Review of data-driven models for quantifying load shed by non-residential buildings in the United States

Author names and affiliations:

Yashvi Malhotra[a]

Ben Polly[b]

Jason MacDonald[c]

Jordan D. Clark[a,*] PhD (Corresponding Author)

a. The Ohio State University
b. National Renewable Energy Laboratory
c. Lawrence Berkeley National Laboratory

*Corresponding Author Contact Information:

221C Bolz Hall

2036 Neil Avenue

Columbus, OH 43210

Phone: (614) 247-8461

Email: clark.1217@osu.edu

Please cite as:

Malhotra, Y., B. Polly, J. MacDonald and J. Clark. (2024) Critical review of models for quantifying load shed by commercial and industrial buildings during demand management events. *Renewable and Sustainable Energy Reviews* 206, 114870. https://doi.org/10.1016/j.rser.2024.114870

Abstract

Shifting and shedding power demand can be cost-effective techniques for grid operators to function reliably and for end users to earn compensation. Grid operators reimburse customers in proportion to the quantity of load shed. Simple data-driven methods are used to quantify this shed, which is the difference between a measured load during the event and modeled “baseline” that would have occurred in absence of the event. These methods have evolved over the years and in many cases have been integrated with building physics, to make them a hybrid between physics based and empirical models. However, there is no comprehensive analysis that provides guidance to building operators, grid operators and researchers in selecting appropriate models based on their specific needs and available data. This work aims to fill this gap by critically assessing the performance of baseline models put forward from the year 2000 through 2023. The literature reviewed includes reports generated by grid operators, reports from national laboratories and academic journal articles.

The work outlines modeling features like the inputs, time-period for modeling, estimation method, adjustments to fine tune the predictions and metrics to evaluate the performance. A comprehensive list of 50 models has been provided. For each model, the study explores the applicability of the model to weather sensitive buildings, variability in the building profile, timing of the event, and whether the building reduces energy consumption before an event. The work identifies the situations in which a particular model works and draws lessons based on evidence of performance. Finally, recommendations to aid decision making in model selection are given.

Keywords

Demand flexibility, demand response, baseline load profile, quantifying load shed, building-grid integration, grid-interactive efficient buildings.

Nomenclature

Abbreviations	Notations
AC	air conditioner
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
C&I	commercial and industrial consumers
CA-ISO	California ISO
CBL	consumer or custom baseline
CDD	cooling degree day
CMTA	California Manufacturers and Technology Association
CP	change point
DADRP	day-ahead demand response program
DOW	day of the week
DRP	demand response program
ERCOT	Electric Reliability Council of Texas
GLD	guaranteed load drop
HDD	heating degree day
HLV	high load variability
HVAC	heating, ventilation, and air conditioning
ISO	independent system operator
ISO-NE	ISO New England Inc.
LLV	low load variability
LV	load variability
medRTE	median of the relative hourly error
MLR	multiple linear regression
MPE	mean percent error
NMBE	normalized mean bias error
NWS	non-weather sensitive
NY-ISO	New York ISO
OAT	outside air temperature
OBMC	optional binding mandatory curtailment
OLS	ordinary least square
PG&E	Pacific Gas and Electric
PJM	PJM Interconnection L.L.C.
PW	piecewise regression
RRMSE	relative root mean square error
RTO	regional transmission organization
SCE	Southern California Edison
SDGE	San Diego Gas and Electric
SLR	simple linear regression
SSE	sum of squared differences
THI	temperature humidity index
TOWT	time of week and temperature
VAV	variable air volume
WM	weather matching
WS	weather sensitive
WSA	weather sensitive adjustment
WWP	wind speed adjusted dry bulb temperature

Indexes and sets	Notations
2P, 3P, 4P and 5P CP	change point models, where the number signifies the number of parameters
i, d	index for input data points, i , d=1,…,n
k	index for hourly temperature in an exponential distribution for two days, k =1,…,48
t	index for 15 min data points in a week, t =1,…,672
w	index for weeks in a year, t =1,…,52

Variables	Notations
$\hat{\epsilon}_-$ (d) and $\hat{\epsilon}_+$ (d)	terms calculated by linear regression on non-DR days for each day "d", - looks back and + looks forward.
$\hat{y}_i$	$\hat{y}_i$ is the predicted load for an account/building.
$\bar{L}$	average load across baseline day-pairs.
Actual Load $_{adj.\ hours}$ and Predicted Load $_{adj.\ hours}$	actual load and predicted load during adjustment/pre-event hours
L	load during the control day-pair period
THI_(DR day) and THI_(non-DR day)	estimate for THI for the peak hours of the event day, and the estimate for the peak hours on non-DR days
y_i	actual load
y_t	average load at time t across all weeks.
$y_{w,t}$	load at time t during the week w

Parameters	Notations
h	starting hour of load shed event
n	number of data points
p	number of independent parameters in the model.

Function	Notations
avg	average
reg	regression
γ	function to remove autocorrelation from the days prior and following the event

1. Introduction

Shifting and shedding power demand can be cost-effective techniques for grid operators to function reliably and for end users to earn compensation. Buildings comprise over 70% of electricity demand in the United States and building demand management is encouraged in various ways by the Federal Energy Regulatory Commission [1], Regional Transmission Organizations (RTO), Independent System Operators (ISO), and utilities across USA. Building demand can be reduced or shaped a number of ways, most of which involve changing thermostat setpoints, but can also include lighting reduction [2], turning off appliances [3], temporary ventilation modulation [4] or reduction [3], sensor-based control [5] and even switching off commercial freezers temporarily [6]. The load shed can either displace a shortfall in generation or assist with grid constraints, e.g., peak demand [7].

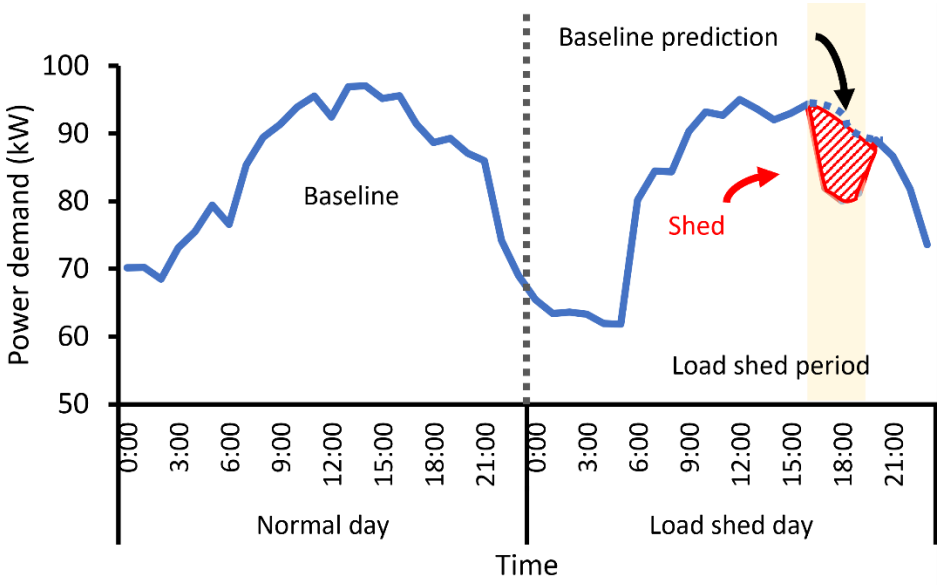

Figure 1: Example time series of power demand in a building and calculated load shed

A load shed event, as shown in Figure 1 in the hatched area, may last from around an hour to several hours [7]. Grid operators reimburse customers in proportion to the quantity of load shed. To quantify this shed, it is necessary to estimate the counterfactual of how much power the building would have used in the absence of the demand shedding event. The actual power used can then be subtracted from this quantity to estimate the shed. There are several methods available for quantifying this counterfactual power demand, which is referred to as the “baseline” from here forward. These methods can have varying levels of accuracy and ease of

implementation. With the increasing need for grid-interactive controls and demand measurement and verification, the selection of methods becomes increasingly important.

No current comprehensive analysis of the multitude of baselining strategies exists. The most comprehensive analysis of various modeling approaches was conducted in the year 2000 [8]. Others conducted in 2011 [9] and 2017 [10], [11] examined many models but only in the context of their applicability in a single region in USA. For these reasons, this work attempts to amalgamate all the current understanding of the performance of all models from different regions of USA that have been put forward for quantifying power demand shedding in a single document and draw lessons as to their performance, domain of applicability, and considerations for their proper use. To meet this need, the objectives of this work are to:

- Provide a comprehensive list of published methods for estimating baseline energy consumption in commercial and industrial buildings in the United States and detail their function, and
- Critically assess the models' performance during load shed events based on published works. This includes the assessment of pros and cons of each approach, domain of applicability, directions in which research and practice are moving, fruitful areas of exploration, and implications for practical application.

This research presents a comprehensive and up-to-date analysis of baseline models used for quantifying load shedding in the USA. It is a repository for best practices in simplified baseline modeling for demand shed events and will support future research directions.

2. Methodology

Baseline estimation has been the subject of many previous works. The current review searched for existing analyses of modeling approaches used in the USA by grid operators. The search was conducted in scientific journal databases as well as national laboratory repositories and grid operators' internally published documents.

In general, buildings can be modeled using physics-based methods [12], [13], data-driven methods [13], and complex methods including machine-learning based methods [14], [15], [16]. However, the vast majority of methods used for quantifying load shed for grid services reimbursement are simplified data-driven methods [8], [9], [17]. These methods are computationally efficient as compared to calibrated simulations [18], [19], [20]. In many cases the simplified methods are integrated with some knowledge of building physics to make a hybrid or semi-physical simplified method [13]. The scope of the current work is therefore confined to these simple data-driven models used by utilities and grid operators in the United States to compensate consumers for events happening on weekdays. It also includes models developed by research labs and energy companies for similar purposes. Section 5 discusses the applicability of this work to locations outside of the United States.

This work does not include weekend-based models, as events generally happen at times when energy demand is high—usually on weekdays for commercial and industrial buildings. It also does not include a review of machine learning algorithms or detailed physics-based modeling as they are rarely if ever used in the applications of interest [18], [19], [20]. This is likely because they require a high level of expertise and computational effort, require inputs that are often unknown for these use cases (e.g., equipment schedules), and have limited applicability in some types of buildings such as those driven by internal loads, e.g., factories. This resulted in exclusion of 7 of 10 models discussed in [21], and 2 of 6 models in [14].

With these constraints, the search returned the 20 works in Table 1. Each of the references in Table 1 examines some aspect(s) of an approach to estimating the baseline power demand and usually focuses on a particular geographical region and climate. In some cases, the performance of a single model was assessed, while in others several models were evaluated for appropriateness for a particular use case. Table 1 shows, for each work, the number and type of baseline models tested, description of the buildings or other loads tested, ISO/RTO territory, climate zone and data used for modeling in each work reviewed. In some works, multiple models were tested to assess the appropriateness of certain models or classes of models for different applications, e.g., weather sensitive buildings.

From these works, the current review extracted the salient features common to the modeling approaches and organizes them below in Section 3. This is done without discussion at first to lay out the problem. Then, Section 4 discusses individual models and their demonstrated performance and domains of applicability. Lastly, Section 5 critically analyzes existing literature on baselining methods and discusses trends and areas for future improvement.

Table 1. Baseline model discussions (2000-2023) in the literature: ‘avg’ denotes average-based models, while ‘reg’ designates regression-based models.

References (Year) and Description	Number of customers and type	Region (and ISO/RTOs)	Climate zones
[8] (2000) compares 18 (‘avg’ and ‘reg’) models used by ISO and utilities for quantifying load shed. The models were selected based on ease of understanding, use, and implementation. The models were evaluated for different customer types and events occurring in both summer and winter.	646 accounts (commercial and industrial C&I)	California, mid-Atlantic, mid-west, north and south USA (CAISO, PJM, and NY ISO)	Dry, warm marine, mixed-humid, mixed-dry, cool humid and cold humid
[22], [23] (2009) tested 7 (5 ‘avg’ and 2 ‘reg’) models developed by ISOs, utilities and consulting companies for quantifying load shed. The authors proposed their own variations to the models. The models were evaluated for events occurring only in the summer.	33 commercial buildings (offices, museums, retail stores, bakery, and detention facility)	PG&E territory, California	Dry, warm-marine, and mixed-marine
[9] (2011) selected 11 (9 ‘avg’ and 2 ‘reg’) models developed by ISOs, utilities and energy companies for quantifying load shed. The models were evaluated for different customer types, customer size and events occurring in both summer and winter.	4,565 DR and 16,002 non-DR C&I customers	Data from PJM service territory; Models from PJM, CAISO, NYISO, ISO-NE and ERCOT	Mixed humid and cold humid
[24], [25] (2008-11) analyzed 6 (‘avg’) models developed by grid operators on the basis of their construct, time periods used for estimation and adjustments used for fine tuning the predictions. The models were evaluated for events occurring in both summer and winter.	306 sites, can be individual or aggregated meter sites	NYISO, ISO-NE, PJM, SCE, PG&E, and ERCOT territory.	Warm-marine and mixed-marine, dry, and humid
[26] (2012) evaluated 5 (1 ‘avg’ and 4 ‘reg’) whole building energy models on granularity of data used for modelling (daily/weekly/monthly) and different training period length used for estimation (6-month, 9 month and 12 month). The models were selected from existing commercial, public domain, and research methods.	29 commercial buildings	California, North Carolina, Washington, Oregon, Colorado, and Idaho	Dry and humid

References (Year) and Description	Number of customers and type	Region (and ISO/RTOs)	Climate zones
[27] (2015) study describes an open-sourced model used by LBNL to establish baseline energy consumption for buildings, The model was evaluated for events in multiple years occurring in summer.	36 commercial building (office, manufacturing, laboratory, retail, museum, jail, and bakery)	PG&E territory, California	Dry, warm-marine, and mixed-marine
[28] (2015) tested the performance of 5 ('reg') whole building energy models, selected from the public domain.	389 commercial buildings	Northern California	Dry, warm-marine, and mixed-marine
[18], [19], [20], [29] (2016) describe 2 ('reg') hybrid whole building energy models for short term measurement and verification in commercial buildings.	40 commercial building (office, school, hospital, and hotel)	Arizona, Texas, Illinois, Canada	Dry and humid
[21] (2016) compares the performance of 10 (3 'reg', others are machine learning algorithms) whole building energy models, selected from existing commercial and research methods.	537 commercial building	California, Northwest and Mid-Atlantic regions of USA	Very hot, hot, warm, mixed, cool, cold, and very cold
[10], [11] (2017) evaluates the performance of 29 ('avg') model used for quantifying load shed. The models were tested for weekend and weekday events occurring in both summer and winter, and different types of customers	104,000 aggregated accounts (C&I, residential, and agricultural), used 100 commercial accounts for simulation	PG&E, SCE, and SDG&E, California	Dry, warm-marine, and mixed-marine
[30] (2019) compares the performance of whole building energy model using smart meter data with utility estimates.	137 commercial buildings	New England	Cool humid
[31] (2021) evaluates 8 models (4 'avg' and 4 'reg') developed by ISOs and research labs for quantifying load shed in commercial buildings.	453 commercial buildings		Marine, cold, and mixed-humid.
[32] (2022) evaluates 5 model (2 'avg' and 3 'reg') developed by either ISOs or commercially used for demand flexibility applications with varying model constructs. The study analyzes events from multiple years.	203 commercial buildings 1)121 retail stores 2) 11 office buildings	11 States	Hot-humid, warm-humid, warm-dry, mild- humid, and cold-humid
[14] (2023) evaluates 6 (2 'avg', 2 'reg', others are machine learning) whole building energy models for demand flexibility applications with varying model constructs. The study analyzes events from multiple years.	120 commercial buildings		Marine and mixed- humid

3. Model features

This section presents the salient features of each the models reviewed, organized by feature. The distinguishing features of the baseline models seen consistently in literature are: 1) input variables(s), 2) time-period and granularity of data, 3) method of estimation, and 4) adjustment(s) used to fine-tune the predictions. Each of these are discussed in the subsections below. In most subsections, a summary table is given with strengths and weaknesses of different modeling choices, and these choices are then discussed in more detail subsequently.

3.1. Input variables.

The most used input variables are summarized in Table *2* and discussed in this section. The impact of the selection of inputs is discussed in much greater detail in Section 5. Baseline models can use 1) weather variables; 2) building variables; and 3) time and calendar variables; or a combination of these, as model inputs for predicting the baseline. Some models use only historical energy data for baseline prediction. A summary of model inputs other than historical power data is shown in Figure 2.

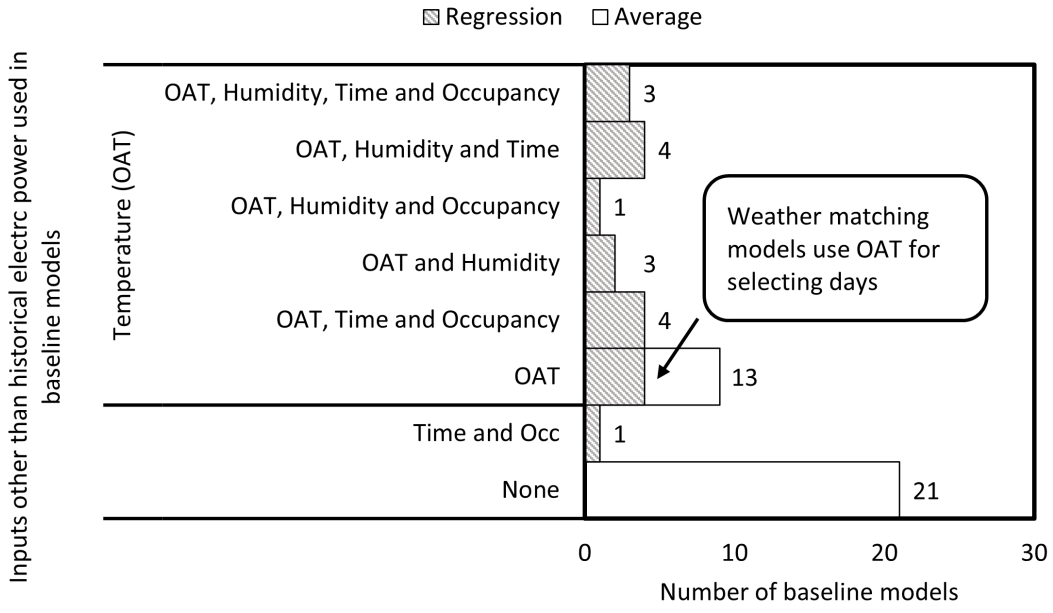

Figure 2: Summary of input variables other than historical power data used in reviewed models.

Table 2: Commonly used model inputs

Category	Sub-category	Model inputs
Weather	Temperature	• Outdoor air temperature (OAT): Hourly dry bulb temperature [33], [34] or aggregated values like average or maximum daily temperature [35], [36], [37]. • Degree day: A measure to quantify the harshness of outdoor climate. [7],[8],[33],[38]– [40]. • Temperature gain variables: Difference between the maximum temperature in the afternoon and minimum temperature in the morning. [33]. • Weather-based day types: e.g., the variable “hotday” =1 when the average temperature is greater than 70°F (21°C), while “coldday” =1 when the temperature is less than 60°F (16°C) [29].
	Humidity	• Dew point, wet bulb temperature [9],[31] and relative humidity [18],[40]. • Difference of outdoor and wet bulb temperature [31],[32]. • Temperature humidity index (THI): A variable that accounts for temperature and humidity [7]–[9], [39].
	Other	• Daylight savings time [33], [35]. • Fraction of dawn and dusk hours [33]. • Wind speed [12], [18], [35]. • Wind speed adjusted dry bulb temperature (WWP): Variable to account for the effect of wind speed and temperature e.g. If wind speed> 10mph (16km/h), WWP=temperature - 0.5*(wind speed-10) [35] • Weather zones: Variable defined for a zone based on multiple weather stations located in it [33].
Building	Structure and operation	• Building type, age and area [26], [32], [38]. • Building schedule and occupancy [9], [28], [36], [38].
	Thermal response of building	• Change or balance point (CP) [33], [35], [38], [42], [43]. • Lagged temperature: Temperature from a previous time e.g., weighted average of degree days or temperature of previous 2 days with weights decreasing exponentially [7], [22], [29].
Time and calendar variables	• Hour [7], [44]. • Month of the year [4]. • Holiday variables [26], [33], [35]. • Day of the week variables [17], [32], [33], [35]. • Season variables: Variable for difference between seasons [33]. • Day of week [19].	

3.1.1. Weather variables

The first category of inputs is weather-related variables. The most used variable by far is outdoor air temperature (OAT), followed by humidity [31], [38], [40]. A few models use degree days instead of OAT to capture the harshness of outdoor climate [17]. ERCOT's regression-based model uses other weather variables like fraction of dawn and dusk hours and time of daylight [33]. In some cases, the weather variables are weighted for an ISO/RTO service territory zone such that the average OAT in a particular zone is the weighted aggregate of OAT from multiple stations, where stations closer to the customer receive greater weight [33], [45]. Many models omit solar radiation as a variable, although it is often a greater driver of thermal dynamics in buildings than humidity. The limited inclusion of solar radiation might stem from challenges in accessing solar data and is discussed in greater detail in subsequent sections.

Some grid operators, such as PJM, utilize wind speed or wind speed adjusted dry bulb temperature for forecasting the demand of its customers instead of quantifying the load shed [35]. Similarly, some hybrid whole building energy models have incorporated wind speed, though not specifically for demand flexibility applications [12], [18], [35]. These inputs have been included in case the reader wishes to assess their performance with their baseline models.

It is well documented that some buildings are not as sensitive to weather variables. These include buildings driven primarily by internal loads such as factories and data centers. For this reason, buildings are often categorized as either weather-sensitive or non-weather sensitive and different models are used to estimate shed in each category of buildings. Table 3 describes the methods used in the literature for classifying these two categories of buildings.

Table 3: Review of methods classifying weather sensitivity (WS) in buildings

Ref	Method for classifying weather sensitivity.
[23]	Spearman rank order correlation coefficient is quantified between load and outdoor air temperature for each hour on admissible days. Cutoff: If the coefficient is greater than 0.7, building is said to be weather-sensitive
[9], [46]	Regress the yearly load with cooling degree hours (CDH) with base temperature as 60°F (16°C), the intercept represents the average non weather sensitive load. The slope represents the increase with each degree increase in temperature above 60°F (16°C). The weather sensitivity (WS) ratio is evaluated between 60 to 90°F (16 to 32°C) and is given as: $$\text{WS Ratio} = \frac{\text{Slope} * (90 - 60)}{\text{Intercept of the equation}} \quad (\text{Eq. } 1)$$ Cutoff: 0.30 or greater.
[8]	Regress the yearly load with hourly degree hour. Cutoff: Weather sensitive if sum of cooling or heating coefficients is positive and the F-statistic for these coefficients is significant at the 0.10 level.
[8]	Fraction of the maximum load or energy used for cooling/heating. Remarks: Such data is not typically available.

3.1.2. Building variables

The second category of inputs includes building variables such as type, age, and floor area. Additionally, building schedules can account for time of operation and occupancy. Alternatively, a few models indirectly integrate buildings' properties into the model using variables like "change point" and "lagged variables". Envelope performance, for example, determines location of the change point to some degree but it is not explicitly accounted for. Similarly, the effects of thermal mass can be captured indirectly by including lagging effects of previous hours.

Several papers discuss the need for accounting for the type of load profile [26], which is often a function of the use type of the building. Similar to the distinction between weather-sensitive and non-weather-sensitive buildings, buildings are often categorized according to their load variability (LV). Highly variable load buildings often have large equipment that is used in a pattern that is not easy to predict, and thus modeling is more difficult. Table 4 describes the ways that load variability has been defined in the literature.

Table 4: Review of methods classifying load variability (LV) in buildings

Ref	Method for classifying load variability
[23]	$\text{Load variability (\%)} = \frac{(\text{Hourly load} - \text{daily load})}{\text{daily load}}$ (Eq. 2) Buildings with load variability greater than 15% are classified as highly variable buildings.
[9]	RRMSE of the residuals of the weather regression for admissible days described in Table *3*. The formula for relative root mean square error (RRMSE) is given below. $\text{RRMSE} = \frac{\sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}}}{\sqrt{\sum_{i=1}^{n} \frac{(y_i)^2}{n}}}$ (Eq. 3) Where y_i is the actual load, $\hat{y}_i$ is the predicted value and n is the number of data points. Cutoff: RRMSE greater than 0.4 was classified as a variable load site.
[8]	Root mean square (RMS) deviation of load from corresponding mean, computed across all load shed hours and normalized by the RMS load during those hours, like RRMSE. Cutoff: 0.29-0.42
[28]	Load variability is computed by determining the average load for each of the 672 15-minute time intervals in a week. For each data point, the squared difference between the load and the average load at that specific time of the week is calculated. The square root of the average of these squared errors defines the LV metric, as expressed by the formula: $\text{LV} = \sqrt{\frac{\sum_{w=1}^{52} \sum_{t=1}^{672} (y_{w,t} - \bar{y}_t)^2}{(52)(672)}}$ (Eq. 4) where $y_{w,t}$ represents the load at time t during the week w, and y_t is the average load at time t across all weeks.

3.1.3. Time and calendar variables

The third category of inputs is time and calendar variables. Including time of day helps to capture the time-dependent variations in energy use [9], [43]. A few models also input the day of week, as occupancy and loads may be greater on weekdays (e.g., in offices). Similarly, variables denoting season are sometimes used, as cooling and heating needs may vary with seasons [33].

3.2. Time-periods

A few modeling decisions must be made regarding time-periods of interest. These include the training period, sampling frequency of the input data used and the prediction period.

The training period should include typical operating conditions of the building like occupancy pattern, population density, schedule for different seasons, and indoor temperature setpoints. The days that are not a good representation of the energy consumption pattern of the building are filtered e.g., previous load shed days, holidays, weekends, off-peak days, and scheduled shutdowns. This gives the “admissible days” that can be input into the model. Some models [23], [26], [47] use additional rules for filtering the “input days”, e.g., only use days that are near the event day in time or have similar OAT [8]. These are called “proxy (event) days” and are used as input instead of the admissible days. Excluding load shed days may pose limitations for buildings with frequent shedding, while omitting holidays and weekends can create temporal gaps between input days and events. Various estimation methods are employed to alleviate these challenges, as detailed in the subsequent sub-section.

The prediction horizon defines the period for quantifying load shed [26]. Models are trained on one set of data and employed on another “unseen” set [27]. How long and how similar this unseen prediction period is to the training period is the subject of some existing work, discussed below. Additionally, studies have explored how the distance from the event day influences predictions, providing insights into model performance variability [14]. Moreover, certain models exhibit varying efficacy depending on the timing of load shed [14], [31], [32], which is explored in Section 5.

Granularity of data also affects model performance e.g., the data can be input in 15-minute intervals, hourly intervals or more [38], [45]. The granularity or the sampling frequency should be uniform throughout the time frame selected for modeling and prediction horizon.

3.3. Baseline estimation method

The third important aspect of the modeling approach is the baseline estimation method. Figure 2 provides an overview of the baseline models reviewed in this work organized by estimation method (regression or averaging). In this section the features of each method have been explained and in Section 5, the strengths and weaknesses and domain of applicability of the two main classes of methods (averaging and regression), as well as the sub-classes of methods within them have been articulated.

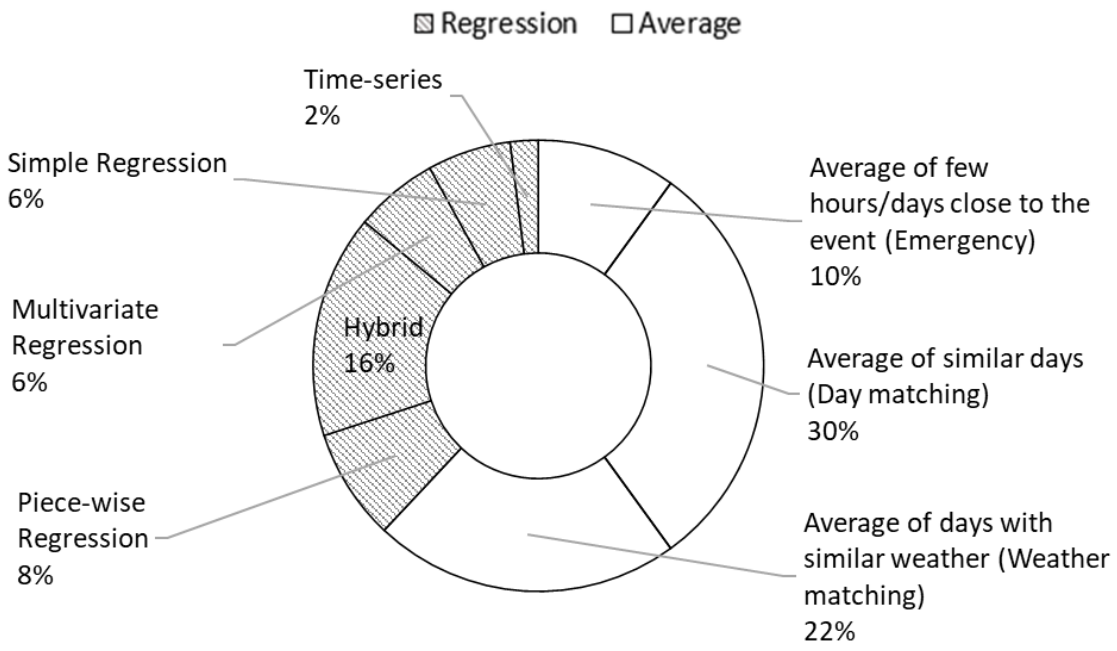

Figure 2: Baseline estimation methods (hatched area represents regression models and unhatched represents average models)

3.3.1. Averaging methods

The simplest method to estimate the baseline is a simple average. The most important choice in selecting an averaging method is the training period, as outlined in Section 3.2. Day-matching models include “High X of Y” [10], [11], [25], [31] where X proxy days with the highest load are selected from Y admissible days; and “Middle X of Y”, where data from moderate weather days is used [25]. In some cases, proxy days refer to a collection of “day-pairs”, which are a pair of 2 days from the preceding year that closely match the “event day-pairs” [8], [9], [10]. On the other hand, weather matching models select proxy days based on weather conditions [10], [11], e.g. days with the daily maximum temperature or days that sufficiently match the cooling degree hours (CDH) [31]. Some models sort the admissible days based on CDH, and then select the top 25% as proxy days [26]. The estimation method then can be a simple average of the hourly data or a weighted average giving more influence to the recent days. Some grid operators refer to averaging methods that use data from the previous 1-5 days or post-event days for baseline estimation as "emergency methods” [9], [48]. This work adopts the same definition.

3.3.2. Regression methods

Beyond simple averaging methods, another class of models uses regression approaches. Linear regression is the simplest regression method and can use single or multiple inputs for prediction. This approach is also referred to as the "energy signature curve". Piecewise regression improves the results by including two linear regression models for temperature below and above the balance point(s). For buildings that have heating and cooling, there will be separate balance points for heating and cooling [17], [34], [37]. However, simple linear regression does not consider the time dependency in the building load data. Advanced approaches like time-series modeling captures autocorrelation and temporal variations in the data (e.g., the effect of past values on future values) and is designed for forecasting [8], [49], [50].

3.4. Adjustments

Baseline models are adjusted to improve predictions. This is also referred to as continuous model calibration. The observed load is compared with the baseline load a few hours before and/or after the event, to calculate a scalar or additive adjustment for the prediction shown below.

$$Predicted\ Load\ after\ adjustment = Scalar\ Adjustment * Predicted\ Load\ (Eq. 5)$$

$$Predicted\ Load\ after\ adjustment = Additive\ Adjustment + Predicted\ Load\ (Eq. 6)$$

A scalar adjustment is a multiplicative adjustment which increases or decreases the prediction by a ratio based on percentage comparison, whereas the additive approach is based on absolute demand difference [25]. The adjustment values can be time-based, such as one hour before the event (h-1) or one hour after the event (h+1). They can be capped in terms of magnitude, e.g., within 80-120% of the (h-1) value, or direction, i.e., symmetrical (baseline adjusted up and down) or asymmetrical (baseline only adjusted up) [24], [25]. Some adjustments are calculated using regression which may or may not be based on outside weather. Those based on weather are called weather sensitive adjustment (WSA) [46]. The different type of adjustments used in the literature are summarized in Table 5.

Table 5:Adjustments used in the literature. “DR” refers to demand response.

Type	Ref.	Adjustment	Formula
Scalar load based	[4]–[6], [29], [33], [41]	Capped (0.8 and 1.2) and based on either first 3 of previous 4 hours or first 2 of previous 3 hours. It can also be called morning adjustment when it is based on the previous 2 hours.	$\text{Adjustment} = \frac{\text{Actual Load during adj. hrs}}{\text{Predicted load in adj. hrs}}$ (Eq. 7) The hours are pre-event hours.
	[10], [11], [31]	Capped and based on 2 hours before load shed and 2 hours after load. The limit can be +/- 20% or +/- 40%	$\text{Adjustment} = \frac{\text{Actual Load during adj. hrs}}{\text{Predicted load in adj. hrs}}$ (Eq. 8) The hours are a combination of pre and post event hours.
Scalar regression based	[8], [51]	Baseline counterparts of control day-pair loads are subjected to regression analysis. The adjustment is used when R sq is greater than 0.5.	$L = a + b\bar{L}$ (Eq. 9) L represents the load during the control day-pair period and $\bar{L}$ is the average load across baseline day-pairs.
	[27]	Two terms $\hat{\epsilon}_{-}(d)$ and $\hat{\epsilon}_{+}(d)$ are calculated by linear regression on non-DR days. For each day “d”, - looks back and + looks forward. γ is the function to remove autocorrelation from the days prior and following the event.	$\widehat{\epsilon_{-}}(d) = \gamma_{-s} \in (d_{-})$ (Eq. 10) $\widehat{\epsilon_{+}}(d) = \gamma_{+s} \in (d_{+})$ (Eq. 11) $\text{Adjustment} = \frac{1}{2}(\hat{\epsilon}_{-}(d) + \hat{\epsilon}_{+}(d)$ (Eq. 12)
	[8], [9]	A model is fitted to the load as a function of THI. The ratio of the estimate for THI for the peak hours of the event day, and the estimate for the peak hours on non-DR days, is the adjustment.	$\text{Adjustment}=\text{THI\_(DR day)}/\text{THI\_(non-DR day)}$ (Eq. 13) This adjustment is weather sensitive (WSA)
Additive load based	[46], [33]	Can be based on either first 3 of previous 4 hours or first 2 of previous 3 hours.	$\text{Adjustment} = [\text{Actual Load} - \text{Predicted Load}]_{(\text{adj. hours})}$ (Eq. 14)
Additive regression based	[52]	Piece wise regression (reg) on OAT- day types and hour load where load reductions are expected.	$\text{Adjustment} = [\text{reg}_{\text{event OAT}} - \text{reg}_{\text{non-event OAT}}]$ (Eq. 15) This adjustment is weather sensitive (WSA)

4. Models

This section compiles models employed by utilities, grid operators, and national laboratories in the United States for quantifying baseline power demand. The models are divided based on estimation method into Table 6 and Table 7. These tables provide details on data selection methods, adjustments, and use cases.

Note that naming conventions vary among models in Table 6 and Table 7. For instance, the second “10” in "High 10 of 10" means either 10 closest days, 10 highest energy days, or 10 days with OAT > or 80°F (27°C) depending on the study. Some studies further filter data based on time, e.g., using data only from 12-6 pm or 12-9 pm. All these nuances have been summarized in Table 6 and Table 7. It should be noted that models used for weekends, or employing machine learning, have been excluded in keeping with the scope of this work. In addition, some models discussed in Table 1 were similar and thus have been grouped together in Tables 6 and 7. Therefore Tables 6 and 7 have fewer entries than Table 1.

In Table 6 and Table 7, “use cases” refer to the applicability of each model to either weather-sensitive (WS) vs. non-weather-sensitive (NWS) buildings, or to low load variability (LLV) or high load variability (HLV) buildings. Methods for categorizing these buildings as such are summarized in Table 3 and Table 4. Similarly, “adjustments”, customized per use case, are also detailed in Table 6 and Table 7.

Table 6: Models based on averaging arranged in the increasing order of time-period used. The abbreviations for ISOs and RTOs have been defined in the nomenclature. Use cases include weather-sensitive (WS), non-weather-sensitive (NWS) buildings, low load variability (LLV), high load variability (HLV) or all of them. Abbreviations for independent parameters and adjustments have been described in Section 3.

	Model	**Data used**	**Adjustments**	**Weather sensitivity**	**Load variability**	**Ref.**
A.1	PJM emergency GLD (guaranteed load drop)	Average of loads in the 2 hours preceding and the 2 hours following the event hour.	Use only pre-event hours and can be additive or scalar.	Both	Both	[9]
A.2	PJM emergency energy settlement	Average load in the hour preceding the start of the event.	Use only pre-event hours and can be additive or scalar.	Both	Both	[9]
A.3	PJM emergency GLD for non-weather sensitive customers	Two days, one day before or one day after an event, which is closest after exclusions. If there is a tie, previous day is chosen.	Additive or scalar. Unadjusted works better for this model.	Non-weather-sensitive	Both	[9]
A.4	PJM emergency GLD (switches to regression for weather sensitive customers)	From the full season select 1 day. The model uses Temperature humidity index (THI) to select the day closest to the event day in terms of weather. The weather terms are retained if significant otherwise dropped.	Additive or weather sensitive adjustment. Scalar adjustment inflates the bias.	Both	Both	[9]
A.5	ISO-NE Emergency model [36]	Simple average of 1 day prior for existing buildings and weighted average of 5 recent days for new buildings.	Additive, scalar or weather sensitive adjustment.	Both	Both	[9]
A.6.	High 3 of 3	Average 3 of last 3 eligible days. Giving weight to the recent days improves predictions.	Scalar adjustment, based on pre-and post-event hours.	Non-weather-sensitive	Low	[10]
A.7	High 3 of 5	Average 3 of last 5 eligible days. Giving weight to the recent days improves predictions.	Capped scalar adjustment based on the combination of pre and post event hours.	Weather-sensitive	High.	[10]
A.8	High 4 of 5	Average 4 of last 5 eligible days. Giving weight to the recent days improves predictions.	Capped scalar adjustment based on the combination of pre- and post-event hours.	Weather-sensitive.	High	[10]

	Model	Data used	Adjustments	Weather sensitivity	Load variability	Ref.
A.9	PJM Economic Customer baseline (CBL), High 4 of 5	From 2 months, select 4 days with the highest energy of 5 most recent admissible days.	Additive or scalar. Weather sensitive adjustment and unadjusted models can inflate the bias.	Both	Both	[9]
A.10	High 5 of 5	Previous five business days.	No adjustment.	Both	Both	[9]
A.11	Middle 4 of 6	Select 4 days by dropping the highest and lowest kWh days from 6 recent days. Preferred over “Middle 8 of 10”, as it uses shorter set of recent days.	Additive or scalar. Weather sensitive adjustment and unadjusted models can inflate the bias.	Both	Both	[9]
A.12	High 3 of 10- Day matching	Select 3 days closest to the event	No adjustment	Both	Both	[10]
A.13	High 3 of 10 – Weather matching	From 1 year, select 3 days with outdoor air temperature (OAT) > 65°F (18°C).	Morning adjustment.	Both	Both	[23]
A.14	High 5 of 10 NYISO Day-Ahead Demand Response Program (DADRP) 2001-2002/ PJM economic load response 2002	From 1 month, select 10 admissible days excluding the days with accepted bids, while ensuring usage exceeds 25% of the previous month's peak hourly load. From these, choose the 5 days with the highest usage, maintaining the 25% threshold.	Adjustments can be additive, scalar or temperature humidity index adjustment. Unadjusted model tends to underestimate.	Both	Both	[8]
A.15	High 5 of 10- Weather matching	Use the data from May-October, sort them based on cooling degree hour (CDH) with 65°F (18°C) as change point (CP). From the top 25 %, select the highest 5 days.	Morning adjustment tends to make the model underestimate.	Both	Both	[23]
A.16	NY-ISO Standard Customer baseline (CBL)- High 5 of 10	From 1 month, select 10 days starting 2 days before the event day excluding event and low usage days. From these, select the highest energy days.	Weather sensitive and highly variable loads should use a capped, scalar adjustment based on the combination of pre and post event hours.	Both	Both	[9]
A.17	ERCOT 2002 Middle 8 of 10	Select 8 days by dropping the highest and lowest kWh days from 10 high energy days.	Additive adjustment for weather sensitive and loads with low variability.	Both	Both	[9], [33]

	Model	Data used	Adjustments	Weather sensitivity	Load variability	Ref.
A.18	High 10 of 10, California Manufacturers and Technology Association (CMTA) proposed Optional binding mandatory curtailment (OBMC) and CA-ISO CBL 2001	Select 10 days closest to the event.	Can have additive or scalar, based on either pre or post hours. Temperature humidity index (THI) adjustment is recommended for weather sensitive buildings.	Both	Both	[9], [31]
A.19	High 10 of 10 ISO-NE 2001- 02	Exclude days with energy levels less than 75% or greater than 125% of the average of a provisional baseline for four or more consecutive hours. From the remaining admissible days, select 10	Can have additive and capped scalar adjustment. The scalar adjustment addresses the problem of zero baseline due to zero usage before curtailment.	Both	Both	[8]
A.20	High 10 of 10- Weather matching	Use the data from May-October, sort them based on cooling degree hour (CDH) with 65°F (18°C) as change point (CP). From the top 25 %, select the highest 5 days. Another study uses 1 year data and selects 10 days with OAT > 80F (26°C). These days could be filtered further to extract the data only from 12-6pm or 12-9pm	No adjustment recommended for events happening at the start of the week and buildings that pre-cool. Morning adjustment, tends to underpredict for buildings with variable load profiles that are not weather sensitive.	Both	Both	[23], [32]
A.21	High 10 of 11, CAISO Demand Response Program (DRP)	10 days with the highest energy consumption from the 11 days prior to the load shed day.	No adjustments recommended.	Both	Both	[8]
A.22	High 3/5/10 of 20	From 1- 2 months, select 20 days with high energy consumption. From these, select the highest a) 3 or b) 5 or c) 10 days. One can	Scalar adjustment.	Both	Both	

	Model	Data used	Adjustments	Weather sensitivity	Load variability	Ref.
		also select the last 20 days [8].				
A.23	EnerNoc	Use the data from May-October, sort them based on CDH with 65°F (18°C) as CP. From the top 25 %, select 20 days (weighs recent days more) with OAT > 65°F (18°C)	Morning adjustment tends to underpredict.	Both	Both	[23]
A.24	Demand based match day	From 3 months, select 20 days with temp >75°F (24°C).	Additive capped (h-1 and h-2). Unadjusted model tends to increase the bias.	Both	Both	[8]
A.25	Match 3 days from 3 months	Sort 3 months data on a) maximum OAT, or b) degree hours, and select 3 days.	Scalar capped adjustment.	Weather sensitive.	Both.	[10]
A.26	Match 4 days from 3 months	Sort 3 months data on a) maximum OAT, b) degree hour or c) degree day.	Scalar capped 20% (h-3 and h-4)	Weather sensitive.	Both.	[10], [31]
A.27	Match 5 days from 3 months	Sort 3 months data on a) maximum OAT, b) degree hour or c) degree day and select 5 days.	Scalar capped adjustment.	Both	Both	[10]
A.28	Binning	Create 3 bins for temp less/equal/greater than 80°F (27°C). Sort data based on a) maximum OAT, b) degree hour or c) degree day. Baseline is the average peak period load on non-DR days in a bin.	Scalar capped adjustment.	Both	Both	[10]
A.29	Pulse adaptive model	1-year, weighted average. Predicts monthly or weekly quantity better than daily. Captures temporal periodicity.	No adjustment. This is a proprietary algorithm.	Both	Both	[26]
A.30	Matching Model	Identify a control day-pair and from the preceding year, select 10 matching day-pairs with lowest sum of square errors. The model selects multiple days to calculate the final baseline.	Adjustments can be additive or regression-based	Weather sensitive.	Both.	[8], [33]
A.31	Weather matching method	From full season, select days with temperature greater than the pre-defined limit based on local weather conditions.	Additive adjustment based on h-1 and h-2. Unadjusted model tends to underpredict.	Both	Both	[8]

Table 7 : Models based on regression arranged in the increasing order of time-period used. “SLR” is simple linear regression, “MLR” is multiple linear regression, “PW” is piece-wise regression and “OLS” is ordinary least square. The abbreviations for ISOs and RTOs have been defined in the nomenclauture. Use cases include weather-sensitive (WS), non-weather-sensitive (NWS) buildings, low load variability (LLV), high load variability (HLV) or all of them. Abbreviations for independent parameters and adjustments have been described in Section 3.

#	Model	Data used	Adjustments	Weather sensitivity	Load variability	Ref.
R.1	Hybrid model [34]	Annual utility bill data, 2-week data of electricity from swing months, OAT, specific humidity potential and/or occupancy. Variation of ASHRAE 3P (three parameter) change point model.	No adjustments.	Both.	Low	[13] – [15]
R.2	MLR: KEMA Customer baseline (CBL) [8]	Select 20 days for regression from a month. Time variables, temperature humidity index (THI) and interaction between them.	Time-based, either additive or scalar. Unadjusted models tend to inflate bias.	Both	Both	[9]
R.3	OLS: Degree Day	1-2 months, degree day used as input. The model works well with daily data as well. It performs better when the full seasons data is used. Base temperature is 65°F (18°C) for both heating and cooling.	Additive or scalar based on h-1 and h-2. Unadjusted models tend to inflate bias.	Both	Both	[8]
R.4	OLS: Lagged temperature	Summer data from Jun-Sep and time 5-10 am. Lagged temperature is used as input. $$\text{Lagged Temp} = \frac{\sum_{k=1}^{48} OAT\ e^{-k/48}}{\sum_{k=1}^{48} e^{-k/48}} \quad (\text{Eq. } 16)$$	Additive or scalar, based on h-3 and h-4. Unadjusted models tend to inflate bias.	Both	Both	[8]
R.5	SLR: OAT or THI	1-2 months, OAT or THI can be used as input. $$THI = \text{OAT} - 0.55 * \left(1 - \frac{\text{RH}}{100}\right) * (\text{OAT} - 58°\text{F}) \quad (\text{Eq. } 17)$$ The data is filtered to select on days with OAT > 58°F (14°C)	Additive or scalar based on h-1 and h-2. Unadjusted models tend to inflate bias.	Both	Both	[8]
R.6	Time series	Summer data from Jun-Sep, with weather variables.	No adjustments.	Both	Low	[8]

#	Model	Data used	Adjustments	Weather sensitivity	Load variability	Ref.
R.7.	Hybrid: Day temperature and time [34]	1-2 years, OAT, time of day, day of the week and 2 CP (50 and 65°F or 10 and 18°C). Combination of ASHRAE 5P CP and OLS.	No adjustments.	Both.	Low.	[26], [28]
R.8	PW: Change point model 2P [34]	1-2 years, OAT, two variations, a) all data is used for modeling, b) Filter the days based on average or maximum daily temperature and select either 10 or 30 days. Selecting filtered days works better than using the entire season for training. Recommended for buildings that require year-round heating or cooling.	Capped scalar (h-1 and h-2). Can also use morning adjustment. Unadjusted models tend to inflate bias.	Both	Both	[23], [32]
R.9	PW: Change point model 3P [34]	1-2 years, OAT, and day of week. Filtered four weeks of data, 12-9pm with OAT>80ºF (27°C). Recommended for buildings using electric AC or gas for heating.	No adjustments.	Both	Both	[32]
R.10	Hybrid model CDD-HDD [34]	1-2 years, OAT, and annual utility bill. Predicts monthly energy usage as a function of Cooling degree day (CDD)=55°F (12°C) and Heating degree day (HDD)= 65°F (18°C). Variation of 3P CP.	Adjustment should be scalar and time-based.	Both	Both	[28]
R.11	PW: Change point model 4P	1-2 years, OAT, and can have more inputs. Has better goodness of fit than 3P models for appropriate cases. Recommended for buildings with VAV HVACs (Variable air volume heating, ventilation, and air conditioning system).	No adjustments.	Both	Both	[34]
R.12	PW: Change point model 5P [34]	1-2 years, OAT, CP determined by optimization and 4°F apart. Can have more inputs. Recommended for buildings with simultaneous heating and cooling.	No adjustments.	Both	Both	[26]
R.13	MLR [17], [19], [34], [53]	Variable, OAT and can use up to six inputs. One variation is OAT with day of week (OAT+DOW) using four weeks of admissible days before the event. Recommended for commercial buildings	No adjustments.	Both	Low	[28], [32]
R.14	MLR: ERCOT Customer	1 year, weather, and calendar variables. The model consists of two equations, daily energy, and 24-hourly energy fraction equation.	Adjustments can be additive, scalar or weather sensitive	Both	Both	[9]

#	Model	Data used	Adjustments	Weather sensitivity	Load variability	Ref.
	baseline (CBL) [33]		adjustment. Unadjusted models tend to inflate bias.			
R.15	MLR: Mean week model	From 1-year, average data for each day of the week and hour. Separate model for occupied and unoccupied period. The model works better with longer modeling periods. This is a simple mean only model.	No adjustments.	Weather sensitive	Low	[21], [26], [28]
R.16	Hybrid: TOWT Model with five change points [54]	1 year, OAT, time of week, and five CP (45, 55, 65, 75, 80°F or 7, 12, 18, 24, 27°C). It is a combination of mean week and change point model. Separate model for occupied and unoccupied modes. If occupancy is not recorded, it is determined based on the electric load profile of the building. Captures weekly periodicity and intra-day temperature dependence.	No adjustments.	Both	Both	[21], [26]
R.17	Hybrid: TOWT Model (7-day baseline) [55]	Like R.15 but uses 7 weekdays for regression. It is recommended to use 5 or 10 weekdays, as the arrangement creates uneven representation for 2 weekdays.	Capped scalar (h-3 and h-4). Unadjusted model increases the bias.	Both	Both	[21], [31]
R.18	Hybrid: Weighted TOWT (70-day baseline) [55]	Like R.15 but uses 70 weekdays for regression and weights them. Two variations, 1) Weight 14 days and 2) Weight 10 days. The weights decrease as the distance from the central day increases in both directions (before and after the event day). No major improvement observed by including a longer baseline.	No adjustments.	Both	Both	[21], [31]
R.19	Hybrid: TOWT model with two CP [27]	1-2 years, OAT with 2 CP, and day of the week. The CP should be at least 2.2°C (4°F) apart. and a minimum of 10% OAT measurements should exceed the first change point and a minimum of 10% should be below the second change point.	Regression based adjustment that removes auto-correlation from the data.	Both	Both	[28]

4.1. Reported model performance.

To quantitatively evaluate model performance, the normalized mean bias error (NMBE) [10], [11], [26], [31], [32] and/or the median relative error (medRTE) [8], [23] for averaging and regression models is illustrated in Figure 3 and Figure 4, respectively. The formulas for these metrics are shown below. For a comprehensive overview of metrics used by building modelers in the literature, refer to [56]. Figure 3 and Figure 4 only show the best adjustment, and the effect of adjustment is discussed in Section 5.

$$\text{Mean percent error (MPE) or normalized mean bias error (NMBE)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\bar{y}} * 100 \quad \text{(Eq. 18)}$$

Some models have n-p (where p=1) instead of n. "n" is the number of data points and "p" is the number of independent parameters in the model.

$$\text{Median of the relative hourly error (medRTE)} = \left(\frac{y_i - \hat{y}_i}{y_i}\right) \quad \text{(Eq. 19)}$$

y_i is actual load and $\hat{y}_i$ is the predicted load for an account/building.

It should be noted that these values are collected from multiple studies, which used different C&I building types, different numbers of buildings/customers for evaluation and different climatic conditions. In some cases, the studies reported the model accuracy for multiple accounts aggregated together instead of individual accounts [10]. Further, the objective of some of the studies may not to be limited to evaluation of different baseline models, but also to analyze the impact of different model elements on the accuracy of the models [31]. Also, some studies performed data cleaning and removed buildings/accounts that were erroneous [9]. As a result, if the reader were to compare the number of buildings reported in Figure 3 and Figure 4, with those reported in these studies or Table 1, they will find some variation.

The effect of climatic conditions on the performance is not shown. But it was observed that the range of errors were wider for studies conducted in humid areas [9], [31] as compared to dry conditions [10], [21], [23]. The studies [8], [19], [26], [32] provide a wider range of climatic conditions and can be referred to for a mixture of climatic conditions.

Figure 3 shows, in general, that adding more information to averaging models generally leads to

improved model performance, among the models reported in the literature. While fluctuations occur based on the value of X in "High X of Y models," detailed discussions on these variations are presented in the subsequent section. In weather matching methods, greater specificity offered by the selection of days with similar weather is associated with improved performance. It should be noted that one of the studies [10], had very few readings for extreme summer afternoons, and most of reported errors were in the range of 0 to 0.001. While it appears to be performing very well, it is advised to have more references before drawing conclusions.

Figure 4 appears to show that for regression-based models adding more information decreases model performance. The reader is cautioned that several explanations for this apparent phenomenon may exist. While some degradation in model performance may be due to over-parameterization, other factors likely contribute as well. These factors are discussed in detail in Section 5.

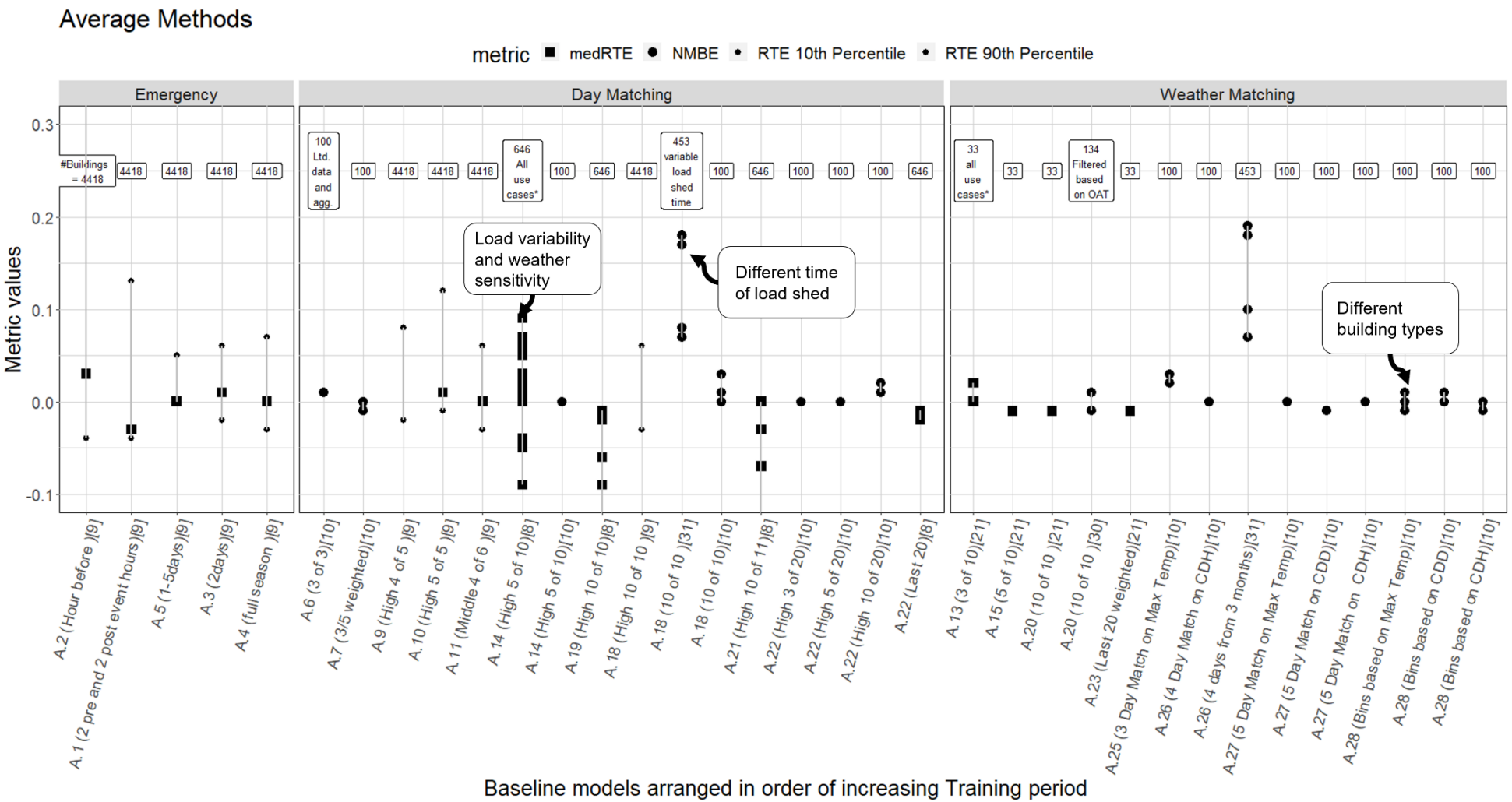

Figure 3: Reported performance of averaging models, generally organized in ascending order of training period and model complexity (reasons for multiple values provided in the callouts).

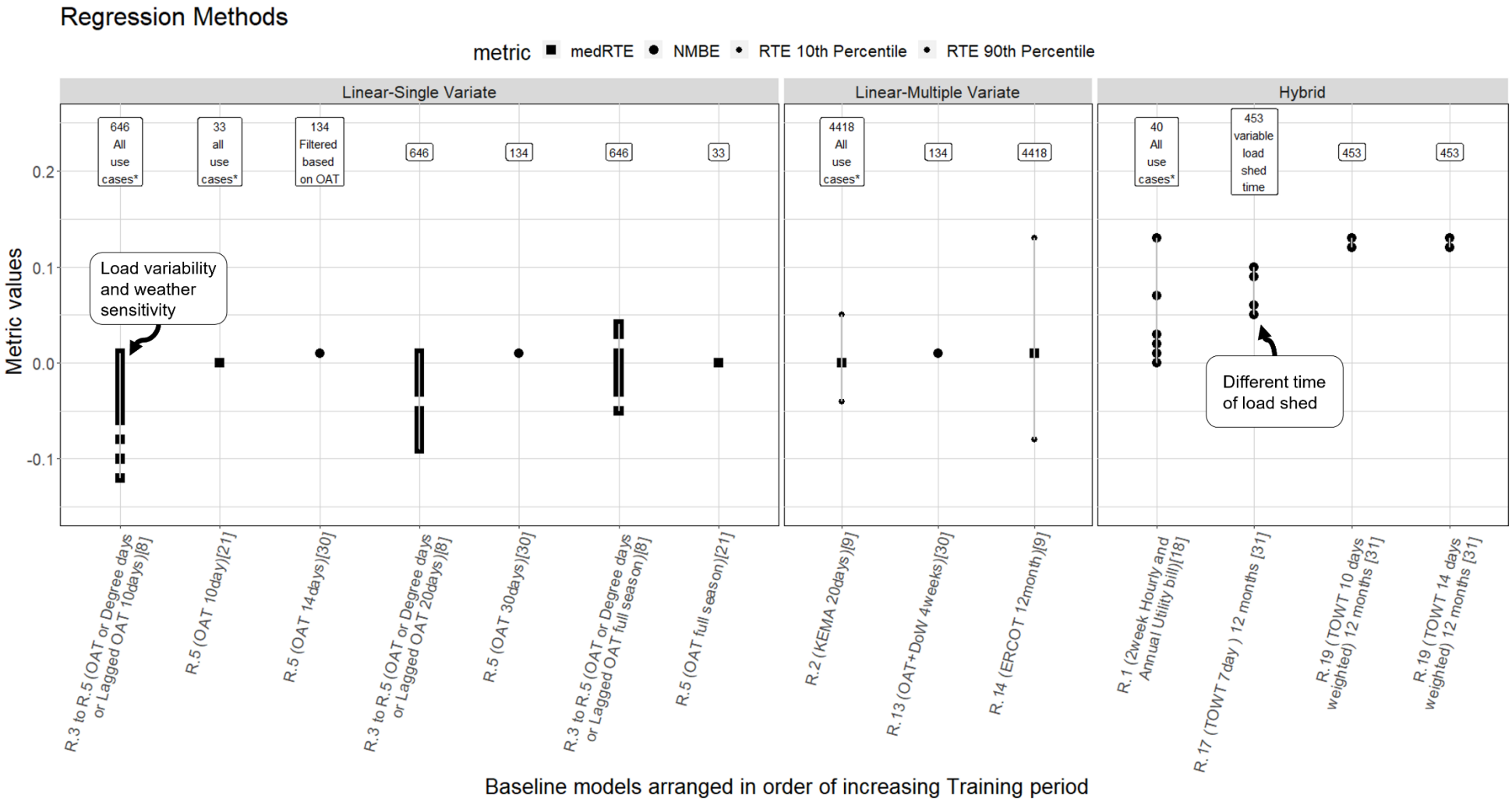

Figure 4: Reported performance of regression models, generally organized in ascending order of training period and model complexity (reasons for multiple values provided in the callouts).

5. Discussion

This section analyzes the different modeling elements and provides recommendations from the literature along with discussions about their application.

5.1. Input variables.

5.1.1. Weather variables

As seen in Section 3, temperature is used for identifying input days in weather matching models and as a weather variable for most regression-based models. This is done for a few reasons, including OAT's physical relationship to conduction and infiltration loads, and because of the high correlation between OAT and other weather variables that influence building energy consumption such as humidity and solar radiation [8], [34]. Table *7* illustrates that ISO/RTOs generally prefer to use OAT as the sole input, which has been shown to be sufficient in many cases. Ref [45], [57], [58], [59], [60] mention that most researchers use OAT as the single independent variable for predicting energy in commercial buildings.

Some models use degree days instead of OAT, as an input for models to normalize the effect of weather. For example, the Xenergy OLS regression model uses both heating and cooling degree days (Model R.3) [17], [37], [49], [61]. However, there is evidence that predictions can suffer if the balance point is not properly selected [17], [61]. This can be improved by choosing a reference temperature after regressing load data with OAT. It reduces the systematic error and is suitable for weather sensitive buildings but can be complex to implement and requires long term data [8]. This is a perennial problem in simplified changepoint modeling and would benefit from development of quick algorithms for determining balance points that do not require multi-linear regression with a great deal of data, as do current methods.

Ideally, a robust model would include other relevant weather variables such as solar radiation, outdoor humidity, wind speed, etc. [8], [34], [49], but these are often omitted for a few reasons. First, some of these variables are co-linear with OAT and assigning coefficients to these variables in regression models presents statistical challenges. ASHRAE Guideline 14–2014 [34] explicitly recommends against using both OAT and humidity in a model because the collinearity can affect model stability and reduce the predictive accuracy. This can also be seen in Figure 4, where simple OAT models had similar or even better performance than multivariate models.

However, excluding humidity from an OAT univariate model can lead to heteroscedasticity. One study [62] found that the model showed large errors at greater temperatures when the influence of humidity was high. A way around this issue could be to do a factor analysis for all the weather

variables and drop those which are not correlated to the energy data [63]. Another work around is illustrated in the PJM emergency GLD model for WS customers (Model A.4) [9], [31], which uses both OAT and RH as model inputs, discussed in Table 6. In this model, the data is regressed with the THI. If the coefficients are significant, the models use a THI-based regression method. Otherwise, it averages the days with least sum of squared differences between THI values, for the event and admissible days.

One glaring lack of input variable is solar radiation, as many works have shown that often building thermal dynamics are driven primarily by solar radiation [38], especially for more contemporary construction with large glazing areas [64]. This lack is almost certainly caused by a lack of measured radiation [65]. While most weather stations measure temperature, wind speed, and humidity and thus these data are ubiquitous, much fewer have the multiple expensive pyranometers needed to properly characterize solar radiation. It is expected that as data-sharing becomes easier because of greater connectivity, this variable will be more widely available, and researchers should look to include it in baseline models. However, an additional obstacle to doing so is solar radiation's collinearity with OAT. More advanced statistical techniques that manage to avoid putting unnecessary burden on modelers would be beneficial in solving some of the issues articulated above.

5.1.2. Building variables

High-level metadata such as age and floor area are sometimes available. For the most part, this information has only been used to normalize meter data, or benchmark buildings against other similar buildings [32]. The analysis did not find any instances of this information being used to improve model performance, suggesting a possible area for improvement as has been done previously for detailed physics-based models [66]. Again, as this data becomes more available it may be beneficial to incorporate it.

Occupancy improves models with hourly or sub-hourly temporal resolution. While most physics-based approaches and calibrated simulation require occupancy schedules, it is less common in the literature on baseline models [62], likely because it is typically unavailable. A few models have attempted to use occupancy as a variable by fitting a regression model on the metered data and defining a surrogate occupancy variable when residuals are positive (Model R.15) [26] [32], [33]. Another model derived the occupancy variable from the lighting and equipment load of the building (Model R.1) [18], [19], [20], while a few defined the variable based on visual inspection [45], [54].

However, the impact of including occupancy in the studies reviewed is not as effective as one would expect, as shown in Figure 4. This could be because the occupancy model is lacking, and perhaps even introducing additional error. To help address this, many means of measuring occupancy are being developed and deployed, such as CO_2 measurement, Wi-Fi access point device counting, infrared sensing, and photogrammetric methods [67]. It is expected that this variable will be more available moving forward and can be instructive for baseline models. It should be noted that occupancy patterns often change during demand response events (e.g., occupants are told to go home early) and this must be considered.

Discussion about including "business operations" is also found in the literature [32]. Building operations refers to the changes in the building schedule and changes made to the building equipment, e.g., retrofitting lights with energy efficient alternates. Though these variables affect building performance, they do not have a standard definition yet, nor have they been used in any baseline models [32]. One area of exploration that may prove fruitful is in expanding and updating the limited database of typical load profiles defined by building type [68].

Other properties of buildings such as envelope construction, internal construction, furniture, etc. can add autocorrelation of time-lagged errors through the effect of building thermal mass. This challenge is generally overlooked by grid operators and those interested in remuneration of grid services. One way of dealing with these issues is to use "lagged variables," which integrate building properties into the models without direct knowledge of physical parameters. When the model uses these variables, they may not need adjustments, as they are already incorporating the effects of time and building characteristics (Model R.4) [8].

Similarly, envelope construction can affect the relationship between weather variables and the building's response (e.g. more glazing leads to greater sensitivity to solar radiation and outdoor temperature). One way to indirectly include information about the envelope is to use derived parameters such as "change point temperature" to capture the building response [17], [20], [34], [53]. Change point or balance temperatures define the minimum outdoor temperature at which a building requires cooling and speak indirectly to the performance of the building envelope and operating conditions.

5.1.3. Time and calendar variables

Time and calendar variables enable models to predict energy for different temporal resolution e.g., ERCOT CBL predicts the building energy on a daily scale (Model R.14) [9] [33], while the mean week model (Model R.15) [26] [21] [54] creates a different load profile for each day of the

week. These variables enable models to account for periodicity due to operation schedule [69], e.g. the model DDT performs separate regression for occupied and unoccupied periods (Model R.7) [26], [70]. Many studies [27], [31], [40], [71] have developed separate statistical models for different periods using the time and calendar variables. Greater availability of data is likely to allow for greater understanding of these effects [12].

5.1.4. Baseline aggregation

Lastly some models use aggregated data from customers instead of using the individual readings. Aggregation is preferred when the buildings being modeled are highly variable and lower computational power is required [9]. The precision improves with the size of aggregation and if the aggregated accounts have similar weather and customer mix [10], [25], [28]. This finding is supported by another study for a different region in the USA. In this study [55], the difference between actual energy consumption and model-predicted consumption for a year was calculated, categorizing buildings based on the percentage difference.

This method, akin to the load variability assessment discussed in Table 4, employed a cutoff of less than 35% for low load variability (LLV) and greater than 35% for high load variability (HLV). While the approach yielded unbiased and precise baselines through portfolio baselines, evidence suggests a potential underestimation compared to models using individual meter readings as inputs [14], [61]. Care should be taken while using aggregated data as averaging removes nuances in the data [8]. It is recommended to review time-series plots of a significant sample of demand response events before drawing conclusions [32].

5.2 Time periods

Modeling decisions related to relevant time periods- data granularity, training period, prediction horizon- have been shown to affect model performance and are discussed presently.

5.2.1. Data granularity

Most models discussed ingest hourly data. One exception is the day-matching model (A.30), which uses 15-minute resolution to identify day pairs with the lowest SSE [8]. Of the regression models reviewed, only the ERCOT model (R.14) requires 15-minute resolution [33]. Generally, regression-based models perform effectively with hourly data. It is not likely that models will be improved by granularity in data finer than 15 minutes [45], though this resolution of data is likely to be available for many buildings in the near future, as the characteristic time scales in buildings are longer than this in most cases. In fact, finer resolution data is likely to lead to overfitting issues

and this is likely the reason there has been very little investigation of finer resolution models in the literature.

5.2.2. Training days for averaging methods

As discussed previously, models can ingest as few as three days' data up to more than a year's data. There is evidence that the number of input days significantly affects accuracy [26]. A few studies suggest that length of training period affects model performance more than even length of prediction horizon [27], [31]. This can be seen in Figure 4, where the performance of OAT models generally improves with increasing training period. One study found accuracy improves as the number of input days increases from 0 to 18 [14], but gets worse from 18 to 39, and flattens beyond 39 days. This was backed by another study, that found when a 20-day period is used, factors like load changes or temporary variations in building operation [32] [18], which are generally overlooked by models ingesting fewer than 10 days' data, can be included and improve the model.

Despite the general trend of more data improving model performance up to around 20 days, most day matching methods use 10 or fewer days. In general, 10 days captures near-term trends while mitigating potential of gaming, which is deliberate action by customers during nonevent hours designed to manipulate baselines. [24]. However, fewer-day models may perform poorly for buildings that have highly variable building profiles, shed regularly, or have swings in weather and power. Adding adjustments can account for temporary variations to some extent [10], [24], and Middle X of Y methods can be used to discard outliers [33], [52], [72], but it is recommended to filter bad data like shutdowns or large swings in consumption before using day matching methods [25], [45]. There are numerous recommendations for accomplishing this filtering process in the literature, discussed previously. Nevertheless, regression methods are recommended for customers who shed load regularly or have variable load profiles.

5.2.3. Training period for regression-based methods

Regression models capture the effect of operative variables such as outdoor temperature, and therefore days with conditions unlike the shed day can be included. This leads to more data being used for regression models than averaging methods in almost all cases, and the amount of data used has been shown to affect model performance. Regression models have reported lower bias and variability with longer training periods for both WS and NWS accounts [8], [71]. Especially with buildings that shed load regularly and have variable load profiles, improvements have been documented from including a longer and diverse dataset [8], [19], [73].

However, like with any modeling endeavor, there exists tradeoffs among data availability, computational expense, and model performance and using a great deal of data is sometimes not possible or feasible. One study [26] also suggests that using long-term data for training may not necessarily improve the accuracy of predictions but using too little data has been shown to degrade model performance [21], [69], [71].

To address these competing constraints, there is an emerging consensus around an optimum training period of around nine months for regression models. [28] suggest the 6–9 months of data is ideal for aggregated models like mean week (Model R.15). [26] similarly showed improved performance with longer training periods, with performance peaking at nine months of data. Similarly, [21] increasing performance with additional data up to 9-12 months. It should be noted that this amount of data may be insufficient for advanced regression methods and bin modeling [19], [73].

Another way to reduce computational expense is to use filtering to reduce the number of input days while ensuring the days used to train the model are relevant and provide for accurate predictions. This is done in Model R.7, which sorts the input days based on average or maximum temperature and selects the 10 hottest days [23], [34]. Another model (Model R.19) filters and weighs the days closer to the event days [31], improving the model performance, also seen in Figure 4. In one study [18], a hybrid model using only two weeks of carefully selected data resulted in a NMBE of 12% for medium-sized offices and 7% for large-sized offices.

It is not expected that constraints on availability of training data or computational power will play much of a role in determining model accuracy moving forward. Data availability is increasing rapidly with the deployment of advanced metering infrastructure, and computational power continues to increase as it has for decades. Moving forward, there is likely to emerge a constraint on the amount of data used that has to do with changing patterns of operation of buildings, i.e., the building is not used in exactly the same way it was 5 years ago. While this phenomenon was not mentioned in the literature, other studies [74] have shown a shift in how buildings are being used, especially after the Covid-19 pandemic.

5.2. Baseline estimation method

As with most modeling problems, different methods are appropriate for different baselining problems. Averaging models are simpler [3] and thus have been used more often, as can be seen in Figure 2. Different averaging methods have strengths, weaknesses and domains of applicability as described in the literature and summarized in Table 8.

Table 8: Strengths and weakness of average based methods

Approach	Strengths	Weakness
Emergency models	• Efficient for quick calculations with short datasets.	• Evidence that it can be ineffective for frequent load shedding due to the same baseline estimation for all load shed days [24]. • Can be susceptible to pre-cooling and notable high bias observed in predictions [8], [9].
High X of Y (Day matching)	• Simple model construct [24], [25]. • Only 20 days needed for best performance [14]. • Further improvement achieved by assigning weight to days closer to the event [25], [26]. • High 10 of 10 excels in summer with low intra-day variation, serving as an industry benchmark for predicting annual peak load days, particularly in summer DR events [14], [24]. The model performs better than a few regression methods for WS and LV buildings [25], [32]	• Can be ineffective for frequent load shedding, as the same baseline is estimated for all the load shed days [24]. • Has high gaming potential [24]. • Can be susceptible to pre-cooling and notable high bias observed in predictions [8], [9]. • Prior study [14] showed potential to over predict in swing seasons. • Data like shut downs or large swings in consumption need to be filtered [25], [45]. • Need to be adjusted for an accurate estimate [25]
Middle X of Y (Day matching)	• Can be effective for swing season predictions as it removes the high and low energy consumption from the input [25]	• Susceptible to pre-cooling impact [9]. • Less commonly used, can be ineffective for frequent load shedding.
Demand based matching (Day matching)	• Identifying matching day pairs in energy consumption provides a theoretically simple method for baseline estimation [8].	• Requires 15-minute data for matching [8]. • Infrequently implemented and lacks widespread use. • Ineffective for frequent load shedding due to challenges in finding multiple distinct days each year that match the load shed energy consumptions [8].
Weather matching	• Customized for local climatic conditions [8].	• Challenging to automate for multiple climates, requiring specific temperature threshold consideration for each climate.

The regression approach has a higher complexity and larger data requirement, but the same model can be applied to all event times or consecutive load shed days, unlike averaging methods [9]. Regression models allow the inclusion of multiple inputs to capture effects of weather, building and time and thereby improve the accuracy of the baseline estimate [18], [25]. Evidence suggests that gaming a regression model is difficult [24]. It should be noted, however, that there is still a preference for average-based methods simply because regression requires more data and is not as easy to quickly use to remunerate customers [8], [9], [24], [25]. Within the class of regression methods, several approaches have been articulated and tested in the literature. Their strengths and weaknesses are summarized in Table 9.

Table 9: Strengths and weakness of regression based methods

Approach	Strengths	Weakness
OAT models (Simple linear regression or SLR)	• Needs only one input. • Can use degree days to normalize weather effects [17], [37], [49], [61]. • Can use "lagged variables" to integrate thermal mass indirectly [8].	• Models disregard other weather variables [71]. • Integration of lagged variables and degree day tends to increase baseline estimate variability [17], [61].
OAT with other variables (Multiple linear regression or MLR)	• Permits inclusion of interaction terms, building variables, and time variables [9]. • Documented improvement in performance from greater model specificity. • Capture intra-day load changes due to weather more accurately than average based methods [32].	• Too many parameters can hinder computation and reduce accuracy [71]. • Collinearity among variables poses challenges [34]. • May require 15-minute data [9], [33]. • With fewer input days, it is less likely to capture weather conditions [14]. • Adjustment calculations can be cumbersome [27].
Change point models (Piece-wise Regression)	• Adaptable to seasonal shifts • Indirectly includes building envelope performance and operational conditions. • Can be extended to multiple variables [34]. • Serve as the industry benchmark [34].	• Requires an efficient algorithm for determining balance points without relying on complex multi-linear regression with extensive data [17], [71]. • The change point behavior introduces nonlinearity since CP is a parameter to be determined [71]. • Difficult to automate for multiple buildings.
Time series	• Cleans data to eliminate noise [50]. • Indirectly incorporates thermal mass effects.	• Challenging to handle missing data [50]. • Implementation can be complex, lacks guidance.
Hybrid models	• Can adapt to use short-term data [18], [19], [20]. • Recommended when error terms exhibit first-order autocorrelation.	• If using electricity data, inconsistent formats across utilities may pose challenges for automation [14].

5.3. Adjustments

Baseline models can be adjusted by a scalar or additive adjustment or left unadjusted. Unadjusted models may be beneficial for grid operators when the chances of gaming are high [8]. However, adjustments can reduce modeling error by more than half across all prediction time windows [9], [24], [31] as shown in Figure 6, especially for WS accounts [8]. Particularly for averaging methods, adjustments are recommended. A maximum adjustment of ±20% is recommended for day matching methods while ±40% has been recommended for weather matching methods [10], [18], [24], [31].

Interestingly, a study by LBNL [27], managed to use adjustment to bypass time series modeling. Two error terms were calculated to remove the autocorrelation effect from days prior to the event and days following the event. The study added the error terms as adjustments, which made the model (Model R.17) robust to outliers and stable in its prediction performance [54], [71]. However, there is evidence that a full season lag model gives comparable predictive accuracy to an adjusted model [8], although some models with long training periods use adjustments nonetheless to reduce autocorrelation in the data [27], [71].

Figure 6 illustrates the effect of adjustments on model performance as discussed in the literature. A few general trends are evident. First, unadjusted models show a tendency to underpredict baseline for all cases (NWS/WS/Variable load/non-variable load). Adjustments help reduce this bias. Secondly, buildings with highly variable loads present challenges for baseline estimation, regardless of adjustment approach and thus the error reported for these buildings is greater in almost all cases. Weather sensitive adjustment appears to reduce model error for WS buildings and NWS as well. It should be noted that even though the building is classified as NWS, each building has some level of weather sensitivity which may not be captured by the method used for distinguishing. No broad trend is evident to recommend either scalar or additive adjustments over their counterpart.

If models are to be adjusted, a decision must be made as to how they are adjusted. Additive adjustments are recommended for buildings with constant loads that are not weather sensitive [3],[6]. Multiplicative adjustments are recommended for both WS and NWS buildings. The adjustments rely on the scale of the building's load profile and are believed to capture weather sensitivity and fluctuations in NWS accounts better than additive adjustment [8]. WS buildings lend themselves to weather sensitive adjustments (WSA), but WSA are not suitable for buildings dominated by internal loads [8]. Time-based adjustment was recommended in such cases [3],[6].

The adjustment process is a place where the issue of gaming often arises. Adjusting based on

pre-event hours has been recommended over post-event hours, or the combination of pre- and post-event hours [10], [31]. However, the use of pre-event hours can be problematic. It increases the chances of gaming and accidental inflation of baseline, since if buildings pre-cool it would increase the baseline [27]. Conversely, sometimes office buildings cancel operations when faced with a curtailment notice or industrial buildings, which take time to shut down, might start the process of load shedding early. When this happens, the load in the 2 hours before the shed is less than typical and modeled baseline is low. In such cases, either scalar adjustment based on third and fourth hour before the load shed or unadjusted full season daily degree day models is recommended [8], [25], [54]. Another study [24] recommends two hours prior to event notification instead of two hours prior to the event start, i.e., adjustment window should not overlap with the "ramp period" (period right before the shed) [25]. Others recommend adjustments based on morning operation, although this has issues as the building may be unoccupied [27], [31] and may have night setback strategies in place [40].

Besides gaming, adjusting in a way that does not punish energy efficiency, or variable load patterns, is a subject of discussion. Adjustments can be either symmetric (adjust either up or down) or asymmetric (only adjusted up). Disallowing symmetric adjustments punishes customers with variable loads that may happen to be low during an event (common for industrial customers) [24]. Asymmetric adjustments (only adjusted up) allow room for energy-efficient operation of buildings. It also has challenges like enabling customers to receive compensation for a planned shutdown that may coincide with an event day [24]. If gaming was not an issue for grid operators, nor the issue of eroding the baseline due to responsible operation for building operators, asymmetric adjustments would be suitable for both winter and summer programs [8], [9], [25]. It is not clear that a comprehensive answer that accomplishes the three competing objectives— disincentivizing gaming, incentivizing long-term energy efficiency, and properly remunerating load shedding— has been developed. This is an area of research that may be fruitful.

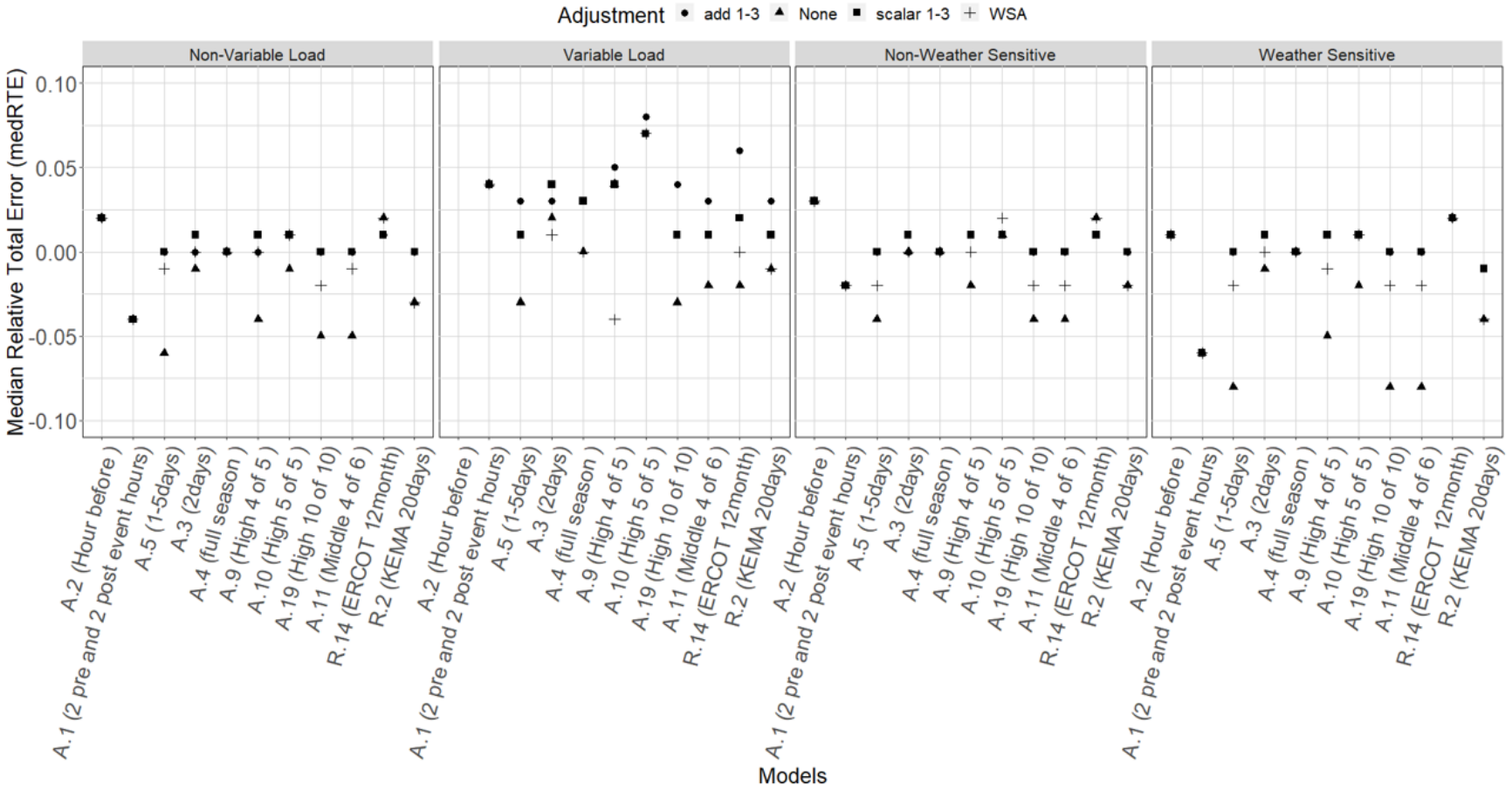

Figure 5: Effect of adjustments and impact of weather sensitivity and load variability on baseline models. The data has been extracted from *[9]* and the graph has been clipped between -0.1 to 0.1. This has resulted in the removal of few data points from variable loads for A.2 and A.10. The emergency model (A.2) had medRTE value of 0.3 for all adjustment and the High 4 of 5 (A.10) had a value of 0.15 for the unadjusted variation

5.4. Limitations

This study and the models discussed within in have several inherent limitations. First, the study is confined to models used by grid operators in the United States. The study discusses models used in different regions of the United States, which spans several climates, but primarily lies in a temperate climate region. In particular, locations with very humid climates may be expected to see model performance different from that described herein.

The models described and the published results of their testing are also heavily weighted toward summer hours when cooling demand is greatest. In regions with substantial heating loads that are served via electrically driven heat pumps, a direction in which many countries including the United States are moving, it is likely that additional model refinement and/or development may be necessary. In regions of the world where space cooling is less prevalent, load profiles in the summertime will be affected less by variables such as weather. Space cooling is in 80% of commercial buildings [75] in the United States and thus it is expected that summer peak demand is strongly influenced by cooling, but this may be less so in other locations such as northern Europe.

Conversely, in areas of the world such as South Asia, where cooling demand may account for over 70% of all grid demand [76] at peak, model performance may be expected to vary. As cooling of buildings is expected to double or triple the required grid capacity in these regions of the developing world with hot and humid climates such as South Asia [66], while at the same time energy efficiency measures and the adoption of electric cars reduces the relative portion of electricity used for cooling in developed countries, model performance will need to be assessed and updated periodically in the near future. It is likely that regressions using only weather variables will be less useful in the United States in the future for these reasons, and the need for tools enabling demand response programs in the developing world is likely to increase vastly.

Finally, another issue is that many buildings are served by a central plant, such as occurs on many campuses and some entire cities, primarily in Europe. Any reduction in cooling load is thus not fully realized as a reduction in electric power at the individual building, but instead can occur mostly at the plant. This reduction at the plant may be lagged in time from any action taken at the building, necessitating adjustment to models.

6. Conclusions

While several previous works have discussed implications of using various baseline models in different applications, there has been no systematic and comprehensive evaluation of the rationale for choosing a particular baseline, which is provided herein. This work provides a comprehensive list of 50 baseline models used in research and industry. The salient features of the baseline models and the strengths and limitations of each method are discussed below.

A few broad lessons can be drawn from a survey of extant literature, which may be instructive for improvement and application of models:

- A consensus as to the proper set of input variables has not emerged, even for a particular building type. There are several reasons for this, including lack of availability of some data and concerns over statistical issues around using collinear inputs. The second issue is likely something that can be addressed using more advanced statistical methods but may then move the models into a degree of complexity that is beyond what is desired/acceptable for most users. Creating a robust model with all relevant inputs that nonetheless retains necessary simplicity is a challenge waiting for interested modelers.
- This research pointed to a lack of "building" information in most models in practice, meaning metadata on construction of the building in question or information on how it is operated. In many cases this is because of lack of the metadata, and in others it is intentional to ensure models are agnostic to building type being modeled, or even whether the load being modeled is attributable to a building. In cases where greater accuracy is desired, a few purely data-driven workarounds have been developed including the use of a time-of-week variable as a surrogate for occupancy and building operation patterns, and the use of change-points, which are data-driven indicators of building performance that do not require explicit information about the building. These both seem to improve model predictions. Similar creative workarounds, as well as direct input of building information from central repositories, may further improve baselines as well.
- As in any modeling endeavor, in baseline modeling there exists a tradeoff between accuracy and computational complexity. Extensive filtering processes are reviewed herein and can reduce the amount of data needed to produce accurate predictions. Where little data is available, best practices can maximize model performance, nonetheless. These include aggregation of several similar load profiles and adding some physics to the model where possible, even if it is a simple regressed relationship between outdoor temperature and load.

This also alleviated the problem of having to determine appropriate adjustments and weighting factors as exists with simple averaging methods. Including other variables as discussed above will also likely help in these situations. This should be less of an issue as more advanced metering infrastructure is installed and data collected.

- The issue of gaming remains a concern for some parties. A few mitigations were suggested in the literature, including making adjustments on days that are not near the shed day and using larger datasets. This also seems to be an area where improvement can be made, for example, in development of an algorithm that can detect gaming and adjust accordingly.
- Perhaps the opposite of the issue of gaming is the concern over not punishing users who are developing energy efficiency programs that may already be shaping loads in a way that would reduce their compensation for additional demand shaping during a demand response event. Some work has been done looking at this issue, but it can likely use additional work.
- It should be noted that all the methods studied may break down for customers with highly variable loads, and especially for customers whose loads are both highly variable and driven by internal loads such as equipment. Methods for predicting highly variable loads are still lacking and are likely much different than those for other buildings. One potential method for compensation of these customers might be an absolute target rather than a modeled shed, though more work is necessary in this area.
- The integration of more advanced statistical techniques incorporating factors such as solar radiation, occupancy, and time variables holds promise for enhancing predictive accuracy. Given the observed relationship between the time and season of load shedding, future investigations should explore the inclusion of these temporal dimensions in modeling efforts. Similarly, considering the size and type of buildings could provide valuable insights for refining baseline models.

As participation in demand response program grows and is starting to include residential, agricultural, and other customers not traditionally participating, further development of appropriate methods will be required. This review identifies the state of the art for C&I applications and a starting point for those investigations.

Acknowledgements

The work was funded by the U.S. Department of Energy under Grant # EE0009776: Connected Communities. This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE- AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

References

[1] "Order No. 719 | Federal Energy Regulatory Commission." Accessed: Apr. 20, 2023. [Online]. Available: https://www.ferc.gov/media/order-no-719

[2] M. A. Piette, D. Watson, N. Motegi, and S. Kiliccote, "Automated Critical PeakPricing Field Tests: 2006 Pilot ProgramDescription and Results," Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), LBNL-62218, Jun. 2007. doi: 10.2172/919387.

[3] M. Young, B. D. Less, S. M. Dutton, I. S. Walker, M. H. Sherman, and J. D. Clark, "Assessment of peak power demand reduction available via modulation of building ventilation systems.," *Energy and Buildings*, vol. 214, p. 109867, May 2020, doi: 10.1016/j.enbuild.2020.109867.

[4] B. D. Less, S. M. Dutton, I. S. Walker, M. H. Sherman, and J. D. Clark, "Energy savings with outdoor temperature-based smart ventilation control strategies in advanced California homes," *Energy and Buildings*, vol. 194, pp. 317–327, Jul. 2019, doi: 10.1016/j.enbuild.2019.04.028.

[5] J. D. Clark, B. D. Less, S. M. Dutton, I. S. Walker, and M. H. Sherman, "Efficacy of occupancy-based smart ventilation control strategies in energy-efficient homes in the United States," *Building and Environment*, vol. 156, pp. 253–267, Jun. 2019, doi: 10.1016/j.buildenv.2019.03.002.

[6] A. Hirsch, J. Clark, M. Deru, K. Trenbath, I. Doebber, and D. Studer, "Pilot Testing of Commercial Refrigeration-Based Demand Response," National Renewable Energy Lab. (NREL), Golden, CO (United States), NREL/TP-5500-65009, Oct. 2015. doi: 10.2172/1226469.

[7] M. Neukomm, V. Nubbe, and R. Fares, "Grid-Interactive Efficient Buildings Technical Report Series: Overview of Research Challenges and Gaps," NREL/TP-5500-75470, DOE/GO-102019-5227, 1577966, Dec. 2019. doi: 10.2172/1577966.

[8] M. L. Goldberg and G. K. Agnew, "Protocols for Demand Response Calculation - Findings and Recommendations," KEMA-XENERGY, Consultant Report 400-02-017F, Feb. 2003.

[9] C. Lake, "PJM Empirical Analysis of Demand Response Baseline Methods," 2011.

[10] J. Bode and A. Ciccone, "California ISO Baseline Accuracy Assessment (2017)," Nexant, Nov. 2017.

[11] Nexant, "California ISO Baseline Work Group Proposal (2017)," Nexant, Jun. 2017.

[12] K. Oh, E.-J. Kim, and C.-Y. Park, "A Physical Model-Based Data-Driven Approach to Overcome Data Scarcity and Predict Building Energy Consumption," *Sustainability*, vol. 14, no. 15, Art. no. 15, Jan. 2022, doi: 10.3390/su14159464.

[13] I. Qaisar and Q. Zhao, "Energy baseline prediction for buildings: A review," *Results in Control and Optimization*, vol. 7, p. 100129, Jun. 2022, doi: 10.1016/j.rico.2022.100129.

[14] J. Granderson, S. Fernandes, E. Crowe, M. Sharma, D. Jump, and D. Johnson, "Accuracy of hourly energy predictions for demand flexibility applications," *Energy and Buildings*, vol. 295, p. 113297, Sep. 2023, doi: 10.1016/j.enbuild.2023.113297.

[15] N. Javaid, A. Naz, R. Khalid, A. Almogren, M. Shafiq, and A. Khalid, "ELS-Net: A New Approach to Forecast Decomposed Intrinsic Mode Functions of Electricity Load," *IEEE Access*, vol. 8, pp. 198935–198949, 2020, doi: 10.1109/ACCESS.2020.3034113.

[16] N. Somu, G. Raman M R, and K. Ramamritham, "A deep learning framework for building energy consumption forecast," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110591, Mar. 2021, doi: 10.1016/j.rser.2020.110591.

[17] J. K. Kissock, J. S. Haberl, and D. E. Claridge, "Inverse Modeling Toolkit: Numerical Algorithms," *ASHRAE Transactions*, vol. 109, pp. 425–434, 2003.

[18] B. Abushakra and M. T. Paulus, "An hourly hybrid multi-variate change-point inverse model using short-term monitored data for annual prediction of building energy performance, part I: Background (1404-RP)," *Science and Technology for the Built Environment*, vol. 22, no. 7, pp. 976–983, Oct. 2016, doi: 10.1080/23744731.2016.1215222.

[19] B. Abushakra and M. T. Paulus, "An hourly hybrid multi-variate change-point inverse model using short-term monitored data for annual prediction of building energy performance, part III: Results and analysis (1404-RP)," *Science and Technology for the Built Environment*, vol. 22, no. 7, pp. 996–1009, Oct. 2016, doi: 10.1080/23744731.2016.1215659.

[20] B. Abushakra and M. T. Paulus, "An hourly hybrid multi-variate change-point inverse model using short-term monitored data for annual prediction of building energy performance, part II: Methodology (1404-RP)," *Science and Technology for the Built Environment*, vol. 22, no. 7, pp. 984–995, Oct. 2016, doi: 10.1080/23744731.2016.1215199.

[21] J. Granderson, S. Touzani, C. Custodio, M. D. Sohn, D. Jump, and S. Fernandes, “Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings,” *Applied Energy*, vol. 173, pp. 296–308, Jul. 2016, doi: 10.1016/j.apenergy.2016.04.049.
[22] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, “Estimating Demand Response Load Impacts: Evaluation of BaselineLoad Models for Non-Residential Buildings in California,” LBNL--63728, 928452, Jan. 2008. doi: 10.2172/928452.
[23] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, “Statistical analysis of baseline load models for non-residential buildings,” *Energy and Buildings*, vol. 41, no. 4, pp. 374–381, Apr. 2009, doi: 10.1016/j.enbuild.2008.11.002.
[24] EnerNoc, “The Demand Response Baseline (2008),” EnerNOC, Inc., Oct. 2008.
[25] Enernoc, “The Demand Response Baseline (2011),” EnerNOC, Inc., North America, White Paper, 2011.
[26] J. Granderson and P. Price, “Evaluation of the Predictive Accuracy of Five Whole Building Baseline Models,” LBNL--5886E, 1172955, Aug. 2012. doi: 10.2172/1172955.
[27] P. N. Price, N. Addy, and S. Kiliccote, “Predictability and Persistance of Demand Response Load Shed in Buildings,” p. 22, 2015.
[28] J. Granderson, P. N. Price, D. Jump, N. Addy, and M. D. Sohn, “Automated measurement and verification: Performance of public domain whole-building electric baseline models,” *Applied Energy*, vol. 144, pp. 106–113, Apr. 2015, doi: 10.1016/j.apenergy.2015.01.026.
[29] V. Singh, T. A. Reddy, and B. Abushakra, “Predicting Annual Energy Use in Buildings Using Short-Term Monitoring and Utility Bills: The Hybrid Inverse Model Using Daily Data (HIM-D),” vol. 119, 2013.
[30] E. Crowe, J. Granderson, and S. Fernandes, “From Theory to Practice: Lessons Learned from an Advanced M&V Commercial Pilot,” Apr. 2023, Accessed: Apr. 17, 2023. [Online]. Available: https://escholarship.org/uc/item/2hd384rc
[31] J. Granderson *et al.*, “Assessment of Model-Based peak electric consumption prediction for commercial buildings,” *Energy and Buildings*, vol. 245, p. 111031, Aug. 2021, doi: 10.1016/j.enbuild.2021.111031.
[32] Liu, Jingjing; Yu, Lili; Yin, Rongxin; Piette, Mary Ann; Pritoni, Marco; Casillas, Armando; Neukomm, Monica; Roth, Amir, “Benchmarking Demand Flexibility in Commercial Buildings and Flattening the Duck – Addressing Baseline and Commissioning Challenges,” 2022, doi: 10.20357/B7M89Q.
[33] “ERCOT: Demand Response Baseline Methodologies (Version 19.0).” Accessed: Feb. 28, 2023. [Online]. Available: https://www.ercot.com/services/programs/load
[34] D. R. Landsberg, J. A. Shonder, K. A. Barker, C. R. L. Hall, and D. T. Reindl, “ASHRAE Guideline 14-2014,” p. 150.
[35] “PJM Manual 19: Load Forecasting and Analysis.” PJM: Resource Adequacy Planning, Jun. 01, 2016.
[36] “ISO New England: Measurement and Verification of Demand Reduction Value from Demand Resources,” ISO New England Inc., Jun. 2014.
[37] J. Haberl, C. Culp, and D. Claridge, “ASHRAE’s Guideline 14-2002 for Measurement of Energy and Demand Savings: How to Determine what was really saved by the retrofit,” p. 13, 2005.
[38] M. Agenis-Nevers, Y. Wang, M. Dugachard, R. Salvazet, G. Becker, and D. Chenu, “Measurement and Verification for multiple buildings: An innovative baseline model selection framework applied to real energy performance contracts,” *Energy and Buildings*, vol. 249, p. 111183, Oct. 2021, doi: 10.1016/j.enbuild.2021.111183.
[39] I. Qaisar and Q. Zhao, “Energy baseline prediction for buildings: A review,” *Results in Control and Optimization*, vol. 7, p. 100129, Jun. 2022, doi: 10.1016/j.rico.2022.100129.
[40] P. Price, “Methods for Analyzing Electric Load Shape and its Variability,” LBNL-3713E, 985909, May 2010. doi: 10.2172/985909.
[41] V. Singh, T. A. Reddy, and B. Abushakra, “Predicting annual energy use in buildings using short-term monitoring: 2014 ASHRAE Winter Conference,” *ASHRAE Transactions - ASHRAE Winter Conference*, pp. 397–405, 2014.
[42] “CalTRACK Methods — CalTRACK Technical Documentation 2.0 documentation.” Accessed: Apr. 17, 2023. [Online]. Available: http://docs.caltrack.org/en/latest/methods.html#section-3-b-modeling-hourly-methods
[43] A. Miller and K. Carbonnier, “New Metrics for Evaluating Building-Grid Integration,” New Buildings Institute, 2020. [Online]. Available: https://newbuildings.org/wp-content/uploads/2020/11/NewMetricsForEvaluatingBuildingGridIntegration.pdf

[44] “Order No. 719 | Federal Energy Regulatory Commission.” Accessed: Apr. 11, 2023. [Online]. Available: https://www.ferc.gov/media/order-no-719
[45] N. Addy and J. L. Mathieu, “Understanding the Effect of Baseline Modeling Implementation Choices on Analysis of Demand Response Performance”.
[46] “Weather Sensitive Adjustment Using the WSA Factor Method.”
[47] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, “Estimating Demand Response Load Impacts: Evaluation of Baseline Load Models for Non-Residential Buildings in California,” LBNL--63728, 928452, Jan. 2008. doi: 10.2172/928452.
[48] PJM, “PJM Manual 11 : Energy & Ancillary Services Market Operations.” PJM, Feb. 09, 2023. [Online]. Available: https://www.pjm.com/~/media/documents/manuals/m11.ashx
[49] D. K. Ruch, J. K. Kissock, and T. A. Reddy, “Prediction Uncertainty of Linear Building Energy Use Models With Autocorrelated Residuals,” *Journal of Solar Energy Engineering*, vol. 121, no. 1, pp. 63–68, Feb. 1999, doi: 10.1115/1.2888144.
[50] Y. Yang and Y. Yang, “Hybrid Method for Short-Term Time Series Forecasting Based on EEMD,” *IEEE Access*, vol. 8, pp. 61915–61928, 2020, doi: 10.1109/ACCESS.2020.2983588.
[51] C. E. Commission, “Home Page,” California Energy Commission. Accessed: Dec. 26, 2023. [Online]. Available: https://www.energy.ca,gov
[52] “PJM Operating Agreement,” AMENDED AND RESTATED OPERATING AGREEMENT OF PJM INTERCONNECTION, L.L.C. [Online]. Available: https://www.pjm.com/library/governing-documents
[53] S. Katipamula, T. A. Reddy, and D. E. Claridge, “Multivariate Regression Modeling,” *Journal of Solar Energy Engineering*, vol. 120, no. 3, pp. 177–184, Aug. 1998, doi: 10.1115/1.2888067.
[54] J. L. Mathieu, P. N. Price, S. Kiliccote, and M. A. Piette, “Quantifying Changes in Building Electricity Use, With Application to Demand Response,” *IEEE Trans. Smart Grid*, vol. 2, no. 3, pp. 507–518, Sep. 2011, doi: 10.1109/TSG.2011.2145010.
[55] E. Crowe, J. Granderson, and S. Fernandes, “From Theory to Practice: Lessons Learned from an Advanced M&V Commercial Pilot,” 2019, doi: 10.20357/B73591.
[56] H. Johra, M. Schaffer, G. Chaudhary, H. Kazmi, J. Le Dréau, and S. Petersen, *What Metrics Does the Building Energy Performance Community Use to Compare Dynamic Models?* 2023.
[57] J. M. MacDonald and D. M. Wasserman, “Investigation of metered data analysis methods for commercial and related buildings,” ORNL/CON-279, 6261458, May 1989. doi: 10.2172/6261458.
[58] D. K. Ruch and D. E. Claridge, “A development and comparison of NAC estimates for linear and change-point energy models for commercial buildings,” *Energy and Buildings*, vol. 20, no. 1, pp. 87–95, Jan. 1993, doi: 10.1016/0378-7788(93)90041-R.
[59] S. Thamilseran and J. S. Haberl, “A Bin Method for Calculating Energy Conservation Retrofit Savings in Commercial Buildings,” 1994, Accessed: Jul. 25, 2023. [Online]. Available: https://oaktrust.library.tamu.edu/handle/1969.1/6640
[60] J. K. Kissock, T. A. Reddy, and D. E. Claridge, “Ambient-Temperature Regression Analysis for Estimating Retrofit Savings in Commercial Buildings,” *Journal of Solar Energy Engineering*, vol. 120, no. 3, pp. 168–176, Aug. 1998, doi: 10.1115/1.2888066.
[61] J. H. Eto, “On using degree-days to account for the effects of weather on annual energy use in office buildings,” *Energy and Buildings*, vol. 12, no. 2, pp. 113–127, Sep. 1988, doi: 10.1016/0378-7788(88)90073-4.
[62] D. Ruch and D. E. Claridge, “A Four-Parameter Change-Point Model for Predicting Energy Consumption in Commercial Buildings,” *Journal of Solar Energy Engineering*, vol. 114, no. 2, pp. 77–83, May 1992, doi: 10.1115/1.2929993.
[63] J. C. Lam, K. K. W. Wan, S. L. Wong, and T. N. T. Lam, “Principal component analysis and long-term building energy simulation correlation,” *Energy Conversion and Management*, vol. 51, no. 1, pp. 135–139, Jan. 2010, doi: 10.1016/j.enconman.2009.09.004.
[64] J. Pfafferott, S. Herkel, and J. Wapler, “Thermal building behaviour in summer: long-term data evaluation using simplified models,” *Energy and Buildings*, vol. 37, no. 8, pp. 844–852, Aug. 2005, doi: 10.1016/j.enbuild.2004.11.007.
[65] J.-H. Ko, D.-S. Kong, and J.-H. Huh, “Baseline building energy modeling of cluster inverse model by using daily energy consumption in office buildings,” *Energy and Buildings*, vol. 140, pp. 317–323, Apr. 2017, doi: 10.1016/j.enbuild.2017.01.086.

[66] S. Valovcin, A. S. Hering, B. Polly, and M. Heaney, “A statistical approach for post-processing residential building energy simulation output,” *Energy and Buildings*, vol. 85, pp. 165–179, Dec. 2014, doi: 10.1016/j.enbuild.2014.07.060.
[67] M. Jin, N. Bekiaris-Liberis, K. Weekly, C. Spanos, and A. Bayen, “Sensing by Proxy: Occupancy Detection Based on Indoor CO2 Concentration,” 2015.
[68] “Commercial Reference Buildings,” Energy.gov. Accessed: Nov. 24, 2023. [Online]. Available: https://www.energy.gov/eere/buildings/commercial-reference-buildings
[69] H. Fu, J.-C. Baltazar, and D. E. Claridge, “Review of developments in whole-building statistical energy consumption models for commercial buildings,” *Renewable and Sustainable Energy Reviews*, vol. 147, p. 111248, Sep. 2021, doi: 10.1016/j.rser.2021.111248.
[70] “Time Dependent Valuation of Energy for Developing Building Efficiency Standards.” Energy and Environmental Economics, Inc., Jul. 2014.
[71] T. A. Reddy, J. K. Kissock, and D. K. Ruch, “Uncertainty in Baseline Regression Modeling and in Determination of Retrofit Savings,” *Journal of Solar Energy Engineering*, vol. 120, no. 3, pp. 185–192, Aug. 1998, doi: 10.1115/1.2888068.
[72] “NYISO: Day-Ahead Demand Response Program Manual.” Nov. 2022.
[73] V. Singh, T. A. Reddy, and B. Abushakra, “Predicting annual energy use in buildings using short-term monitoring and utility bills: The hybrid inverse model using daily data (HIM-D),” *ASHRAE Transactions*, vol. 119, pp. 169–180, Jan. 2013, doi: 10.1201/b15398-7.
[74] K. Chaloeytoy, V. Inkarojrit, and A. Thanachareonkit, “Electricity Consumption in Higher Education Buildings in Thailand during the COVID-19 Pandemic,” *Buildings*, vol. 12, no. 10, p. 1532, Sep. 2022, doi: 10.3390/buildings12101532.
[75] “Energy Information Administration (EIA)- Commercial Buildings Energy Consumption Survey (CBECS) Data.” Accessed: Mar. 25, 2023. [Online]. Available: https://www.eia.gov/consumption/commercial/data/2018/index.php?view=microdata
[76] “Keeping cool in a hotter world is using more energy, making efficiency more important than ever – Analysis,” IEA. Accessed: Dec. 14, 2023. [Online]. Available: https://www.iea.org/commentaries/keeping-cool-in-a-hotter-world-is-using-more-energy-making-efficiency-more-important-than-ever