Applied Energy 280 (2020) 115981

Contents lists available at ScienceDirect

Applied Energy

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

SynCity: Using open data to create a synthetic city of hourly building energy estimates by integrating data-driven and physics-based methods

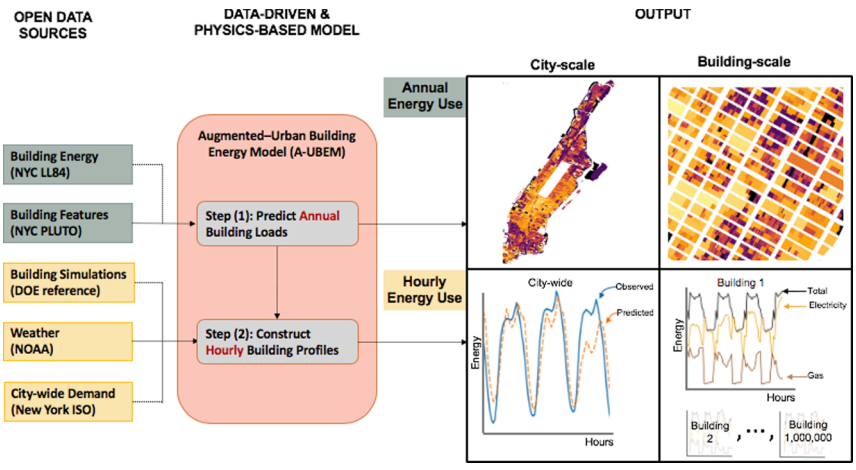
Jonathan Roth [a,c,*], Amory Martin [b], Clayton Miller [c], Rishee K. Jain [a]

[a] Urban Informatics Lab, Department of Civil & Environmental Engineering, Stanford University, United States
[b] Blume Earthquake Engineering Center, Department of Civil & Environmental Engineering, Stanford University, United States
[c] Building and Urban Data Science (BUDS) Lab, Dept. of Building, School of Design and Environment (SDE), National University of Singapore (NUS), Singapore

HIGHLIGHTS

- A-UBEM generates synthetic smart meter data for all buildings in a city.
- We integrate data-driven and physics-based simulation methods to create A-UBEM.
- A-UBEM is built using solely open-data sources, with New York City as a case study.
- We validate our model using Monte Carlo simulations and city-wide hourly load.
- We highlight applications where A-UBEM can be used to plan sustainable cities.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Urban building energy model
Supervised machine learning
Convex optimization
Smart meter
Energy efficiency
Energy prediction

ABSTRACT

Cities officials are increasingly interested in understanding spatial and temporal energy patterns of the built environment to facilitate their city's transition to a low-carbon future. In this paper, a new Augmented-Urban Building Energy Model (A-UBEM) is proposed that combines data-driven and physics-based simulation methods to produce synthetic hourly load curve estimates for every building within a city—similar to data an hourly smart meter would measure. By using only publicly available data, a generalizable *two-step* process is implemented—that other cities with similar available data can replicate—using New York City as a case study. Step (1) estimates the annual energy use for every building in the city using supervised machine learning algorithms. Step (2) extends these results and leverages physics-based simulation models through a convex optimization formulation that minimizes the squared difference between the aggregated building demand and the observed city-wide hourly electricity demand. Results from step (1) show that the Random Forest algorithm performs best with a mean log squared error of 0.293, while the convex optimization in step (2) results in a mean training error of 6.11% mean absolute percentage error (MAPE). To validate the stability of the produced load

* Corresponding author at: Urban Informatics Lab, Department of Civil & Environmental Engineering, Stanford University, United States.
E-mail address: jmroth@stanford.edu (J. Roth).

https://doi.org/10.1016/j.apenergy.2020.115981
Received 11 May 2020; Received in revised form 21 September 2020; Accepted 2 October 2020
Available online 10 October 2020
0306-2619/© 2020 Elsevier Ltd. All rights reserved.

curves, Monte Carlo simulations are conducted, using random subsets of buildings from the city, which produce an out-of-sample error averaging 6.41% MAPE across each simulation. Particle swarm optimization is also explored—using the results from the Monte Carlo simulation—to assess if the model could be improved by relaxing certain constraints, but marginal error reductions are found, further proving the stability of the proposed model. Overall, A-UBEM is a first step towards creating highly granular urban-scale synthetic hourly load curves solely using open data. Such load curves are integral for planning sustainable cities and accelerating the adoption of low-carbon distributed energy resources (DERs) and district energy systems.

Nomenclature:

The following list of symbols, with their corresponding dimensions, are used in this paper:

$\boldsymbol{\alpha}$	Lasso: estimated coefficients from Lasso $[[N \times 1]]$
$\boldsymbol{\beta}$	Design variables, mapped weights for DOE $[[6I \times 1]]$
ε	SVM: insensitivity parameter
η	Gradient Boosting: Learning rate
λ	Lasso: penalization hyperparameter
$\varphi(\boldsymbol{x})$	SVM: kernel function
A	Annual time step$\{A = 1\}$
a	PSO: particle number
b	SVM: bias term
$\boldsymbol{C}$	Cooling degree hour for NYC
C_{reg}	SVM: regularization parameter
c_1	PSO: the cognitive learning factor $[[H \times 1]]$
c_2	PSO: the social learning factor
$\boldsymbol{D}$	DOE three reference buildings mapped to PLUTO building class i $[[H \times 1]]$
$\boldsymbol{E}$	Total city-wide hourly load (NYISO)$[H \times 1]$
$\widehat{\boldsymbol{E}}(\boldsymbol{\beta})$	Predicted city-wide hourly load (aggregated building load)$[H \times 1]$
G_l^{best}	PSO: best value obtained by any particle in iteration$l$
i	PLUTO building class$\{1, \cdots, i, \cdots, I = 25\}$
j	Building in randomly sampled NYC dataset$\{1, \cdots, j, \cdots, J = 1000\}$
k	Used to denote the three $\boldsymbol{\beta}$ parameters associated with$\boldsymbol{D}$
l	PSO: iteration number$\{1, \cdots, l, \cdots, L\}$
m	Number of observations (buildings)
m_{PLUTO}	The total number of buildings in the PLUTO dataset$\{1, \cdots, m, \cdots, M_{PLUTO} = 1,000,000\}$
m_{ll84}	Total number of buildings in LL84 dataset$\{1, \cdots, m, \cdots, M_{LL84} = 15,000\}$
n	Building characteristics (e.g., area, type)$\{1, \cdots, n, \cdots, N = 38\}$
$\boldsymbol{P}$	Annual energy use for J buildings from step (Section 4.1) $[[J \times 1]]$
$\overline{\boldsymbol{P}}$	Normalized annual energy use for J buildings from step $[[J \times 1]]$
P_l^{best}	PSO: best solution acquired by each particle in all the previous l iterations
r_1, r_2	PSO: two independent random uniform numbers
S	The total number of Monte Carlo simulations$\{1, \cdots, s, \cdots, S = 500\}$
$\boldsymbol{T}$	Temperature vector for NYC $[[H \times 1]]$
t	Hourly time step$\{1, \cdots, t, \cdots, T = 8784\}$
v_{l+1}^a	PSO: velocity of particle a at iteration$l$
$\boldsymbol{W}$	Weekend, business day and holiday vector $[[H \times 1]]$
$\boldsymbol{w}$	SVM: weight, or unit vector $[[H \times 1]]$
$w_{inertia}$	PSO: the inertia used to control the effect of the particles' velocity
$\boldsymbol{X}$	The building characteristics from PLUTO for those buildings appearing in the LL84 data $[[M_{ll84} \times N]]$
$\boldsymbol{x}_{m_{LL84}}$	One observation from matrix $\mathbf{X}$ $\boldsymbol{[[N \times 1]]}$
x_{l+1}^a	PSO: position of particle a at iteration$l$
$\boldsymbol{Y}_{ij}^t(\boldsymbol{\beta})$	Hourly energy demand for building j of PLUTO building class i $[[I \times J]]$
$\boldsymbol{y}$	The log transformation of annual building energy use (kBtu) $[[M_{ll84} \times 1]]$
$\widehat{\boldsymbol{y}}$	The estimated log transformation of annual building energy use (kBtu $[[M_{ll84} \times 1]]$)
*SVM	support vector machines
*PSO	particle swarm optimization
*DOE	Department of Energy

1. Introduction

Driven by rapid growth in urban populations, cities are increasingly becoming the nexus of economic activity and civic engagement and, therefore, a driving force for environmental stewardship. Cities are also facing increased threats from fossil fuel induced climate change including deadly heatwaves—exacerbated by the urban heat island effect—and rising sea levels that threaten two-thirds of the world's major cities that are on the coast [1,2]. Furthermore, according to the World Bank, the burning of fossil fuels caused 5.5 million deaths in 2013 from air pollution, which accounts for 1 in 10 deaths globally, and is the third most import health risk leading to early death in low- and lower-middle-income countries [3]. Despite these costs, the International Monetary Fund estimates fossil fuel subsidies totaled $4.7 trillion in 2015 with $649 billion coming from the U.S. alone [4]. In response to these threats, coupled with rising energy prices, many cities are taking the lead and implementing new sustainability initiatives.

City officials are adopting long-term plans centered around electrification and new power generation from wind and solar to reduce emissions from combustion-based generation. Though renewables are curbing cities' dependence on fossil fuels, these non-dispatchable resources are also increasing fluctuations in electricity demand and adding uncertainty to energy markets [5]. In places with large numbers of solar panels, like California, the timing imbalance between solar production and peak demand has created volatile electricity prices and decreased grid reliability [6]. Furthermore, traditional power generation—like coal plants or utility-scale solar stations—are beginning to be supplanted by distributed energy resources (DERs) that decentralize power generation and storage through rooftop solar and behind-the-meter batteries. DERs therefore, paradoxically, pose potential solutions and issues for the grid. Depending on where DERs are installed, their generation and/or storage capabilities could either mitigate strain on the grid induced by periods of high demand or add instability by generating excess power when none is needed. Consequently, the future grid must cope with this changing energy landscape by finding solutions to enhance grid resilience, reduce electricity prices, and decrease or shift energy demand as required.

Beyond the need to better model the electricity grid, the increased adoption of DERs is also creating a strong need to understand spatial and temporal patterns of building energy use [7]. Buildings consume between 30 and 70% of total primary energy use in cities, but their energy consumption fluctuates over time and varies greatly between buildings [8]. As a result, many municipalities are focused on better understanding their energy usage patterns and finding ways to reduce or shift their demand through energy efficiency retrofits, new energy storage technologies, or access to district energy systems, among other solutions. A first step to many of these solutions is the city-wide installation of *smart meter* devices that can measure energy consumption at the sub-hourly level. These data are invaluable for the analysis of temporal energy use patterns; however, accessing this data from the utility is often restricted and has been a substantial barrier for their use in policy-making. One major reason why is that smart meter data are often the subject of concern for many building owners and tenants due to potential privacy abuses. But without this data, widespread adoption of DERs, targeted policymaking, and solutions to reduce or shift building energy demand will prove difficult.

One strategy that has proven effective to reduce building energy use is through energy benchmarking [9]. The goal of this practice is to identify inefficient buildings—with large opportunities for energy savings—by measuring and comparing the energy use of various types of buildings [10]. Cities have recently begun to understand the value of energy benchmarking; over 30 cities throughout the United States have passed ordinances requiring benchmarking and energy disclosure for a subset of their city's building stock [11]. In addition to achieving measurable energy reductions, these ordinances often require the annual release of a publicly available open dataset of energy usage data and characteristics of benchmarked buildings. These open datasets enable the public to track annual building energy performance over time [12], but perhaps more importantly, they allow this highly sought after data to be used more widely. This open data initiative empowers researchers and policy-makers—who are often in need of data—by allowing them to find new insights, create more collaborations between institutions and governments, and more easily build off other work that uses the same datasets [13–15].

Provided with these new open data sources, this paper aims to provide detailed spatial and temporal patterns of building energy use in cities to empower policymakers with the information they need to accelerate the energy transition. To achieve this goal, the research objectives for this paper are as follows:

- Produce hourly energy use estimates for each individual building within a city
- Combine physics-based simulations and data-driven techniques to leverage advantages offered by both types of methodologies
- Design a flexible and extensible model such that it can be applied to any city with open data available
- Validate the model using Monte Carlo simulations and city-wide hourly data

1.1. Novelty of approach

In this paper, we create a generalizable *two-step* Augmented-Urban Building Energy Model (A-UBEM) that produces hourly electricity, heating, and cooling profiles for all buildings within a city using only publicly available data. These synthetic hourly energy predictions at the building level are similar to the data that would be generated by *smart meters*. Using New York City (NYC) as a case study, we show the advantages of our model—combining data-driven techniques with physics-based simulations models—that other cities with similar public data can replicate. Although several previous studies have combined physics-based and data-driven techniques, this study is one of the first to do it at the urban-scale, using solely open data, and provide hourly energy estimates for each individual building in a city. In step (1) we construct a supervised machine learning model to predict the *annual* energy use values for every building in NYC using observed energy use values from a small subset of buildings. In step (2) we match each building in the city to three building archetypes from the Department of Energy's (DOE) reference building dataset. We then construct hourly loads for each building by fitting a weighted average of these three building types' load curves, adjusting for weather and weekday effects, and fit the model using city-wide electricity hourly demand from the New York Independent System Operator (NYISO). We run 500 Monte Carlo Simulations to validate the model by examining its stability by creating a distribution of building load profiles for each building type. By modeling hourly values of energy consumption, city officials and planners can more adequately evaluate energy system alternatives, such as district energy and distributed energy systems.

2. Literature review

Several studies have attempted to model urban energy consumption using top-down approaches—techniques that use external metrics, such as economic activity, to estimate energy consumption at the city-level. Dhakal examined urban energy consumption for several Chinese cities using energy values derived from economic activity. The study found that the 35 largest cities in China account for nearly 40% of national energy consumption [16]. Bentzen and Engsted also used economic indicators to examine the annual energy consumption for Denmark, finding that long-term energy consumption was strongly affected by income and previous years' consumption [17]. Brownsword et al., examined the savings effects of energy-management measures and associated reductions in $C0_2$ using a linear programming module based on energy supply data and zip code information [18]. These top-down models are useful for providing high-level estimates but rely on derived values from economic activity, instead of actual energy measurements, or overlook smaller scale variation from urban land use and the diversity of building typologies.

Urban building energy modeling (UBEM) is an emerging field with the goal of creating more detailed models of city energy consumption. UBEM primarily relies on new physics-based simulation techniques, which have typically focused on modeling individual buildings using geometric, construction, weather, and usage schedule data. One review study looked at the handful of bottom-up UBEM tools that have been created in the last ten years—from CitySim in 2009 to TEASER in 2018—and identified a number of limitations that continue to persist. For example, the study named several key components for further research: data availability, people movements within cities, district energy modeling, microclimate effects, heat exchange between buildings, validation, and life cycle assessment. Since constructing a physics-based model for one building can be quite time intensive, extending it to the urban scale is computationally challenging [19]. To circumvent these computational challenges, one study simulated a subset of 47 buildings within a city and used them as archetypal buildings for mapping the remainder of the city [20]. Some UBEMs require GIS shapefiles to extract the footprint of every building within a city, which are ideally combined with building heights or LiDAR, to produce a more holistic model of each building [21]. This modeling approach depends on uncommon city data, plus it can fail to account for the urban context, neglecting aspects like shading, local wind patterns, traffic flows, etc. To account for the urban heat island effect, specifically, Mavrogianni et al. integrated localized weather files instead of using the meteorological year profiles often used [22]. Going a step further, Xu et al. incorporated other aspects of urban morphology—such as building density, canyon aspect ratio, building height, and sky view factor—to assess changes in building energy demand from changing pavement albedo [23].

Despite these studies accounting for other building types and local contexts, simulation modeling still requires several assumptions to extend the results to the urban-scale. Furthermore, these types of

simulation-based energy models lack energy data for validation, leading to uncertainty in the produced results.

Instead of using physics-based simulation techniques, researchers have focused on alternative methods to model annual energy use at the building-level using data-driven techniques [24,25]. Previous studies have shown that machine learning models outperform simpler linear models for predicting total energy consumption [26]. Popular and effective models in this field have included support vector regression, random forests, gradient boosting regression, and artificial neural networks [27–31]. These models are able to capture non-linear effects of energy consumption in buildings. One recent Kaggle competition on building energy prediction even found that most winning models used gradient boosting [32]. The effectiveness of these methods, however, can vary depending on the level of aggregation present—performance increases with building aggregation, indicating that predicting many buildings is easier than one [28]. As such, researchers have begun to examine how these machine learning algorithms can be extended from individual buildings to the urban context [33]. In one study, geographically weighted regression was used to analyze determinants of water consumptions in New York City [34], while another study combined spatial analysis with neural networks to find connections between urban form and energy use [35]. Furthermore, researchers are also looking at long-term forecasting—rather than just hourly, monthly, or yearly—to predict district heating loads in cities which can be helpful for general energy planning purposes [36].

Using supervised machine learning models, like the ones previously discussed, are not a silver bullet for energy prediction, despite the recent rapid advancements in these tools. These models are often difficult to interpret and provide little physical explanatory power. Often times it is critical to understand what the driving mechanisms are for building energy use so that decision-makers—like facility managers, building owners, or policymakers—can make more informed decisions [37–39]. Unsupervised machine learning algorithms can also provide different types of insights into building energy performance by extracting patterns from datasets without the need for a target variable [40]. Clustering algorithms can be used to divide large groups of buildings into similar groups, while network analysis can be used to identify reference buildings from specific groupings [41,42]. The growing availability of data has also resulted in researchers exploring dimensionality reduction methods to reduce the size of data while maintaining as much information as possible in order to speed up computation [43].

Despite the growing availability of public open data sources, relatively few studies have focused on using these datasets in conjunction with data-driven methodologies, like statistical methods and machine learning algorithms, to predict urban building energy consumption [44]. But these datasets and data-driven techniques offer new opportunities for urban energy use modeling [45]. There are three key studies, which lay the groundwork for this research by examining building energy use for an entire city (New York City) using a data-driven approach, rather than a traditional physics-based UBEM. First, Howard et al. [24] presented a methodology to model building-level annual energy use intensities by downscaling zip code level energy data. Assuming that primary end use is solely dependent on building function and size, the study fits a linear model to the zip code level energy data to find building level energy intensities. The main limitation of this study is the lack of validation for individual building loads. Second, Robinson et al. [25], predicts building energy use for buildings in New York City by building a machine learning model based on a subset of building-specific energy usage data obtained from Local Law 84 (LL84). Gradient boosting regression is found to have the best results when validated at the building-level, though no validation was attempted at larger aggregation levels. Third, Kontokosta and Tull [46] also built a machine learning model using the LL84 dataset but validated it at both the building and zip-code level; results showed that the linear regression OLS model performed best at the zip-code level, while the support vector machine performed best at the building-level. Inspired by these recent efforts, this study aims to not only predict annual building loads but translate them into hourly profiles necessary for informing decision-making using supervised machine-learning and convex optimization. By using both data-driven and physics-based simulation methods, our proposed integrated model leverages the advantages from both types of modeling techniques to obtain higher resolution of energy consumption patterns [47]. Further, we purposefully construct our model using only open data sources to allow other cities with similar data to replicate our model and to enable researchers to more easily build off our work.

3. Data collection and pre-processing

The overarching objective of this study is to construct a generalizable Augmented-Urban Building Energy Model (A-UBEM) that produces synthetic hourly energy demand profiles for every building in New York City (NYC) using only publicly available open data; this allows other researchers and city-officials to replicate the methodology and apply it to other urban areas. The analysis consists of two primary steps as shown in Fig. 1: (1) constructing annual building-level energy estimates for all buildings in NYC; (2) converting annual energy loads into building-level hourly demand profiles. The first step uses supervised machine learning and historical annual energy consumption data from about 15,000 buildings, building off previous work [48]. The detailed methodology for this step is discussed in Section 4.1. The second step uses archetypal simulation-based models and a novel optimization algorithm to match the aggregated building load to the NYC electricity profile. The detailed methodology for this step is discussed in Section 4.2. This section summarizes the datasets used for both steps of A-UBEM using New York City as a case study.

We demonstrate this two-step framework for New York City due to its high energy load, number of buildings, and availability of public data, which has been recognized by other researchers in this domain [25,49]. One significant contribution of this study is to build a A-UBEM capable of estimating building specific hourly energy usage profiles using only publicly available data. We use a total of five different public datasets, summarized in Table 1, to first predict annual energy demand for all buildings in NYC, across all five boroughs, and second convert those predictions into hourly energy demand profiles by leveraging physics-based simulation models from the U.S. Department of Energy's OpenEI database. All the data for this project can be found on our GitHub page or the URLs provided in Table 1.

3.1. Step (1) Data: LL84 and PLUTO datasets

For step (1) of the framework, we combine annual energy consumption data with building characteristics to predict annual energy use for all buildings in NYC. This step requires two publicly available open datasets: (1) annual building-level energy use data for 15,000 buildings (*LL84*, dataset (a)); (2) building characteristics for all 1 million buildings in the city (*PLUTO,* dataset (b)).

In 2009, NYC passed Local Law 84 (LL84) which required all buildings over 50,000 square feet to benchmark the energy performance of their buildings and disclose their annual energy consumption. Every year a public version of this dataset is released on the NYC Open Data portal. This law was recently updated to cover buildings over 25,000 square feet though this data is not yet available. Other works have also used versions of this dataset for building energy analysis, as discussed in Section 2 [46]. For step (1) of the framework, we use the disclosed energy data for the 2016 calendar year which contains about 15,000 buildings (*LL84*, dataset (a)). We opted to use total building *energy* data since the disclosed *electricity* data contained more missing data and erroneous values. To prepare the data, total site energy use was computed for each building in the LL84 dataset by multiplying the site energy use intensity (EUI) by the building area. Outliers were identified by finding all points that were outside four times the interquartile range for site EUI and then removed. We performed this calculation of outliers

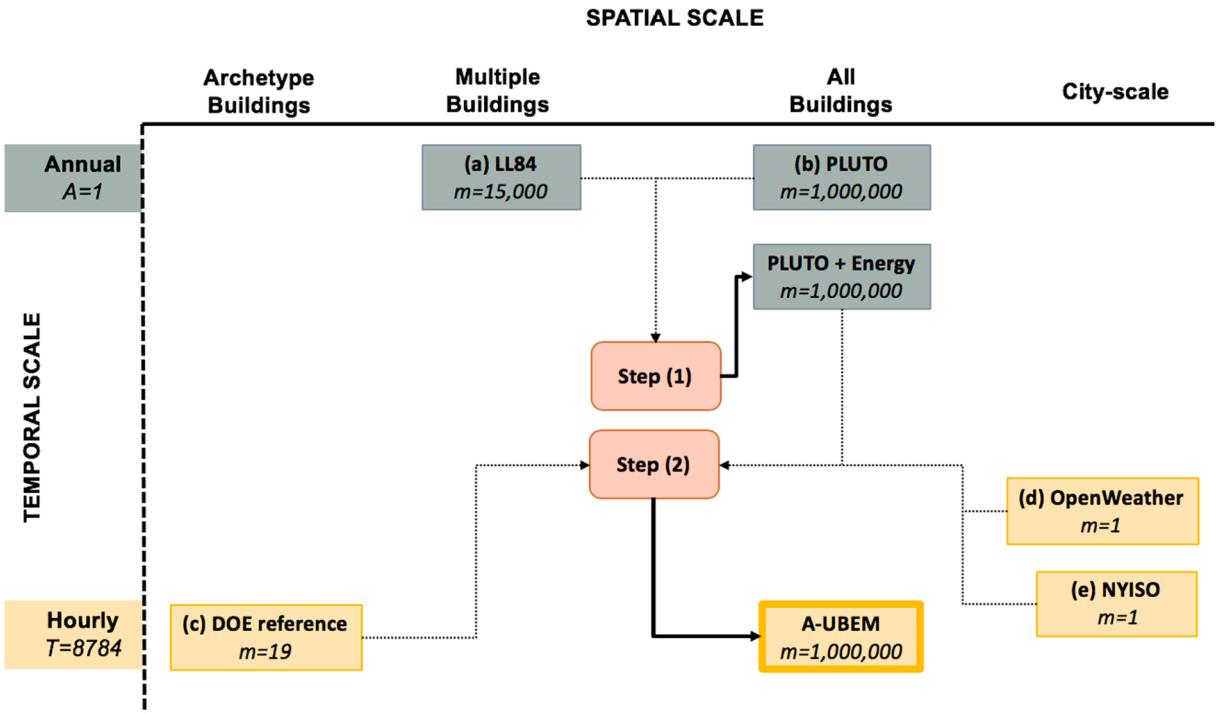


Fig. 1. Overview of proposed Augmented-Urban Building Energy Model (A-UBEM). Starting with annual energy load from a subset of the city building stock, step (1) trains a supervised machine learning model on these data and predicts the annual load for all 1 million buildings. Using this output, step (2) leverages simulation-based models and city-level utility energy consumption to estimate the hourly load for all buildings in the city.

Table 1
All five public datasets that are used to construct the Augmented-Urban Building Energy Model (A-UBEM) for New York City. Each of these datasets are available to download online.

Set	Public Dataset Name	Description	Step	Temporal Scale	Spatial Scale	URL
(a)	Local Law 84 (LL84)	Energy consumption for 15,000 buildings in NYC in 2016	Step (1): Machine Learning	Annual (A = 1)	Building-level (m = 15,000)	https://www1.nyc.gov/html/gbee/html/plan/ll84_scores.shtml
(b)	Primary Land Use Tax Lot Output (PLUTO)	Physical building characteristics for all 1 million buildings in NYC in 2016	Step (1): Machine Learning	Annual (A = 1)	Building-level (m = 1,000,000)	https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page
(c)	Department of Energy (DOE) commercial and residential reference buildings	Hourly energy demands for 19 building archetypes produced by physics-based simulations using EnergyStar software	Step (2): Optimization	Hourly (T = 8784)	Building-level (m = 19)	https://openei.org/doe-opendata/dataset/commercial-and-residential-hourly-load-profiles-for-all-tmy3-locations-in-the-united-states
(d)	OpenWeather	Hourly weather data experienced by NYC in 2016	Step (2): Optimization	Hourly (T = 8784)	City-level	https://openweathermap.org/history
(e)	New York Independent System Operator (NYISO)	Hourly electricity demand experienced by NYC in 2016	Step (2): Optimization	Hourly (T = 8784)	City-level	http://www.energyonline.com/Data/GenericData.aspx?DataId = 13

since our goal was to maintain as much data as possible to ensure that the following supervised machine learning model would generalize well to the entire city. Finally, we also removed any site EUI below an absolute value of one.

NYC's Primary Land Use Tax Lot Output (*PLUTO*, dataset (b)) dataset provides building information for the entire city, auxiliary information used in step (1) to predict annual building energy use. The PLUTO dataset contains information for every tax lot in NYC including basic physical characteristics for all 1 million buildings in the city across all five boroughs. We merge the PLUTO dataset with the LL84 dataset to provide a standardized feature space to train a model to predict annual building energy use for the buildings where public energy data is not available.

The combined LL84 and PLUTO datasets contain the whole feature space for the 15,000 buildings that are used to build our models in step (1), outlined in Section 3.2 (energy data from LL84, building characteristics from PLUTO). We therefore engineered several additional features to help our models obtain the best predictions possible. These additional engineered features were constructed to extract non-linear patterns that might go uncaptured if they were not created. We applied the logarithmic transformation to ten separate features, calculated fractions of floor space by use type, and appended them to the dataset. A summary of the $N = 38$ total features used for step (1) can be found in Appendix B in the supplementary document. We also imputed all missing values in the PLUTO dataset (note, that less than 1% of data was missing from features that are used for modeling). The MICE package in R was used, which generates multiple imputations for the incomplete data through Gibbs sampling [48]. We employed classification and regression tree methods due to their flexibility in handling missing data and ability to find non-linear relationships [50].

3.2. Step (2) Data: DOE reference buildings, NYISO electricity data, and OpenWeather

To extend the model to produce hourly loads for each building in NYC, we compiled three other publicly available open datasets: (1) archetypal hourly building loads produced from physics-based simulation models (*DOE Reference Buildings,* dataset (c)); (2) historical NYC hourly weather data (*OpenWeather,* dataset (d)); (3) historical hourly NYC electricity demand (*NYISO Electricity Data,* dataset (e)).

The archetypal hourly building loads from the Department of Energy (DOE) commercial and reference building dataset (*DOE Reference Buildings,* dataset (c)) shows hourly load profile data for 16 commercial and 2 residential building types (high and medium energy consumption households) for all TMY3 (typical meteorological year, version 3) locations in the United States. The 18 profiles collected were for Central Park TMY3. These profiles breakdown the hourly loads by total facility electricity use, electricity for heating, electricity for cooling, electricity for interior lighting, electricity for interior equipment, total facility gas use, gas for heating, gas for interior equipment, and gas for water heating. The DOE produces loads for these 18 buildings using physics-based simulation models—design and inputs to produce each building type can be found on the DOE website. Given the rapid construction of data centers around the country and a total forecasted U.S. load of 73 billion kWh in 2020 [51] we found it important to include this type of building, which is also known to have a very different energy use profile. We therefore added one more hand constructed profile to the reference, named "Datacenter", bringing the total number of reference buildings to 19. For simplicity, we model this profile as a flat energy curve because datacenters have consistent energy loads [51]. Finally, given that 2016 was a leap year, 24 h of data were appended to the DOE reference buildings dataset for February 29th. The added data took on the same values as the previous day's load.

Although the 19 reference buildings are modeled by the DOE using TMY3 data from Central Park, we use historical weather data from NYC *(OpenWeather,* dataset (d)) in 2016—to use for our step (2) modeling process—to obtain more accurate load curves; because weather is known to have a large impact on building energy consumption, more precise data will help our model achieve more accurate results. To acquire 2016 weather data from NYC, we used the API provided from OpenWeather to gather hourly temperature, humidity, and wind speed data for the entire year. This data is used to adjust the loads of the 19 reference buildings for observed weather patterns in our constructed model. For other cities with high weather discrepancies between locations, like San Francisco, weather data from multiple locations in the city could be used to enhance data quality.

Finally, in order to build and validate the output of our final model, we collected hourly electricity demand data for NYC from the New York Independent System Operator (*NYISO Electricity Data,* dataset (e)). This dataset provides historical electricity consumption data for zone J, which encompasses all of New York City. This data is used to construct and calibrate the Augmented-Urban Building Energy Model (A-UBEM). Several data points were missing for the NYISO hourly electricity load and were linearly interpolated based upon the nearest two hours of load.

4. Methodology

4.1. Step (1): Predicting annual building loads

With the goal to model building-level energy demand for each of the 1 million buildings in NYC, we restrict the feature space to those provided by the PLUTO dataset; we cannot use any other features besides energy consumption from the LL84 dataset since we only have this information for 15,000 buildings. Based upon data from the PLUTO dataset, this step captures certain aspects of the urban context known to affect building energy demand, such as building height (i.e., number of floors) and borough, but overlooks other aspects like canyon aspect ratio and sky view factor. Following a similar approach presented by Robinson, et al., we test several different supervised machine learning algorithms and analyze their error using 5-fold cross validation to prevent overfitting and ensure our models are generalizable [25]. Specifically, we examine linear regression with lasso regularization, support vector machines (regression), random forest, and gradient boosting trees. These four models are some of the most popular in supervised machine learning literature due to their high performance and flexibility from their hyperparameters [52]. Furthermore, each model has shown to be effective for different use-cases for building energy prediction and each can also be tuned precisely, through the use of hyperparameters, to achieve higher performance. Lasso has been used to successfully predict energy savings in school buildings in California [53]; random forests outperformed several other algorithms at predicting building energy use at the urban-scale in NYC [46]; support vector machines have been used to predict multi-family residential building load in NYC at different temporal and spatial scales [28]; and gradient boosting has been found to be effective at predicting university buildings around the globe [32]. Each algorithm has its unique strengths and weaknesses, which is discussed in the subsequent subsections, and allows us to examine which model aspects generalize well to modeling annual NYC building energy demand. Fig. 2 shows the modeling process for step (1).

Comparing these four models allow us to test variable importance, regularization, ensemble learning, bagging, boosting, and the differences between linear and non-linear models. For each of the four models, we perform hyperparameter tuning through a grid search to decrease our prediction errors and identify more thoroughly which model achieves superior performance. We validate the performance of each model using 5-fold cross-validation. The model with the best performance (lowest MSE error) is then selected to predict annual energy use for every building in NYC. Given the high degree of uncertainty introduced in step (2) of the proposed A-UBEM model (described in Section 4.2), we limit our grid search and hyperparameter tuning to a total of twelve combinations for every model excluding lasso regression; due to the fast computation time of this model, we are able to examine a total of 100 models instead of twelve. The errors are evaluated using building-level energy data from the LL84 dataset using Mean Square Error (MSE) as the error metric, where $MSE = \frac{1}{M}\sum_{m-1}^{M_{LL84}}(y_m - \widehat{y_m})^2$. Here, $\boldsymbol{y}$ is the log transformation of annual building energy use (in kBtu), $\hat{\boldsymbol{y}}$ is the predicted output from the model, M_{LL84} is the total number of buildings in the PLUTO dataset $\{1, \cdots, m, \cdots, M_{LL84} = 15,000\}$, and m refers to a specific building. We use the log transformation as this is common in the literature when modeling annual building consumption due to the wide energy consumption range and the heteroskedastic nature of building data [12,46].

4.1.1. Linear regression with lasso regularization

Linear regression is one of the most widely used algorithms in machine learning due to its simplicity and computational speed. Lasso is a regularization method that can be used with linear regression by adding a penalization term to the cost function to simultaneously perform variable selection and regularization by setting coefficients to zero [54]. The lasso penalization adds an L1-norm penalty to the sum of squares cost function in normal linear regression and is controlled by the hyperparameterλ. The linear regression with lasso penalization is as follows:

$$\widehat{\boldsymbol{\alpha}}^{lasso} = \underset{\alpha \in R^n}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \quad (1)$$

where $\widehat{\boldsymbol{\alpha}}^{lasso}$is the vector of coefficients that are being estimated, $\boldsymbol{y}$ is the vector of log transformed annual building energy demand (from the LL84 dataset), and $\boldsymbol{X}$ is a matrix with m buildings and n building features (from the PLUTO dataset).λ can be adjusted to modify the amount of penalty added, thereby changing the cost associated with adding more variables and having increased values for fitted coefficients. For

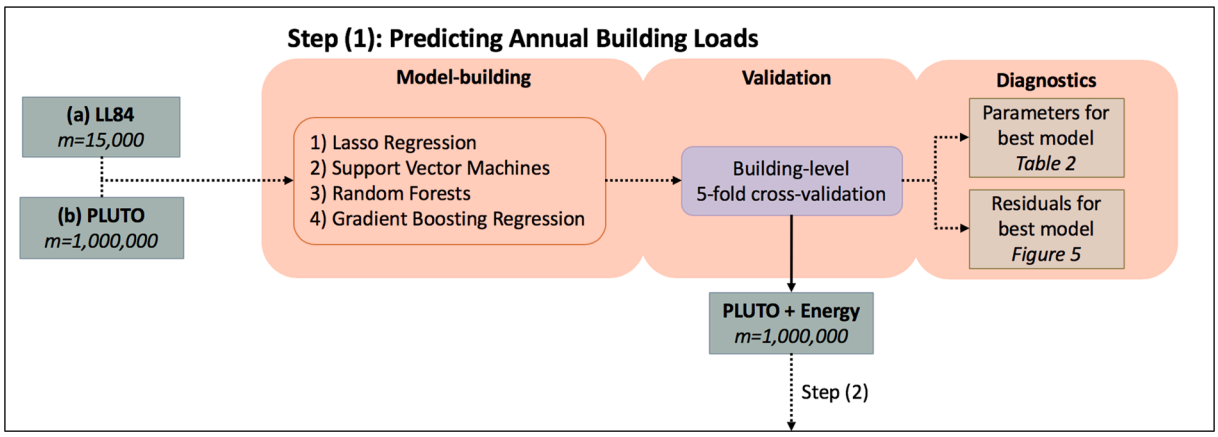


Fig. 2. Flowchart for step (1) in the Augmented-Urban Building Energy Model (A-UBEM), where supervised machine learning is used to predict the annual energy consumption of buildings in NYC.

example, when λ is set to zero, no penalty occurs and a normal linear regression model is fit with all included variables. As λ increases, certain coefficients are forced to zero, effectively choosing a simpler model and excluding those features. In our study, we used the R-package "glmnet" to construct our lasso model and determined the optimal λ through a linear search. Here, the λ parameter is first set to such a high value that all parameters are forced to zero. The parameter, then, is iteratively relaxed letting more parameters obtain a non-zero value, until all features are included in the model. We then observe which value of λ results in the lowest 5-fold cross-validation error.

4.1.2. Support Vector Machines (SVM)

Support vector machines (SVM) is a supervised learning algorithm that can be used for both regression and classification by mapping the input feature space into a higher dimensional plane [28].

$$f(x) = g\left(w^T \varphi(x) + b\right) \tag{2}$$

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C_{reg} \sum_{m}^{M_{LL84}} \max(0, 1 - y_m f(x_m)) \tag{3}$$

$$s.t. \begin{cases} y_m - \langle w, x_m \rangle - b \leq \varepsilon \\ \langle w, x_m \rangle + b - y_m \leq \varepsilon \end{cases} \tag{4}$$

The model uses a kernel function $\varphi(x)$ to do this mapping which allows it to estimate non-linear relationships, according to the regression function $f(x)$, depending on the type of kernel selected. As shown in Eq. (2), $\boldsymbol{w}$ is the weight and b is the bias, which are estimated based on the cost function shown in Eq. (3), where x_m is the training sample and y_m is the annual building energy demand. The support vectors are determined through a discriminating loss function that does not penalize residuals less than a given tolerance, or insensitivity parameter, ε. This means that SVM depends only on a subset of the training data because the cost function ignores training data close to the model prediction, as set by ε. By using a kernel, SVM becomes more computationally efficient by mapping the non-separable feature space to a separable, higher-dimensional space. Much like lasso regression, SVM includes a regularization term C_{reg} to help control overfitting. By choosing a large value for C_{reg}, the optimization will choose a smaller-margin hyperplane, which is constructed based off the support vectors.

4.1.3. Random forest

Random forests is an ensemble supervised learning algorithm that can be used for regression and classification [55]. For regression, the model constructs many regression trees and averages the results from each tree to produce a final prediction. This model addresses the bias-variance tradeoff that many models face by producing many trees—which are weak learners and suffer from high bias—and averaging their results, rather than producing a single model which typically suffers from high variance. To train the regression trees, random forests uses bootstrap aggregating, or bagging, which both randomly selects training data and features with replacement for *each* tree independently. The bootstrapping decorrelates the individual trees and reduces the variance by averaging the results. Each tree is built using about $1 - e^{-1} \approx 2/3$ of the training data, where the error of the model can be calculated using the remaining unseen $^1/_3$ of the data, which is known as the Out-Of-Bag (OOB) estimate. This OOB estimate acts as a type of cross-validation—which can occur in parallel with the training step—and helps ensure that the model is not being overfit [56]. The number of features to include for each tree is set as a hyperparameter, where we will examine which number results in the lowest OOB estimate. In our study, we use the R-package "randomForests", which implements Breiman's random forest algorithm for regression [57]. We used 200 trees to build our models, because random forests has been shown to have a negligible increase in performance above this value [56]. The OOB estimate is calculated using MSE in order to make it comparable to the cross-validation MSE metrics in the other three models.

4.1.4. Gradient boosting regression

Similar to random forests, gradient boosting regression is a tree-based method that combines an ensemble of weak learners to improve prediction accuracy. Unlike random forests, gradient boosting builds the model in a stage-wise fashion, by first building one regression tree and then iteratively constructing new regression trees on the current residual, one after another. The algorithm continues to build new trees until a maximum number of iterations, provided by the user, is reached. Gradient boosting regression is a numerical optimization algorithm that builds an additive model that minimizes the loss function by iteratively adding a new regression tree at each step that best reduces the loss function. For each subsequent tree, the provided learning rate η is used to shrink the contribution of the tree, thereby providing a higher number of small trees, which can provide a higher accuracy than a lower number of large trees [27]. The η parameter takes on a value between 0 and 1, with smaller numbers resulting in a higher number of trees. In our study, we use the R-package "xgboost" which uses an efficient implementation of the gradient boosting framework from Chen and Guestrin to create a scalable tree boosting system [58]. We used the default parameters for tree depth and the fraction of data to be used at each iterative step (100%).

4.2. Step (2): Constructing building hourly profiles

We extend the annual building energy demand from step (1) to predict hourly building energy curves in step (2). To create building hourly profiles for each building in NYC, we leverage the building hourly load profiles from the simulated 19 DOE reference buildings. We assign three of the 19 profiles to each building class (and each building as a result) originally defined by the PLUTO dataset. Other recent work has used a similar strategy of mapping DOE reference buildings to PLUTO classes [59]. We assign three DOE profiles because of the imperfect alignment between the PLUTO dataset's 25 building classes and the 19 DOE reference buildings. This one-to-three mapping was selected to reduce bias introduced by the authors, and can be seen in Appendix A in the supplementary document.

After each building class is assigned three reference buildings, we assign weights for each of the three reference profiles. Here, the explicit assumption is that each building's final profile can be approximated based upon a weighted average of the three reference buildings that it is assigned. To find this weighting, we define a convex loss function that produces six parameters for each PLUTO class: three weights for the DOE reference buildings, two parameters for weather adjustments (based on observed hourly temperature and cooling degree hour), and one parameter for business day adjustments. With 25 PLUTO classes, this results in a total of 150 parameters. To examine the stability of the model, we use a Monte Carlo simulation—using different random subsets of buildings—to solve the convex optimization formulation multiple times. Finally, we utilize particle swarm optimization to assign each building one of the solutions from each Monte Carlo simulation and asses the change in model results and error. Fig. 3 shows the modeling process for step (2).

4.2.1. Convex optimization

The objective of this function is to minimize the square difference between the *predicted hourly aggregated building load* and the NYISO city-wide hourly load, as seen in Eq. (5a). Let $\boldsymbol{E} \in \mathbb{R}^T$ be the normalized vector of the NYISO city-wide hourly load, where $\sum_{t=1}^{T} E(t) = 1$ and $T = 8784$, the total number of hours in the 2016 leap year. Let $\widehat{\boldsymbol{E}} \in \mathbb{R}^T$ be the similarly normalized *predicted hourly aggregated building load,* as seen in Eq. (5b), which is defined as the sum of the individual building profiles—denoted by $\boldsymbol{Y}(t, \boldsymbol{\beta}) \in \mathbb{R}^{T \times J}$—multiplied by their scaled annual energy use $\overline{\boldsymbol{P}} \in \mathbb{R}^J$, as shown in Eq. (5e) (i.e., dot product). The optimization problem is as follows:

$$\underset{\beta}{\text{minimize}} : \|\widehat{\boldsymbol{E}}(\boldsymbol{\beta}) - \boldsymbol{E}\|_2^2 \tag{5a}$$

$$\text{subject to} : \widehat{E}(\boldsymbol{\beta}) = \boldsymbol{Y}(t, \boldsymbol{\beta}) \cdot \overline{\boldsymbol{P}} \tag{5b}$$

$$Y_{ij}(t, \boldsymbol{\beta}) = \beta_i^1 D_i^1(t) + \beta_i^2 D_i^2(t) + \beta_i^3 D_i^3(t) + \beta_i^4 T(t) + \beta_i^5 C(t) + \beta_i^6 W(t) \tag{5c}$$

$$\sum_{t=1}^{T} Y_{ij}(t) = 1 \tag{5d}$$

$$\overline{\boldsymbol{P}} = \frac{\boldsymbol{P}}{\sum_{j=1}^{J} P_j} \tag{5e}$$

$$\beta_i^1 + \beta_i^2 + \beta_i^3 = 1 \tag{5f}$$

$$0 \le \beta_i^k \le 1, \quad \text{for } k = 1, 2, 3 \tag{5g}$$

where $Y_{ij} \in \mathbb{R}^T$ denotes the hourly energy demand for building j of PLUTO building class i, where$i = \{1, 2, \cdots, I = 25\}$, and Y_{ij} is scaled as shown in Eq. (5d)—this ensures that $\sum_{t=1}^{T} E(t) = \sum_{t=1}^{T} \widehat{E}(t) = 1$. All of the design variables—totaling 150 with six variables for each of the 25 PLUTO classes—are expressed in Eq. (5c), as $\{\beta_i^1, \cdots, \beta_i^6\}$: three for the weights of the DOE reference buildings, two for weather adjustments, and one for business day adjustments. Let D_i^1, D_i^2 and D_i^3 denote the three DOE reference buildings mapped to PLUTO building category i, where $i = \{1, 2, \cdots, I = 25\}$. See Appendix A (in the supplementary document) for the mapping of each PLUTO class to the three DOE reference buildings. Let $\boldsymbol{T}$ be the NYC hourly temperature vector and $\boldsymbol{C}$ be the NYC cooling degree hour which is defined as $C(t) = \max(0, T(t) - 65)$ where $t = \{1, 2, \cdots, T = 8784\}$for every hour in the 2016 leap year. Let $\boldsymbol{W}$ be a vector indicating business days, which are given a value of 1, while all weekends and holidays are given a value of 0. Finally, to ensure that we obtain a weighted average of the three DOE reference buildings assigned to each PLUTO class, we add two more constraints, as shown in Eq. (5f) and Eq. (5g).

Let $\boldsymbol{P} \in \mathbb{R}^J$ indicate the annual energy use for all buildings J, where vector $\boldsymbol{P}$ is the result of step (1) described in Section 4.1. For each building j, the mapped DOE reference building for each building is randomly shifted a few hours based on a normal distribution $N(0, 1.5)$ to simulate variations in building schedules. Because we cannot include all 1 million buildings in the optimization function due to computational limitations, we select a random sample of 1000 buildings, meaning $J =$

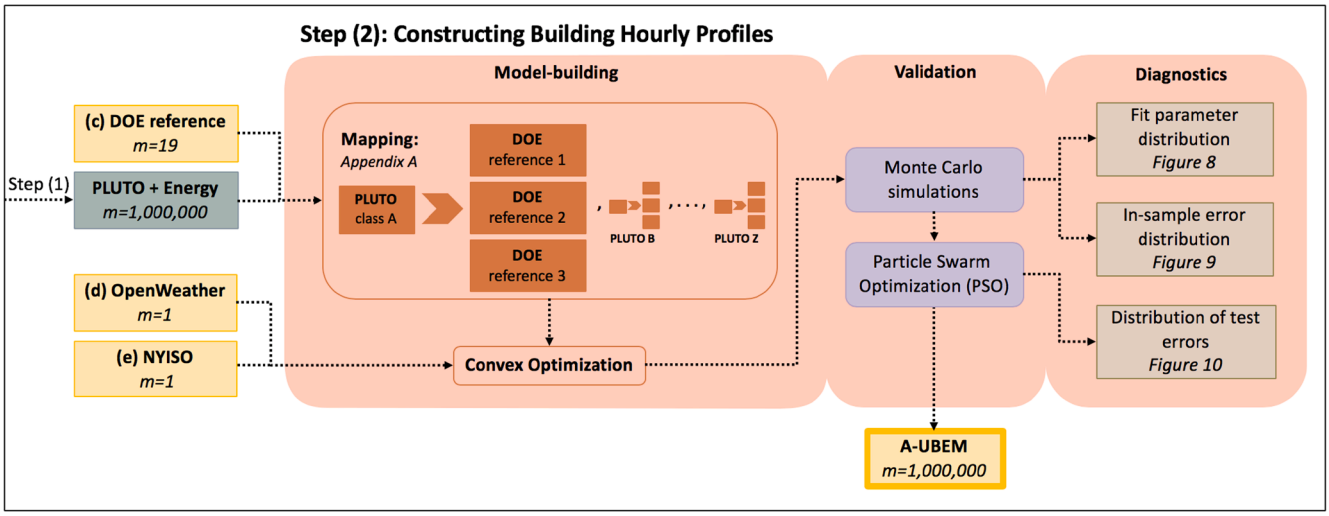


Fig. 3. Flowchart for step (2) in the Augmented-Urban Building Energy Model (A-UBEM), where a Monte Carlo Simulation is used to solve multiple convex optimization formulations producing a set of parameters β, for each simulation, and stored in β_{matrix}. The results for all the simulations, orβ_{matrix}, are then run through a particle swarm optimizer to assign each individual building a set of parameters.

1000 instead of 1 million. The 1000 random buildings are selected through stratified sampling—based on the PLUTO building class since the population of each class varies—to obtain a representative sample of buildings [60]. Creating a subpopulation of the NYC building stock then requires that the NYISO city-wide hourly load vector be on the same scale as the *aggregated hourly building load*. The annual building loads are also normalized (Eq. (5e)), thus ensuring that $\sum_{t=1}^{T} E(t) = \sum_{t=1}^{T} \widehat{E}(t) = \sum_{j=1}^{J} \overline{P}(j) = 1$.

The objective function is a quadratic function, composition of the square of the Euclidean norm and an affine function, thus convex in the design variables $\boldsymbol{\beta}$. In addition, the feasible domain is convex since all equality constraints are affine functions [61]. See Appendix D (in the supplementary document) for more detailed proof of the convexity of the problem. Since the optimization problem is convex, the CVXPY python-embedded modeling system for solving disciplined convex programs is used as the optimization solver [62].

4.2.2. Validation and model stability

Using Monte Carlo Simulations, we can examine the stability of step (2) of the model by examining the distribution of fit parameters and errors, adding a component of validation to the results of the Augmented-Urban Building Energy Model (A-UBEM). Although the defined convex function ensures a global optimum, several limitations arise. First, it assumes that all buildings *within* a PLUTO building class have the same profile shape; buildings of the same type, like apartment buildings for example, should have similar load shapes but not exactly the same. Second, the large size of the dataset, with 1 million buildings and 8784 h in the year, makes this problem computationally infeasible and thus only a subset of buildings was used for the optimization (1000). Third, the optimization function is trying to minimize the distance between two large vectors with several constraints—some of which might be unrealistic (e.g., all buildings of the same building type have the same load profile)—meaning that some error between the two vectors is likely to remain. To address these limitations, we run 500 Monte Carlo Simulations using 1000 randomly selected buildings—through the stratified sampling technique described previously—on the first 1000 h of the year. We determined that 500 simulations would balance computation time with ample variation in sampled buildings to provide adequate understanding of model stability.

Using the$\boldsymbol{\beta}$ results from all 500 simulations, we then use particle swarm optimization (PSO) to assign one of the 500 vectors from the simulation to each individual building. This approach a relaxes the constraint that all buildings of the same PLUTO building class must have the same profile but is still grounded in the original convex formulation. Furthermore, having a more diverse set of building profiles allows the aggregated load profile to more closely match that of NYISO, hence better approximating the true load. With a total of $J = 1000$ buildings, this results in a total of 500^J different combinations of buildings and parameters, hence the reason to use the PSO.

PSO is a non-gradient population-based optimization technique whereby candidate solutions, called particles, iteratively move around the search-space—in this case, the search space is the 500^K different combinations of buildings and parameters [63]. Initially, the particles are randomly placed in the search-space. Then at each iteration (or generation l), the particles move to a new location based on the best solution acquired by each particle in all the previous iterations ($\boldsymbol{P}_l^{best}$) and the best value obtained by any particle ($\boldsymbol{G}_l^{best}$) [64]. The position of the particle $\boldsymbol{x}_l^a$, shown in Eq. (6), is updated based on the velocity $\boldsymbol{v}_{l+1}^a$, shown in Eq. (7).

$$\boldsymbol{x}_{l+1}^a = \boldsymbol{x}_l^a + \boldsymbol{v}_{l+1}^a \tag{6}$$

$$\boldsymbol{v}_{l+1}^a = w_{inertia}\boldsymbol{v}_l^a + c_1 r_1 \left(\boldsymbol{P}_l^{best} - \boldsymbol{x}_l^a\right) + c_2 r_2 \left(\boldsymbol{G}_l^{best} - \boldsymbol{x}_l^a\right) \tag{7}$$

Here, the user-define parameters are: $w_{interia}$ which is the inertia used to control the effect of the particles' previous velocity on the current velocity; c_1 which is the cognitive learning factor used to control the velocity toward the particle's previous best value; and c_2 which is the social learning factor used to control the velocity toward the globally best particle. The variables r_1 and $r_2 \in [0, 1]$ are two independent random numbers used to keep the particles from falling into a local-minima and permit a small percentage of particles to explore the larger search-space.

Because the Monte Carlo simulation relies on random sampling, we obtain 500 unique solutions to the convex optimization formulation described above. With these 500 vectors, each containing coefficients for the 150 decision variables, we can then define another optimization function that assigns each building one of these 500 vectors. For the PSO, we use the mean absolute percentage error (MAPE), as shown in Eq. (8), as the objective function. Both optimization formulations with first and second norm lead to a convex problem, with the same global optimum. However, for practical implementations and convergence issues, the Euclidean norm was chosen for the CVX implementation. Eqs. (5d)–(5g) are still applicable, but Eq. (5a) is substituted for Eq. (8), where the number of buildings represented in this sample is $j = \{1, 2, \cdots, J\}$ where $J = 1000$. The PySwarms python library is used to model and solve the PSO [65]. For each building j, the $\boldsymbol{\beta}$ parameters are determined by selecting from one of the 500 simulations.

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{T}\sum_{t=1}^{T} \frac{\left|\boldsymbol{E} - \widehat{\boldsymbol{E}}(\boldsymbol{\beta})\right|}{\widehat{\boldsymbol{E}}(\boldsymbol{\beta})} \tag{8}$$

$$\text{subject to} \quad Y_{ij}(t, \boldsymbol{\beta}) = \beta_i^1 D_i^1(t) + \beta_i^2 D_i^2(t) + \beta_i^3 D_i^3(t) + \beta_i^4 T(t) + \beta_i^5 C(t) + \beta_i^6 W(t) \tag{9}$$

5. Results

We present our results for the A-UBEM model below and split up the results into two sections, following the model construction outline as shown in Fig. 1. First, the results for the building annual loads (step (1)) are presented and we discuss the prediction errors for each supervised machine learning model. Second, we discuss the hourly building profile results from the convex optimization algorithm (step (2)) and inspect the distribution of the fit parameters from the Monte Carlo simulations to show that our results are robust. Third, we examine the added benefit in running the particle swarm optimization. Finally, we validate our model at the city scale by using the fit parameters from the optimization algorithm and compare our constructed model to the ground-truth NYC electricity demand.

5.1. Predicting annual building loads

Before modeling, each feature from the original collected dataset that showed a high skewness—a total of ten—underwent a log transformation and was added to the original list of features used in the modeling process. Of these ten features, nine were related to measurements of area in buildings and one was the assessed total value of the building. A complete list of the features used in the modeling process can be seen in Appendix B in the supplementary document, along with summary statistics which highlight the skewness of the untransformed features. By including both transformed and untransformed features, each model could then determine the appropriate combination of features needed to represent the non-linear effects that they have on energy consumption. A summary of the four examined models and the MSE from the best set of hyperparameters for each model is shown in Table 2.

All of the models have fairly similar performance, but the random forests model proved to be the best with the lowest MSE of 0.293. Having a non-linear model perform best is consistent with findings from Kontokosta and Tull [46] and Robinson et al. [25]. Because we perform a log transformation of the kBtu values, the MSE shown in Table 2

Table 2
Summary of the parameters examined for each of the models and the final parameters of the best model as tested using 5-fold cross-validation. The shown error rates are for the models with the lowest cross-validation MSE after performing the grid search and selecting the optimal hyperparameters.

Models	Hyperparameters	Final Parameters	Final MSE
Lasso Regression	Penalization: $\lambda = [0,1]$	Penalization:λ = 0.0104	0.312
Random Forest	Max Features: *3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25*	Max Features: *5*	0.293
Gradient Boosting	# Boosting Iterations: *1000, 2000, 3000, 4000* Learning rate:$\eta = 0.01, 0.001, 0.0001$	# Boosting Iterations: *1000* Learning rate:η = 0.001	0.343
Support Vector Machines	Kernel: *Linear* Penalty Factor: $C_{reg} = 1,3,100$ Insensitivity parameter: $\epsilon = 0.1, 0.4, 0.7, 1.0$	Kernel: *Linear* Penalty Factor: $C_{reg} = 1$ Insensitivity parameter: $\epsilon = 0.4$	0.316

should be interpreted as e^{MSE} multiples away of the predicted value from the true energy value for a building on average. For each of the models, we report the MSE of the 5-fold cross validation, except for random forest which uses the OOB estimate as described in Section 3. Random forests also showed to have the most similar performance using different hyperparameters, while the support vector machine models showed to have the highest variability between different hyperparameters. This indicates that not only is the random forest model the best performing overall but that it is also the most consistent, producing stable results. The gradient boosting model was the worst performing model. We postulate that the gradient boosting model may be able to perform better with further hyperparameter tuning, given the high number of hyperparameters in the model. However, we note that even though gradient boosting is often praised for having high performance, this typically comes at the expense of more computation time as it can be difficult to find the optimal hyperparameters [66]. The random forests model requires less hyperparameter tuning than gradient boosting, while still producing an accurate model as our results show.

Fig. 4 below shows the top ten most important features for the best random forests model. The model calculates the feature importance by substituting a vector of noise for each feature and measuring the amount

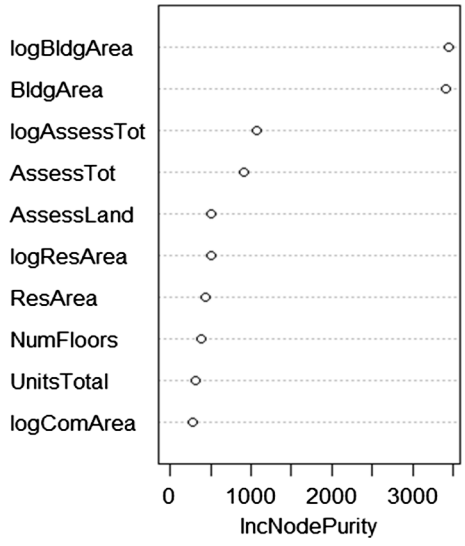


Fig. 4. Summary of the top ten most important features in the best random forests model.

the error increases. The resulting increase in error can be interpreted as the amount of decreased error that particular feature produces.

Year built, building area, assessed price per square foot, and number of floors are the most important variables for building the random forests model. Fortunately, this data is fairly common in cities across the US, though annual building energy data is only available in cities with energy disclosure policies. Since many cities collect this type of data already, the results from Fig. 4 indicate that these cities do not need to go through the effort of collecting the other types of data that NYC possess because they do not add much power to the fit of these models. In short, this shows that other cities can easily replicate the first step of this A-UBEM methodology to create annual energy building load estimates for their entire city. Similar important features were found for a data-driven predictive model for energy use in NYC by Kontokosta et al., however, the authors also found that zipcode level energy consumption was driven by a different set of variables [46].

In Fig. 5, we compare the predicted annual energy use made by the best random forests model to the ground truth data. The figure shows that the random forests model is neither under- nor over-estimating the ground truth energy consumption. The residuals follow a log-normal distribution indicating that the model is not biased.

With the constructed model parameters, we then apply the model to all one million buildings in NYC to predict annual energy use at each building. Fig. 6a shows a map of our predicted annual energy use for just Manhattan aggregated at the block-level by summing the annual energy use for all buildings within a block. Just below Central Park, midtown Manhattan is seen to have some of the highest energy consuming buildings, as shown in the map of building-level annual energy consumption in Fig. 6b. Here, we highlight the Chrysler building as a reference for our results for step (2) where we produce hourly loads of every building.

5.2. Constructing hourly building loads

After predicting individual building hourly load profiles for all of NYC, we aggregate these profiles into a single city-level profile to compare to the actual load from the NYISO dataset. The aggregated building hourly load profile for one of the Monte Carlo Simulations of step (2) is shown in Fig. 7. This figure shows the NYISO actual load (in blue) for 11 days in January and the optimized aggregated building load (in orange) over the same period. Because each simulation is using a subset of the entire NYC building stock, the y-axis in the plot is normalized over the entire year to have a mean of 1. The aggregated load resulting from the optimization accurately captures the overall trend—including intraday fluctuations, weekday/weekend differences, and trends over time—with some minor differences compared to the NYISO observed load. These minor differences stem from several assumptions in the optimization: (1) the manually constructed one-to-three mapping from each PLUTO building class to the DOE reference buildings; (2) that all buildings within a PLUTO class are assigned the same load profile; (3) the scaling of building profiles using their total energy demand rather than electricity demand due to data constraints.

5.2.1. Framework stability and robustness

One of the main reasons to perform a Monte Carlo Simulation is to observe the stability of the fit parameters using different random samples of buildings for the convex optimization formulation. Fig. 8 shows the distribution for the three β parameters associated with the three DOE reference profiles respectively assigned to four PLUTO classes. These four classes are highlighted as they are representative of the four types of distributions observed in each of the 25 PLUTO classes. See Appendix C (in the supplementary document) for the distributions of all 25 classes.

The β parameter distribution for Condominiums (PLUTO class R) as shown in Fig. 8a, is similar to that of nearly half (12/25) of the other PLUTO classes; this distribution type 1 is represented by the background slate gray color and can also be observed in Appendix C. The distribution

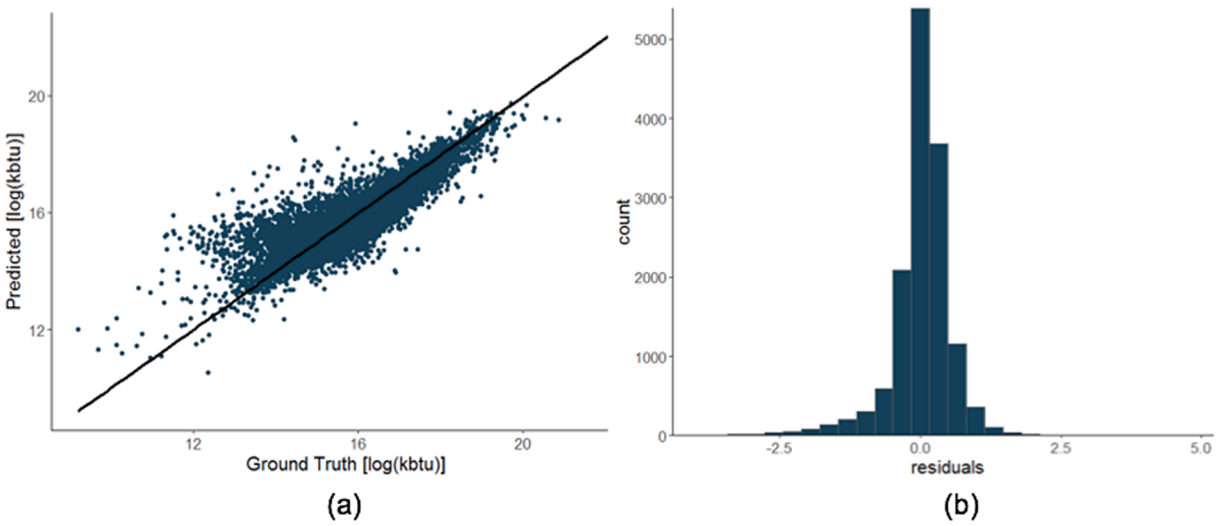


Fig. 5. Error plots comparing the difference between the best random forests model prediction and the ground truth data (a) predicted vs ground truth energy load and (b) residual distribution.

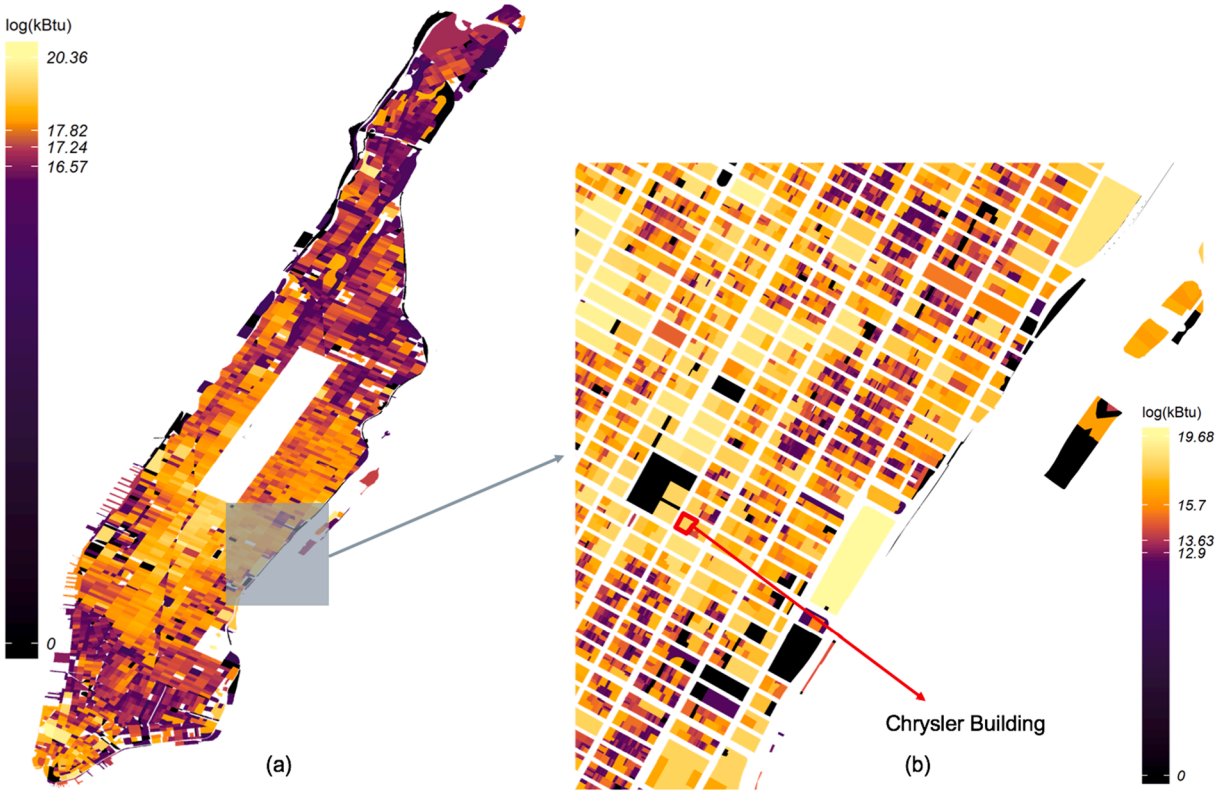


Fig. 6. The map of the entirety of Manhattan (a) is aggregated at the block-level in order to better visualize spatial trends in energy use. The zoomed in map (b) is shown at the building-level to highlight the output of step (1). Both maps use a gradient with breakpoints set at the 5 quantiles of the log energy consumption of the shown blocks and buildings, respectfully.

for this set of PLUTO classes shows that across all 500 simulations, two of the three β parameters are nearly always zero and the other β parameter is nearly always one. The interpretation of this distribution—for condominiums specifically—is that its load profile most closely matches that of the small hotel building from the DOE reference building set, as opposed to a midrise apartment or large residential home, since β_i^3 is receives a value of nearly 1 for nearly all 500 simulations. Appendix A (in the supplementary document) includes the one to three mapping for all other PLUTO classes. Another common distribution for 4 of the 25 PLUTO classes is a near equal weighting for all three parameters, as shown in the β parameters for Theatre (PLUTO class J) in Fig. 8b; distribution type 2 is represented by the background light yellow color and can also be observed in Appendix C. These 16 distributions—exemplified in Fig. 8a and b—account for over 95% of the buildings in NYC. The consistency of these β parameters across all the Monte Carlo Simulations demonstrates the stability of the model for these PLUTO classes.

The plot in Fig. 8d shows a distribution where two β parameters

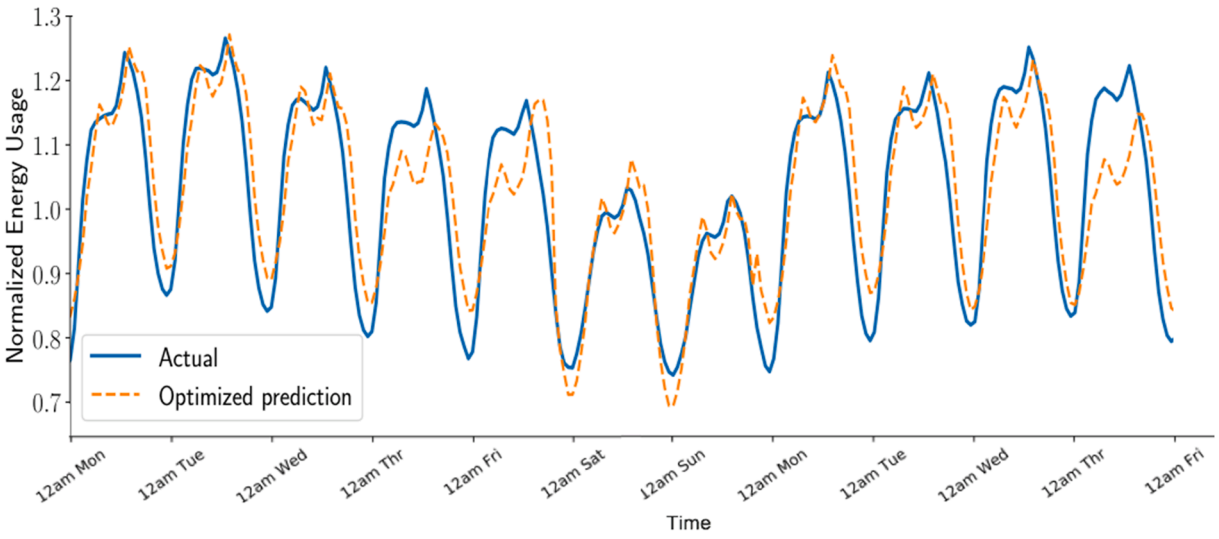


Fig. 7. This is the output for one Monte Carlo Simulation showing the hourly load profile of the actual NYISO electricity demand for NYC (in blue) and the aggregated building load profile (in orange) for 11 days in January.

receive a value of 1 (β_i^1 and β_i^2) for a substantial portion of the simulations, one β parameter is nearly always set to zero (β_i^3), and the remaining values for all three β parameters are uniformly distributed between 0 and 1. This spread indicates that two of the three mapped DOE reference buildings adequately approximate the profile of the respective PLUTO class. There is a total of 6 distributions that follow this pattern. Finally, the distribution shown in the Fig. 8c is the most varied of all the distributions. There are 5 distributions that do not fit into the previously defined three clusters, but buildings in these classes make up less than 1% of the buildings in NYC. Because we used stratified sampling and a total of 1000 buildings for each simulation, the small number of buildings in each class may be one reason why these distributions are much more varied than the others.

The Mean Absolute Percentage Error (MAPE) across all 500 Monte Carlo simulations also exemplifies the consistent performance of step (2), as shown in Fig. 9. The mean MAPE across all the simulations is 6.11% and the distribution has a narrow range between a minimum of 4.57% and a max of 9.30%.

Results shown in Figs. 8 and 9 show that the convex optimization in the step (2) framework produces robust results across various input buildings for the city. In addition, the model only requires a relatively few number of building annual loads (1000) to accurately predict the building hourly load profile of an entire city (1 M buildings), as long as the building samples are representative of the building stock. This is critical since other public datasets, in other geographical locations, may only contain a relatively small data sample of building annual loads.

5.2.2. Validation of the convex optimization

With all the fit parameters from the Monte Carlo Simulation, we then validate their fit using a left-out set (i.e., test set) of 1000 h that the convex function did not use to train the model. Using this same set of 1000 h, we also run the particle swarm optimizer (PSO) to assign each individual building the $\boldsymbol{\beta}$ results from one of the 500 simulations, again to relax the assumption that all buildings of the same PLUTO class must have the same profile. Using the test set of 1000 h, Fig. 10 shows the distribution of MAPE for both the Monte Carlo Simulation and the PSO. Compared to the in-sample training error, shown in Fig. 9, both the Monte Carlo and PSO showed marginally worse out-of-sample errors; given that the parameters were trained using a different set of 1000 h, this is unsurprising.

The PSO results show an improvement over the Monte Carlo results but only by a small amount, with an average reduction of error of 0.09. These results show that the PSO does reduce the error of the model but not significantly. This shows that the assumption that all buildings from a PLUTO class have the same energy profile is actually quite negligible and does not greatly affect the final energy load curve prediction. The distribution of the error also shows that a small set of random buildings is sufficient to get a good prediction for the entire city. The PSO, therefore, is not needed as the original optimization produced nearly as low errors.

5.2.3. Example hourly energy curves for the Chrysler building

Given that the model produces parameters that are weighted averages of DOE reference buildings—where β is the weight—we can extract hourly energy curves for electricity, gas, heating, and cooling for every building in the city. As an example, Fig. 11 shows the produced hourly energy curves for one building in NYC—the Chrysler building—for two sample weeks, one in January and one in June. The figure shows a distribution of estimated energy curves across all 500 simulations. Despite each simulation using a random sample of buildings, each figure shows a consistent trend in estimated loads across all 6 categories: total energy, electricity, gas, heating, cooling, and water heating.

We estimate the expected energy loads for winter (gas and heating) and summer (cooling), along with their daily fluctuations. In January, heating loads are higher in the morning, when it is colder, while in June, cooling loads increase throughout the day. In addition, the total energy loads are lower on the weekends in comparison to weekdays, as the building is primarily comprised of office spaces. For 2016, the Chrysler building used a total of 57.563 million kBtu in energy, with 48.882 million kBtu coming from electricity. The Chryster building has a floor area of 1.04 million square feet, which results in an energy use intensity of 55.6 kBtu/ft^2.

6. Applications & discussion

By better understanding how buildings use energy, both spatially and temporally, policymakers, engineers, and planners can make more informed decisions regarding energy use and urban design. New York City (NYC), like other major cities, imports more electricity than it consumes—using nearly 60% New York state's electricity demand but only creating about 40% of it—thereby requiring city officials to grapple with unique procurement and planning challenges [67]. For example, in different geographical zones in NYC, the amount of energy to procure, as well as its price, changes based on time-of-day. Decision-makers must, therefore, compare energy supply alternatives, evaluate savings from potential retrofits, and investigate impacts from targeted programs with

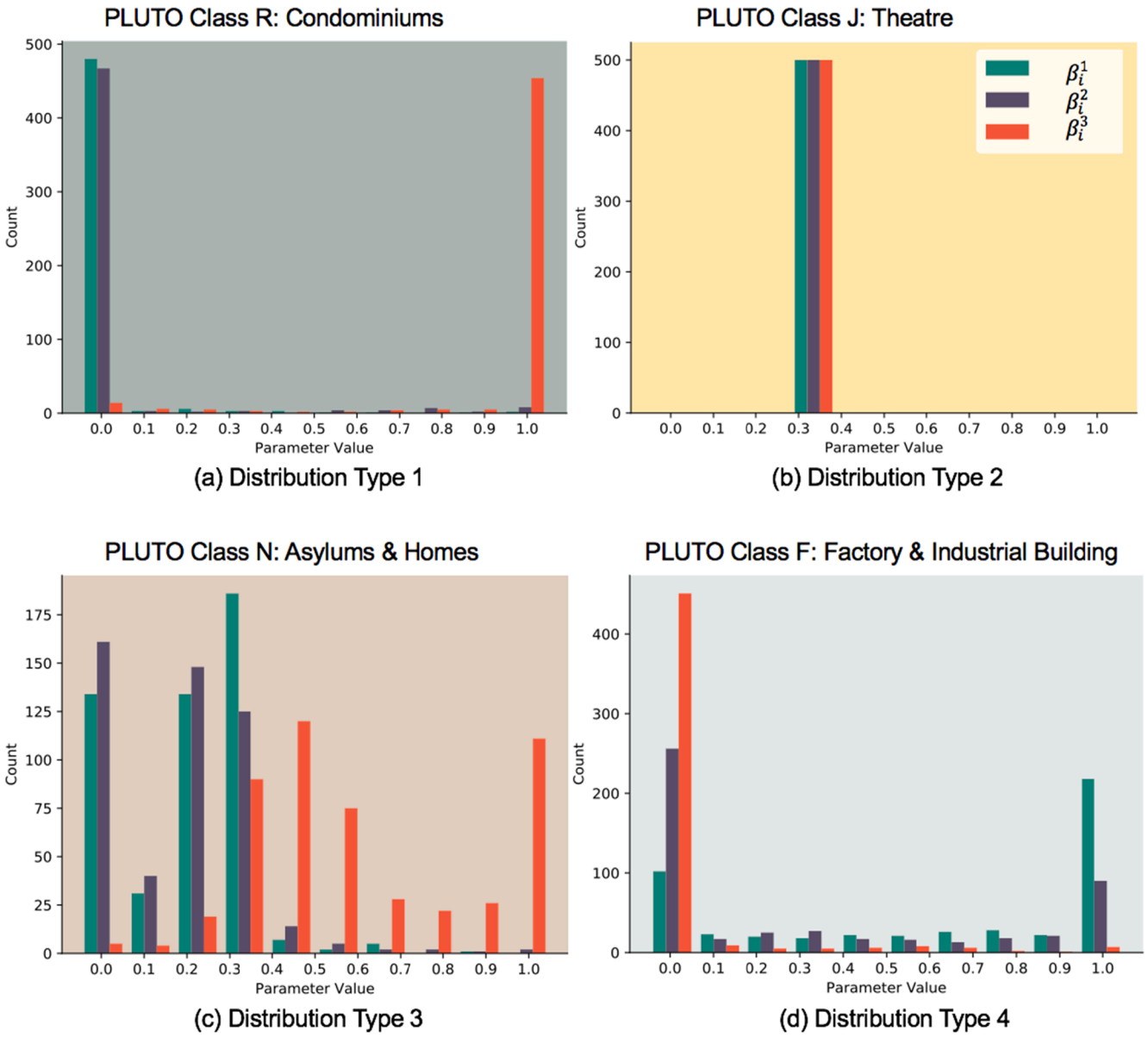


Fig. 8. The distributions for four representative PLUTO classes and their fit parameters across all 500 Monte Carlo simulations: (a) Distribution Type 1 (ash gray); (b) Distribution Type 2 (yellow); (c) Distribution Type 3 (rose); (d) Distribution Type 4 (silver). Of the 25 classes, 16 exhibit similar distributions to the top two plots where nearly all the parameters receive the same value across all the simulations. These 16 classes account for upwards of 95% of the buildings in NYC. The bottom two plots show more variation in the distribution of the fit parameters. See Appendix C (in the supplementary document) for histograms for all 25 PLUTO classes using these four background colors to indicate distribution type. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the goal of providing cheap energy prices and a reliable grid [68]. Cities like New York must also grapple with grid congestion issues, decarbonization efforts, and electrification initiatives which makes this goal even more difficult. But new energy systems are emerging that can be installed in targeted locations within cities which promise to create cheap and clean electricity.

Distributed energy resources (DERs) and district energy systems are two such solutions that can be used to reduce and shift demand, but hourly energy demand information at the block- and building-level are needed to ensure competitive prices. Without this high level-of-detail, these technologies will be implemented piecemeal within cities—as decision-makers will struggle to target locations and buildings that could most benefit—thereby slowing their adoption rates and decreasing their savings potential. The city-wide installation of *smart meters* with sub-hourly measurement capability would be an obvious solution to this challenge. However, the installation expense and lack of strong return-on-investment of such technology can be a barrier to implementation. This paper outlines a process of creating an Augmented-Urban Building Energy Model (A-UBEM) that provides synthetic hourly loads for all the buildings within a city that can help decision-makers systematically deploy clean, cost saving energy efficiency and resource solutions.

6.1. *Distributed energy resources*

New demands for decentralized energy, volatile fossil fuel prices, and technological advancements are leading to increased interest in distributed energy resources (DERs) across the world [69]. DERs—include technologies like solar, storage, and wind—have already experienced precipitous growth in adoption as their costs have fallen to

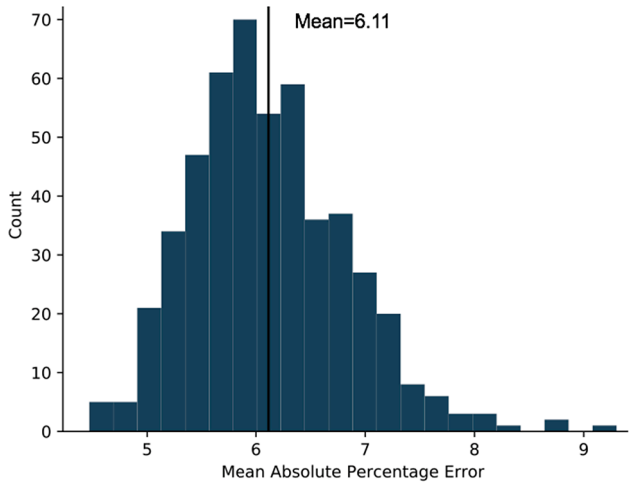


Fig. 9. The distribution of the in-sample Mean Absolute Percentage Error (MAPE) for each simulation in the Monte Carlo.

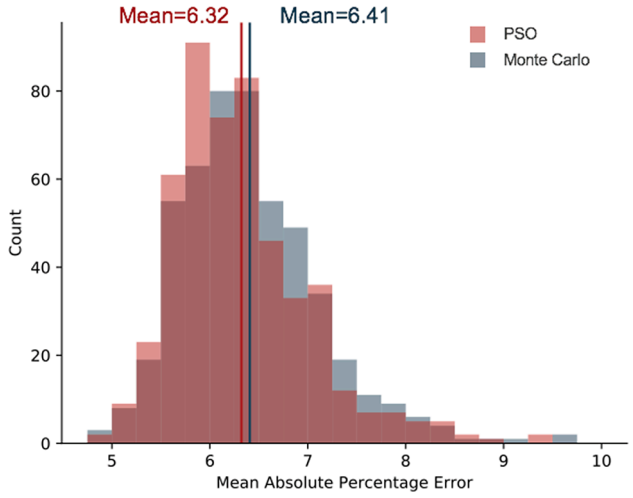


Fig. 10. The distribution of the out-of-sample (i.e., test set) Mean Absolute Percentage Error (MAPE) for the Monte Carlo Simulation and the Particle Swarm Optimizer (PSO). The errors using the test set are only marginally worse than the training set, as shown in Fig. 9. The PSO shows a small improvement in error reduction compared with only using the Monte Carlo without the PSO.

compete with traditional fossil fuel prices [7]. These technologies are helping cities curb greenhouse gas emissions associated with burning fossil fuels but come with new challenges and opportunities that cities must grapple with. Unlike power generated from fossil fuels, solar and wind rely on weather patterns to produce energy which cannot be controlled, leading to intermittency in production and energy forecasting challenges. Indeed, there is a mismatch between timing of renewable energy production, especially wind and solar, and energy consumption. Ultimately, these decarbonization efforts, if unplanned, can destabilize the grid by creating large spikes in energy production which could lead to costly blackouts.

By having a more detailed urban building energy model, city officials can use this resource to identify locations within their city that could benefit from DER installations, ensuring that intermittent energy production is strategically located. Coupling this model with information about transmission and distribution line capacities, cities could install DERs in targeted locations that could prevent the need for costly grid upgrades. Over the next 30 years, NYISO estimates that about 4700 miles of high-voltage transmission lines need to be replaced at a cost of about $25 billion [67]. Furthermore, buildings with large loads, and therefore high utility bill demand charges, can be identified and benefit from the installation of battery storage to shift their load, or solar PV systems to reduce their load. Increased adoption rates and strategic deployment of DERs can replace the need for expensive generation plants, decrease peak electricity prices, improve energy security, raise air quality levels, add local jobs to the market, and enhance operating efficiency and flexibility [70].

6.2. *District energy*

Thermal microgrids, also referred to as 4th generation district energy, are a new and rarely discussed form of energy generation, which is promising to substantially aid in the reduction of fossil fuel use [71]. China, for example, requires by law that all northern cities have district heating systems [72]. Traditional district energy systems rely on some type of fossil fuel (e.g., natural gas) that is burned to generate electricity, while the waste heat is used to warm a fluid (e.g., water) that can then be distributed to buildings through insulated pipes. By heating water at a centralized location, aided by the waste heat from generating electricity, district heating plants can provide higher efficiencies than numerous localized boilers. A new take on district energy has eliminated the need for fossil fuels and instead relies on the use of renewables, large heat recovery chillers, and thermal storage. These new systems simultaneously generate chilled and hot water by transferring heat from one liquid to the other and then storing the hot and cold water in large thermal storage tanks for later distribution. Already, there are a few such systems in place, like the central energy facility on Stanford University's campus [73].

Increases in thermal efficiency for the building sector can have far-reaching effects since heating and cooling make up over 60% of residential and 50% of commercial energy demand [74]. Because buildings typically demand both hot and cold water, these centralized systems can provide higher efficiencies while also shifting electricity demand to off-peak hours.

Assessing whether a location is suitable for this type of thermal microgrid requires data on hourly building loads in order to model the magnitude of savings achieved through district scale electrification and load shifting. Current data limitations have restricted the implementation of such systems to places like universities and campuses which have access to load profiles for large numbers of buildings. A new tool that can model building loads at the city-scale can help engineers better assess the feasibility of thermal microgrids at numerous locations. Our proposed model (A-UBEM) addresses the data limitation challenge associated with district energy assessment which could lead to more of these efficient systems being installed in the future.

7. Limitations and future work

The Augmented-Urban Building Energy Model (A-UBEM) presented in this work makes several key assumptions. First, it assumes that the aggregated building profiles from a small subset of buildings approximates the city-wide profile for NYC; building profiles are scaled using their total energy demand rather than electricity demand due to data constraints. Second, it assumes that each building profile is a linear combination of three mapped DOE reference buildings, as seen in Appendix A in the supplementary document. Third, the model assumes that all buildings within the same PLUTO class have the same energy load curve. Fourth, the model does not distinguish between neighborhood effects for different parts of the city, such as morphology and density, but rather determines the *average* effect of the city-wide urban context. And finally, though each building profile is adjusted for weather and business day operations, the model is not validated at the building-hourly level since this data could not be obtained; acquiring interval-level data for a small subset of buildings would help validate the model. Despite these limitations, this research presents a novel methodology for combining data-driven and physics-based techniques to

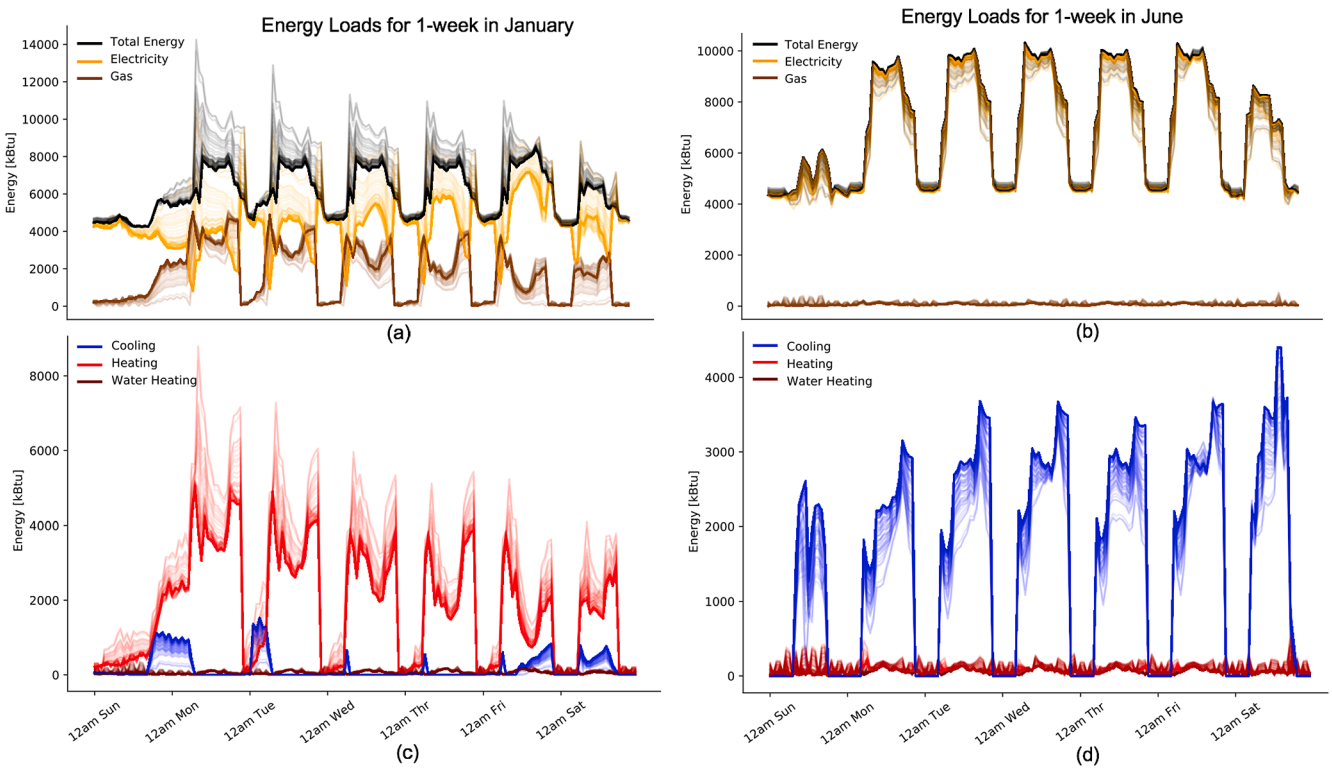


Fig. 11. The calculated hourly energy loads for the Chrysler building in NYC for all 500 simulations. The plots in the left column are for 1-week in January while the plots on the right are for 1-week in June. The plots on the top row show the total energy demand, which is the summation of the electricity and gas demand. The plots on the bottom row show the heating and cooling demand. The gas demand, shown in the top plots, includes energy that comes from NYC's steam-based district energy system.

provide hourly estimates for every building in a city using solely open data. Many of the limitations above could be eliminated by collecting more data on neighborhood context, building geometry, smart meter data from individual buildings, and more; but this data can be very difficult or costly to obtain. Therefore, a major objective of this paper was to build a model using solely open data to provide policymakers with more information on building energy use than they currently have. By incorporating other non-open datasets, the use-cases—as described extensively in Section 6—would no longer hold and the overall usefulness of the model would be diminished. Ideally, more open data would become available so that these other useful types of parameters can be incorporated into our model.

Because we could not validate the model at the single building-level due to data limitations, we took several other measures to address the above limitations. First, by running a Monte Carlo simulation, we examine the variability in results of the model when using different random subsets of buildings. Results from this showed that the results between simulations was fairly consistent and that a small subset of buildings could be used to approximate city-wide consumption; both the train and test errors were low and fit within a narrow range, as shown in Figs. 9 and 10. Second, the distribution of $\boldsymbol{\beta}$parameters, as shown in Fig. 11, shows that over 95% of buildings in NYC receive the same β values across each simulation, suggesting a negligible effect from the manually constructed one-to-three mapping. And third, by using a PSO to assign individual buildings one of the solutions from the Monte Carlo simulation, we showed there is minimal effect arising from the assumption that all buildings within the same PLUTO class have the same energy load curve given that the reduction in error when using the PSO was minimal.

Despite these limitations, this is one of the first attempts at producing synthetic building-level hourly loads for electricity and total energy demand for an entire city, leveraging advantages from both physics-based simulation and data-driven techniques. Future work aims to improve upon the constructed model by expanding the model to other cities across the world where open data is available (e.g., San Francisco, Washington, DC, Singapore, London) and acquiring interval data for several buildings within one of these municipalities to improve validation. This will also allow for us to explore deeper sensitivity analyses between geographic locations and temporal variations in energy loads. Capturing urban context in the proposed model can also be improved upon by acquiring more data on sky view factor, canyon aspect ratio, and building density.

Another avenue for future work is integrating neighborhood morphology and mobility data, such as those from mobile phones to better approximate occupancy levels in buildings—a major source for discrepancies between UBEMs and real-world observations [75]. Extending this further, introducing other mobility information—such as data from electrified fleets of vehicles, trucks, and public transit—can allow for the co-optimization of their energy use, creating a more holistic model that would allow for better planning of sustainable cities and resilient grids.

8. Conclusion

Researchers have acknowledged that one of the largest uncertainties for traditional urban building energy models (UBEM) is the definition and detailed description of archetypes that reliably represent a building stock [76]. Our proposed Augmented-UBEM (A-UBEM) addresses this gap through the design of a *two-step* framework, where step (1) takes annual energy data from a *subset* of buildings in a city and estimates annual energy use for *every* building within the city. In step (2), these energy estimates are converted into hourly load curves through an optimization strategy that constructs profiles for building classes in NYC by fitting a weighted average of possible profiles obtained from the DOE reference building dataset. This allows for greater model flexibility than previous instances using building archetypes, as the optimization

parameters can take a range of weights to appropriately model all buildings within a city. Moreover, we found that the distribution of fitted parameters and, therefore, hourly load curves from a Monte Carlo simulation demonstrates the stability of these results, even when using different subsets of buildings. By only using publicly available data and through the design of our optimization algorithm, our model is generalizable to other cities with energy disclosure policies and open tax assessor databases. Leveraging open data sources provides added value to the public as other researchers can more easily build off our work and easily share the results more widely [77]. As such, the code and data used in this paper is shared on our GitHub page. In short, the A-UBEM model transforms annual energy data from a subset of buildings to hourly energy estimates for every building in a city.

Overall, this work aims to demonstrate the merit of leveraging physics-based simulation modeling, machine learning, and optimization to produce accurate building specific synthetic hourly energy profiles of every building in New York City (1 + million) using only publicly available data. Most importantly, insights from temporally and spatially detailed data can help policymakers, engineers, and planners compare alternative energy systems and programs (e.g., district energy systems, DERs) leading to potential energy efficiency opportunities, lowering of system costs, and reduction in environmental emissions. Urban buildings form the backbone of our major economic centers and represent a significant portion of our energy usage and emissions. Ensuring a pathway to a more sustainable energy future will require deep insights into spatial and temporal distribution of building energy use inside our cities.

CRediT authorship contribution statement

Jonathan Roth: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Formal analysis. **Amory Martin:** Software, Validation, Formal analysis, Writing - review & editing. **Clayton Miller:** Supervision, Funding acquisition, Writing - review & editing. **Rishee K. Jain:** Project administration, Writing - review & editing, Supervision, Funding acquisition.

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.

Acknowledgments

This material is based upon the work supported in part by the Stanford School of Engineering under a Terman Faculty Fellowship, the Precourt Institute for Energy, the National Science Foundation under Grant Nos. 1642315, 1941695, and the Singapore Ministry of Education (MOE) Tier 1 Grant (R296000181133). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The authors would also like to thank Aimee Bailey and Sonika Choudhary, as well as the EDF Innovation Lab, for their support in the development and execution of this study. Computing for this project was partially performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apenergy.2020.115981.

References

[1] Mora C, et al. Global risk of deadly heat. Nat Clim Chang 2017;7(7):501–6.
[2] Aerts JCJH. Connecting delta cities. VU University Press; 2009.
[3] World Bank Group and IHME. The cost of air pollution: strengthening the economic case for action; 2016.
[4] Global Fossil Fuel Subsidies Remain Large: An Update Based on Country-Level Estimates. [Online]. Available: https://www.imf.org/en/Publications/WP/Issues/2019/05/02/Global-Fossil-Fuel-Subsidies-Remain-Large-An-Update-Based-on-Country-Level-Estimates-46509. [Accessed: 05-Dec-2019].
[5] Jain RK, Qin J, Rajagopal R. Data-driven planning of distributed energy resources amidst socio-technical complexities. Nat Energy 2017;2(8).
[6] Denholm P, O'connell M, Brinkman G, Jorgenson J. Overgeneration from solar energy in California: a field guide to the duck chart; 2013.
[7] Hirsch A, Parag Y, Guerrero J. Microgrids: a review of technologies, key drivers, and outstanding issues. Renew Sustain Energy Rev 2018;90:402–11.
[8] Chen Y, Hong T, Piette MA. Automatic generation and simulation of urban building energy models based on city datasets for city-scale building retrofit analysis. Appl Energy 2017;205(July):323–35.
[9] Roth J, Rajagopal R. Benchmarking building energy efficiency using quantile regression. Energy 2018;152:866–76.
[10] Meng T, Hsu D, Han A. Estimating energy savings from benchmarking policies in New York City. Energy 2017;133:415–23.
[11] Map: U.S. Building Benchmarking and Transparency Policies. Institute for Market Transformation; 2017. Available: http://www.imt.org/resources/detail/map-u.s.-building-benchmarking-policies [accessed: 10-May-2017].
[12] Yang Z, Roth J, Jain RK. DUE-B: Data-driven urban energy benchmarking of buildings using recursive partitioning and stochastic frontier analysis. Energy Build 2018;163:58–69.
[13] Roth J, Lim B, Jain RK, Grueneich D. Examining the feasibility of using open data to benchmark building energy usage in cities: a data science and policy perspective. Energy Policy 2020;139:111327.
[14] Miller C, Meggers F. The building data genome project: an open, public data set from non-residential building electrical meters. Energy Procedia 2017;122:439–44.
[15] Roth J, Brown IV HA, Jain RK. Harnessing smart meter data for a Multitiered Energy Management Performance Indicators (MEMPI) framework: a facility manager informed approach. Appl Energy 2020;276:115435.
[16] Dhakal S. Urban energy use and carbon emissions from cities in China and policy implications. Energy Policy 2009;37(11):4208–19.
[17] Bentzen J, Engsted T. A revival of the autoregressive distributed lag model in estimating energy demand relationships. Energy 2001;26(1):45–55.
[18] Brownsword RA, Fleming PD, Powell JC, Pearsall N. Sustainable cities – modelling urban energy supply and demand. Appl Energy 2005;82(2):167–80.
[19] Cerezo C, Dogan T, Reinhart C, Towards standarized building properties template files for early design energy model generation. in: Proceedings of ASHRAE/IBPSA Conference 2014 (Atlanta, Georgia).
[20] Firth SK, Lomas KJ, Investigating CO2 emission reductions in existing urban housing using a community domestic energy model. in: Proceedings of Building Simulation 2009 (Glasgow, Scotland).
[21] Bahu JM, Koch A, Kremers E, Murshed SM, Towards a 3D spatial urban energy modelling approach. in: Proceedings of ISPRS 8th 3DGeoInfo Conference 2013 (Istanbul, Turkey).
[22] Mavrogianni A, Davies M, Kolokotroni M, Hamilton I, A gis-based bottom-up space heating demand model of the london domestic stock. in: Proceedings of Building Simulation 2009 (Glasgow, Scotland).
[23] Xu X, AzariJafari H, Gregory J, Norford L, Kirchain R. An integrated model for quantifying the impacts of pavement albedo and urban morphology on building energy demand. Energy Build 2020;211:109759.
[24] Howard B, Parshall L, Thompson J, Hammer S, Dickinson J, Modi V. Spatial distribution of urban building energy consumption by end use. Energy Build 2012;45:141–51.
[25] Robinson C, et al. Machine learning approaches for estimating commercial building energy consumption. Appl Energy 2017;208:889–904.
[26] Ahmad AS, et al. A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renew Sustain Energy Rev 2014;33:102–9.
[27] Touzani S, Granderson J, Fernandes S. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy Build 2018;158:1533–43.
[28] Jain RK, Smith KM, Culligan PJ, Taylor JE. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. Appl Energy 2014;123:168–78.
[29] Ma J, Cheng JCP. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. Appl Energy 2016;183:193–201.
[30] Wang Y, Gan D, Sun M, Zhang N, Lu Z, Kang C. Probabilistic individual load forecasting using pinball loss guided LSTM. Appl Energy 2019;235:10–20.
[31] Miller C, Meggers F. Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. Energy Build 2017;156:360–73.
[32] Miller C, et al. The ASHRAE Great Energy Predictor III competition: Overview and results. Sci Techn Built Environ 2020;26:1427–47.
[33] Pasichnyi O, Wallin J, Kordas O. Data-driven building archetypes for urban building energy modelling. Energy 2019;181:360–77.
[34] Kontokosta CE, Jain RK. Modeling the determinants of large-scale building water use: Implications for data-driven urban sustainability policy. Sustain Cities Soc 2015;18:44–55.

[35] Silva MC, Horta IM, Leal V, Oliveira V. A spatially-explicit methodological framework based on neural networks to assess the effect of urban form on energy demand. Appl Energy 2017;202:386–98.
[36] Kristensen MH, Hedegaard RE, Petersen S. Long-term forecasting of hourly district heating loads in urban areas using hierarchical archetype modeling. Energy 2020; 201:117687.
[37] Park JY, Yang X, Miller C, Arjunan P, Nagy Z. Apples or oranges? Identification of fundamental load shape profiles for benchmarking buildings using a large and diverse dataset. Appl Energy 2019;236:1280–95.
[38] Miller C. What's in the box?! Towards explainable machine learning applied to non-residential building smart meter classification. Energy Build 2019;199: 523–36.
[39] Roth J, Jain RK. Data-driven, multi-metric, and time-varying (DMT) building energy Benchmarking using smart meter data, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), LNCS, vol. 10863, Jun. 2018. p. 568–93.
[40] Miller C, Nagy Z, Schlueter A. A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. Renew Sustain Energy Rev 2018;81:1365–77.
[41] Zhan S, Liu Z, Chong A, Yan D. Building categorization revisited: a clustering-based approach to using smart meter data for building energy benchmarking. Appl Energy 2020;269:114920.
[42] Xu X, Wang W, Hong T, Chen J. Incorporating machine learning with building network analysis to predict multi-building energy use. Energy Build 2019;186: 80–97.
[43] Fan C, Sun Y, Zhao Y, Song M, Wang J. Deep learning-based feature engineering methods for improved building energy prediction. Appl Energy 2019;240:35–45.
[44] Zhao H, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16(6):3586–92.
[45] Abbasabadi N, Ashayeri M, Azari R, Stephens B, Heidarinejad M. An integrated data-driven framework for urban energy use modeling (UEUM). Appl Energy 2019; 253:113550.
[46] Kontokosta CE, Tull C. A data-driven predictive model of city-scale energy use in buildings. Appl Energy 2017;197:303–17.
[47] Nutkiewicz A, Yang Z, Jain RK. Data-driven Urban Energy Simulation (DUE-S): a framework for integrating engineering simulation and machine learning methods in a multi-scale urban energy modeling workflow. Appl Energy 2018;225:1176–89.
[48] Roth J, Bailey A, Choudhary S, Jain RK. Spatial and Temporal modeling of urban building energy consumption using machine learning and open data. In: Comput. civ. eng. 2019 smart cities, sustain. resil. – sel. pap. from ASCE int. conf. comput. civ. eng. 2019; 2019. p. 459–67.
[49] Papadopoulos S, Bonczak B, Kontokosta CE. Pattern recognition in building energy performance over time using energy benchmarking data. Appl Energy 2018;221: 576–86.
[50] Dahan H, Cohen S, Rokach L, Maimon O. Proactive data mining with decision trees, vol. 81, no. 1. Springer Science and Business Media; 2014.
[51] Shehabi A, Smith S, Sartor D. Lawrence Berkeley National Laboratory Recent Work Title United States Data Center Energy Usage Report: Permalink https://escholarship.org/uc/item/84p772fc Publication Date.".
[52] Deng H, Fannon D, Eckelman MJ. Predictive modeling for US commercial building energy use: a comparison of existing statistical and machine learning algorithms using CBECS microdata. Energy Build 2018;163:34–43.
[53] Burlig F, Knittel C, Rapson D, Reguant M, C. Wolfram C. Machine learning from schools about energy efficiency; 2017.
[54] Tibshirani R. Regression shrinkage and selection via the Lasso. Source J R Stat Soc Ser B 1996;58(1):267–88.
[55] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. arXiv:1510.04342.
[56] Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling; 2003.
[57] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[58] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785–94.
[59] Wei P, Jiang X. Data-driven energy and population estimation for real-time city-wide energy footprinting. In: BuildSys 2019 – proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation; 2019. p. 267–76.
[60] Imbens GW, Lancaster T. Efficient estimation and stratified sampling. J Econom 1996;74(2):289–318.
[61] Boyd S, Vandenberghe L. Convex optimization. Cambridge, UK: Cambridge University Press; 2004.
[62] Diamond S, Boyd S. CVXPY: a python-embedded modeling language for convex optimization. J Mach Learn Res 2016;17(83):1–5.
[63] Kerdphol T, Fuji K, Mitani Y, Watanabe M, Qudaih Y. Optimization of a battery energy storage system using particle swarm optimization for stand-alone microgrids. Int J Electr Power Energy Syst 2016;81:32–9.
[64] Delgarm N, Sajadi B, Kowsary F, Delgarm S. Multi-objective optimization of the building energy performance: a simulation-based approach by means of particle swarm optimization (PSO). Appl Energy 2016;170:293–303.
[65] James L, Miranda V. PySwarms a research toolkit for Particle Swarm Optimization in Python. J Open Source Software 2018;3:433.
[66] Feurer M, Klein A, Jost KE, Springenberg T, Blum M, Hutter F. Efficient and robust automated machine learning 2015. In NIPS*29.
[67] Power Trends: New York's Evolving Electric Grid, New York City; 2017.
[68] Sokol J, Cerezo Davila C, Reinhart CF. Validation of a Bayesian-based method for defining residential archetypes in urban building energy models. Energy Build 2017;134:11–24.
[69] Burke MJ, Stephens JC. Political power and renewable energy futures: a critical review. Energy Res Soc Sci 2018;35:78–93.
[70] Kakran S, Chanana S. Smart operations of smart grids integrated with distributed generation: a review. Renew Sustain Energy Rev 2018;81:524–35.
[71] Lund H, et al. 4th Generation District Heating (4GDH): integrating smart thermal grids into future sustainable energy systems. Energy 2014;68:1–11.
[72] Lo K. A critical review of China's rapidly developing renewable energy and energy efficiency policies. Renew Sustain Energy Rev 2014;29:508–16.
[73] "Stanford Energy System Innovations (SESI) – Sustainable Stanford - Stanford University." Available: https://sustainable.stanford.edu/campus-action/stanford-energy-system-innovations-sesi [accessed: 29-Apr-2020].
[74] Ürge-Vorsatz D, Cabeza LF, Serrano S, Barreneche C, Petrichenko K. Heating and cooling energy trends and drivers in buildings. Renew Sustain Energy Rev 2015;41: 85–98.
[75] Barbour E, Davila CC, Gupta S, Reinhart C, Kaur J, González MC. Planning for sustainable cities by estimating building occupancy with mobile phones. Nat Commun 2019;10(1):3736.
[76] Reinhart CF, Cerezo Davila C. Urban building energy modeling – a review of a nascent field. Build Environ, Feb 2016;97:196–202.
[77] Jetzek T, Avital M, Bjørn-Andersen N. The sustainable value of open government data. J Assoc Inf Syst 2019;20(6):702–34.