Junker, A.G. Azar, R.A. Lopes, K.B. Lindberg, G. Reynders, R. Relan, et al., Characterizing the energy flexibility of buildings and districts, Appl. Energy 225 (2018) 175–182, https://doi.org/10.1016/j.apenergy.2018.05.037, https://www.sciencedirect.com/science/article/pii/S030626191830730X.
[23] V.M. Nik, A. Moazami, Using collective intelligence to enhance demand flexibility and climate resilience in urban areas, Appl. Energy 281 (2021) 116106, https://doi.org/10.1016/j.apenergy.2020.116106, https://www.sciencedirect.com/science/article/pii/S0306261920315270.
[24] F. D'Ettorre, M. Banaei, R. Ebrahimy, S.A. Pourmousavi, E. Blomgren, J. Kowalski, et al., Exploiting demand-side flexibility: state-of-the-art, open issues and social perspective, Renew. Sustain. Energy Rev. 165 (2022) 112605, https://doi.org/10.1016/j.rser.2022.112605, https://www.sciencedirect.com/science/article/pii/S1364032122005007.
[25] X. Ge, F. Xu, Y. Wang, H. Li, F. Wang, J. Hu, et al., Spatio-temporal two-dimensions data based customer baseline load estimation approach using lasso regression, IEEE Trans. Ind. Appl. 58 (3) (2022) 3112–3122.
[26] F. Lu, Z. Yu, Y. Zou, X. Yang, Energy flexibility assessment of a zero-energy office building with building thermal mass in short-term demand-side management, J. Build. Eng. 50 (2022) 104214, https://doi.org/10.1016/j.jobe.2022.104214, https://www.sciencedirect.com/science/article/pii/S2352710222002273.
[27] H. Sha, P. Xu, M. Lin, C. Peng, Q. Dou, Development of a multi-granularity energy forecasting toolkit for demand response baseline calculation, Appl. Energy 289 (2021) 116652.
[28] Z. Xuan, X. Gao, K. Li, F. Wang, X. Ge, Y. Hou, Pv-load decoupling based demand response baseline load estimation approach for residential customer with distributed pv system, IEEE Trans. Ind. Appl. 56 (6) (2020) 6128–6137.
[29] C. Lake, PJM Empirical Analysis of Demand Response Baseline Methods, 2011.
[30] T.K. Wijaya, M. Vasirani, K. Aberer, When bias matters: an economic assessment of demand response baselines for residential customers, IEEE Trans. Smart Grid 5 (4) (2014) 1755–1763.
[31] S. Mohajeryami, M. Doostan, A. Asadinejad, P. Schwarz, Error analysis of customer baseline load (CBL) calculation methods for residential customers, IEEE Trans. Ind. Appl. 53 (1) (2016) 5–14.
[32] F. Wang, K. Li, C. Liu, Z. Mi, M. Shafie-Khah, J.P. Catalão, Synchronous pattern matching principle-based residential demand response baseline estimation: mechanism analysis and approach description, IEEE Trans. Smart Grid 9 (6) (2018) 6972–6985.
[33] K. Li, J. Yan, L. Hu, F. Wang, N. Zhang, Two-stage decoupled estimation approach of aggregated baseline load under high penetration of behind-the-meter PV system, IEEE Trans. Smart Grid 12 (6) (2021) 4876–4885.
[34] Y. Chen, P. Xu, Y. Chu, W. Li, Y. Wu, L. Ni, et al., Short-term electrical load forecasting using the support vector regression (SVR) model to calculate the demand response baseline for office buildings, Appl. Energy 195 (2017) 659–670.
[35] A. Estebsari, R. Rajabi, Single residential load forecasting using deep learning and image encoding techniques, Electronics 9 (1) (2020) 68.
[36] R. Khalid, N. Javaid, F.A. Al-Zahrani, K. Aurangzeb, E.u.H. Qazi, T. Ashfaq, Electricity load and price forecasting using Jaya-long short term memory (JLSTM) in smart grids, Entropy 22 (1) (2019) 10.
[37] L. Fan, J. Li, X.P. Zhang, Load prediction methods using machine learning for home energy management systems based on human behavior patterns recognition, CSEE J. Power Energy Syst. 6 (3) (2020) 563–571.
[38] J. Huang, M. Algahtani, S. Kaewunruen, Energy forecasting in a public building: a benchmarking analysis on long short-term memory (LSTM), support vector regression (SVR), and extreme gradient boosting (XGBoost) networks, Appl. Sci. 12 (19) (2022) 9788.
[39] T. Cerquitelli, G. Malnati, D. Apiletti, Exploiting scalable machine-learning distributed frameworks to forecast power consumption of buildings, Energies 12 (15) (2019) 2933.
[40] R. Wang, S. Lu, Q. Li, Multi-criteria comprehensive study on predictive algorithm of hourly heating energy consumption for residential buildings, Sustain. Cities Soc. 49 (2019) 101623.
[41] A.A. Al-Shargabi, A. Almhafdy, D.M. Ibrahim, M. Alghieth, F. Chiclana, Buildings' energy consumption prediction models based on buildings' characteristics: research trends, taxonomy, and performance measures, J. Build. Eng. 54 (2022) 104577.
[42] Advanced metering infrastructure and customer systems, Tech. Rep., U.S. Department of Energy, 2016.
[43] A. Emery, D. Erne, E. Lyon, Demand side grid support program: Proposed draft guidelines first edition, Tech. Rep., California Energy Commission, 2022, Publication Number: CEC-300-2022-008.
[44] C. Miller, A. Kathirgamanathan, B. Picchetti, P. Arjunan, J.Y. Park, Z. Nagy, et al., The building data genome project 2, energy meter data from the ASHRAE great energy predictor III competition, Sci. Data 7 (2020) 368.
[45] W.S. Chandler, J.M. Hoell, D. Westberg, T. Zhang, P.W. Stackhouse Jr., Nasa prediction of worldwide energy resource high resolution meteorology data for sustainable building design, Tech. Rep., 2013.
[46] D.B. Crawley, L.K. Lawrie, F.C. Winkelmann, W.F. Buhl, Y.J. Huang, C.O. Pedersen, et al., Energyplus: creating a new-generation building energy simulation program, Energy Build. 33 (4) (2001) 319–331.
[47] Input output reference, Energyplus version 9.6.0 documentation, Tech. Rep., U.S. Department of Energy, 2021.
[48] ANSE/ASHRAE/IES standard 90.1-2010, Energy standard for buildings except low rise residential buildings, Tech. Rep., American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Atlanta, Georgia, 2010.
[49] M. Deru, K. Field, D. Studer, K. Benne, B. Griffith, P. Torcellini, et al., US Department of Energy Commercial Reference Building Models of the National Building Stock, 2011.
[50] A.D. Smith, B. Stürmer, T. Thurber, C.R. Vernon, diyepw: a Python package for do-it-yourself energyPlus weather file generation, J. Open Sour. Softw. 6 (64) (2021) 3313, https://doi.org/10.21105/joss.03313.
[51] B. Ratner, The correlation coefficient: its values range between +1/−1, or do they?, J. Target. Meas. Anal. Mark. 17 (2) (2009) 139–142.
[52] R. Kissell, Chapter 6 - Price volatility, in: R. Kissell (Ed.), The Science of Algorithmic Trading and Portfolio Management, Academic Press, San Diego, ISBN 978-0-12-401689-7, 2014, p. 211, https://www.sciencedirect.com/science/article/pii/B9780124016897000064.
[53] S. Kasemsumran, Y.P. Du, B.Y. Li, K. Maruo, Y. Ozaki, Moving window cross validation: a new cross validation method for the selection of a rational number of components in a partial least squares calibration model, Analyst 131 (2006) 529–537, https://doi.org/10.1039/B515637H.
[54] C.J. Eubel, A reinforcement learning characterization of thermostatic control for hvac demand response and experimentation framework for simulated building energy control, Ph.D. thesis, The Ohio State University, 2022.
[55] K. Pearson , VII. Note on regression and inheritance in the case of two parents, Proc. R. Soc. Lond. 58 (347-352) (1895) 240–242.
[56] N.S. Tzannes, J.P. Noonan, The mutual information principle and applications, Inf. Control 22 (1) (1973) 1–12, https://doi.org/10.1016/S0019-9958(73)90448-8, https://www.sciencedirect.com/science/article/pii/S0019995873904488.
[57] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (2004) 066138, https://doi.org/10.1103/PhysRevE.69.066138, https://link.aps.org/doi/10.1103/PhysRevE.69.066138.
[58] P. Laarne, M.A. Zaidan, T. Nieminen, ennemi: non-linear correlation detection with mutual information, SoftwareX 14 (2021) 100686, https://doi.org/10.1016/j.softx.2021.100686, https://www.sciencedirect.com/science/article/pii/S2352711021000315.
[59] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., SciKit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[60] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[61] X. Su, X. Yan, C.L. Tsai, Linear regression, Wiley Interdiscip. Rev.: Comput. Stat. 4 (3) (2012) 275–294.
[62] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, IEEE, 1995, pp. 278–282.

[63] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[64] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.
[65] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, in: M. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, vol. 9, MIT Press, 1996, https://proceedings.neurips.cc/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.
[66] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1) (1967) 21–27, https://doi.org/10.1109/TIT.1967.1053964.
[67] A.G. Ashrae, Measurement of energy and demand savings, ASHRAE Trans. 35 (2002) 41–63.
[68] A.G. Ashrae, Measurement of energy, demand, and water savings, ASHRAE Trans. 4 (2014) 1–150.
[69] G. Ramos Ruiz, C. Fernandez Bandera, Validation of calibrated energy models: common errors, Energies 10 (10) (2017) 1587.
[70] J. Cowan, International performance measurement and verification protocol: concepts and options for determining energy and water savings-vol. I, International Performance Measurement & Verification Protocol, 2002, p. 1.
[71] M.W. Ahmad, J. Reynolds, Y. Rezgui, Predictive modelling for solar thermal energy systems: a comparison of support vector regression, random forest, extra trees and regression trees, J. Clean. Prod. 203 (2018) 810–821, https://doi.org/10.1016/j.jclepro.2018.08.207, https://www.sciencedirect.com/science/article/pii/S0959652618325551.
[72] A. Liu, F. Li, J. Li, B. Qian, Y. Che, M. Zhou, A reconstruction method for missing data of electricity users using extremely randomized tree, in: 2021 International Conference on Computer, Internet of Things and Control Engineering, CITCE, 2021, pp. 74–77.
[73] M. Gong, J. Wang, Y. Bai, B. Li, L. Zhang, Heat load prediction of residential buildings based on discrete wavelet transform and tree-based ensemble learning, J. Build. Eng. 32 (2020) 101455, https://doi.org/10.1016/j.jobe.2020.101455, https://www.sciencedirect.com/science/article/pii/S2352710219324799.
[74] L. Zhang, Y. Ren, P.N. Suganthan, Towards generating random forests via extremely randomized trees, in: 2014 International Joint Conference on Neural Networks, IJCNN, 2014, pp. 2645–2652.