IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 January 2023, accepted 17 January 2023, date of publication 23 January 2023, date of current version 27 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3239328

RESEARCH ARTICLE

A Framework to Predict Variability Characteristics in Building Load Profiles

SAM MOAYEDI[1], ANDREW PARKER[2], KEVIN JAMES[1], XIAOYUE CHENG[3], MICHAEL HEMPEL[4], (Member, IEEE), HAMID SHARIF[4], (Fellow, IEEE), AND MAHMOUD A. ALAHMAD[1], (Senior Member, IEEE)

[1]Durham School of Architectural Engineering and Construction, University of Nebraska–Lincoln, Scott Campus, Omaha, NE 68182, USA
[2]National Renewable Energy Laboratory, Golden, CO 80401, USA
[3]Department of Mathematics, University of Nebraska Omaha, Omaha, NE 68182, USA
[4]Electrical and Computer Engineering, University of Nebraska–Lincoln, Scott Campus, Omaha, NE 68182, USA

Corresponding author: Mahmoud A. Alahmad (malahmad2@unl.edu)

This work was supported in part by the National Renewable Energy Laboratory, University of Nebraska–Lincoln Layman Award, and in part by the University of Nebraska at Omaha Office of Research and Creative Activity.

ABSTRACT The expansion of Advanced Metering Infrastructure (AMI) has provided building operators and researchers detailed information on building energy consumption. The majority of AMI systems, however, record data at relatively low resolutions of 15, 30, or 60 minutes, due to cost, storage and bandwidth limitations. Emerging applications in power flow analysis, Quasi-Static Time-Series Simulation (QSTS), smart grid integration and load matching, however, require data at higher resolutions. Short-term energy demand can deviate significantly from long-term averages, with an unknown magnitude and frequency when only low-resolution load profile data is available. This paper presents a novel data-driven approach to predict characteristics of the missing high-resolution information in a low-resolution signal, applicable to both measured and modeled building load profile data, utilizing machine learning regression algorithms. In the proposed framework, the relationship between characteristics of high-resolution and low-resolution signals is learned from the decomposition and characterization of a subset of high-resolution building data. This paper validates the underlying hypotheses and methodology of this approach through a single-building case study, training a variety of machine learning models on one year of data, and using the resulting models to predict high-resolution characteristics in a different year. An Ensemble Tree regression model demonstrates a high predictive accuracy (R^2 of 0.79-0.92) for several statistical metrics of the high-resolution load profile. These results support the broader potential for leveraging low-resolution information to accurately constrain predictions of missing high-resolution information in building load profiles, which may greatly increase the utility of both measured and modeled data in many practical and research applications. Generalizing such models will require analysis of high-resolution data from a diverse set of building types.

INDEX TERMS Building load profile modeling, decomposition, DWT, regression, variability prediction.

I. INTRODUCTION AND RELATED WORK

With an increasing number of connected distributed energy resources (DERs) in the grid, and electrical loads in the built environment, the real-time impact on voltage, frequency, and power generation/consumption should be evaluated [1]. In such an effort, measured or modeled time-series data is typically being used for power system analysis. The accuracy of many of these studies depends on the available resolution (the time interval between two consecutive measurements) of the time-series data. For example, researchers in [2] studied the modeling of load profiles at the distribution feeder level for a realistic Quasi-Static Time-Series (QSTS) simulation. They have demonstrated the need for load profiles to

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

provide at least a one-second resolution to accurately model the distribution system and capture key component operations such as the number of regulator tap changes. The author in [3] discussed the importance of the availability of 1-minute resolution data on the demand side management analysis, while the impact of time resolution on load matching in the presence of DERs such as photovoltaics (PV) was discussed in [4] and [5]. They also concluded that high-resolution data (with resolution in the 1 to 5 minute range) carries more information about the generation/consumption than hourly data and can significantly reduce the model's error. As reported in [6], accurate power flow and voltage regulation analysis relies on high-resolution data, since the controller of the current-voltage regulation equipment like capacitors requires a time delay of around 30 seconds. Therefore, for power flow analysis purposes, the resolution of load data should match the controller's defined delay.

Building load profiles are time-series load profile data that vary across different observation time horizons from sub-minute to hourly data [7], [8] measured using AMI devices. AMIs provide critical spatio-temporal energy usage information, applicable to energy efficiency measures [9] or building energy analysis. The number of deployed AMIs across the U.S. continues to grow. As of 2016, approximately 76 million customers (out of roughly 152 million electricity customers) had these smart meters [10]. By 2020, the number of installed meters had grown to about 102.9 million [11]. These smart meters are capable of providing near real-time power and energy consumption information with high temporal granularity. This information can benefit building services and assets in the smart grid domain by offering detailed information, facilitating the realization of grid-interactive energy efficient buildings (GEB) [12].

However, current practices to measure this information vary widely as AMI measurement intervals vary across the built environment. Every utility across the United States collects data at different time resolution scales ranging from one second to one hour with typical values being one, five, 15, 30, and 60 minutes. In 2016, with approximately 16.3 million AMI meters analyzed, the Smart Grid Investment Grant Program (SGIG) found that 52% of residential meters collect data at the 60-minute interval, 6% at the 30-minute interval, and 42% at the 15-minute interval. For commercial meters, the data shows 22% of meters collect data at the 60-minute interval, 5% at the 30-minute interval, 72% at the 15-minute interval, and 1% at the one-minute interval. For industrial meters, 7% collect data at the 60-minute interval, 3% at the 30-minute interval and 90% at the 15-minute interval [13].

The time resolution of AMI data is often limited due to hardware or financial constraints, preventing AMIs from collecting, processing, transferring, or storing power consumption data at higher resolution. Also, building owners do not tend to share their power consumption in open access databases due to privacy concerns. To overcome the lack of load profiles at high resolutions, researchers often rely on data from nominally similar buildings, allocated and scaled feeder data [14], or modeled load profiles [15], [16], [17], [18]. Each of these approaches suffers from a potentially significant loss of accuracy at high resolutions for a given building load profile.

Some of the high-resolution features of a low-resolution load profile may be recovered through signal-processing techniques such as compressed sensing [19]. In [20] a compressed version of a load profile is used to forecast 15-minute demand with high accuracy, indicating that some high-resolution load features may be predictable without requiring measurement at that resolution.

There are several potential approaches to characterize or model the high-frequency components of load profiles. Authors in [21] analyzed load profiles in several higher education buildings by applying wavelet decomposition and constructing typical load patterns through two-stage clustering. The uncertainty in these load patterns is quantified by entropy, with a higher uncertainty observed in the high-frequency patterns, but a high probability for patterns to remain consistent over time. In [22], researchers developed a diversity and variability library from high-resolution transformer data, by applying Discrete Wavelet Transformation (DWT). Such variability libraries can be scaled to the feeder level to produce a more realistic load representation, but the library approach is difficult to generalize for an arbitrary building. Similar to the library approach is the generation of synthetic, representative high-resolution load profiles by training models on real data. Researchers in [23] applied generative adversarial networks and kernel density estimators to generate realistic synthetic profiles based on real 1-minute data.

Future short-term fluctuations may be highly correlated to past short-term fluctuations in a given building, such that short-term forecasting is possible with a reasonable degree of accuracy, as long as historical data is measured at sufficient resolution. In [24] researchers apply a long short-term memory recurrent neural network-based framework to predict short-term demand for individual residential users, with promising accuracy compared to previous approaches.

One challenge for the prediction of high-resolution load profile behavior, however, is the scalability of observed profiles to an arbitrary system for which high-resolution data may not be available. In [25], researchers characterize the short-term fluctuations in several datasets using the standard deviation. The relationship between the short-term demand uncertainty and the base demand level is modeled as a polynomial function. Such a model shows a significant increase in accuracy over the assumption that uncertainty is a fixed percentage of the base demand.

The present research seeks to generalize this scalability, by examining the dependence of high-resolution characteristics on the low-resolution profile as a function of time, as well as other building or external characteristics. The underlying problem of resolution is illustrated in Figure 1.

This Figure represents a load profile of an actual building for three consecutive days in 2019, measured at two different time resolutions of one minute and one hour. Energy (kWh) measurements are converted to average kW to allow direct comparison across time resolutions. As can be seen, the load at the one-hour resolution (red), interpolated to one-minute, is not an accurate representation of the consumption at the one-minute resolution (blue) because high-frequency fluctuations are missing in the one-hour resolution. At the hourly scale, many load profiles follow somewhat deterministic patterns, exhibiting a predictable rise and fall over the course of a day. Below this scale, short-term energy demand can exhibit highly random characteristics, becoming more similar to noise. The term ''load profile variability'' is used to describe this highly random, high-frequency component of the load profile. This variability exists at multiple timescales, up to and including the actual instantaneous power. The variability may be affected by equipment, occupancy behavior, building size, or any number of factors. Henceforth, the low-resolution load profile data will be referred to as the ''base load,'' and the high-frequency phenomena missing from this base load as the ''variability.''

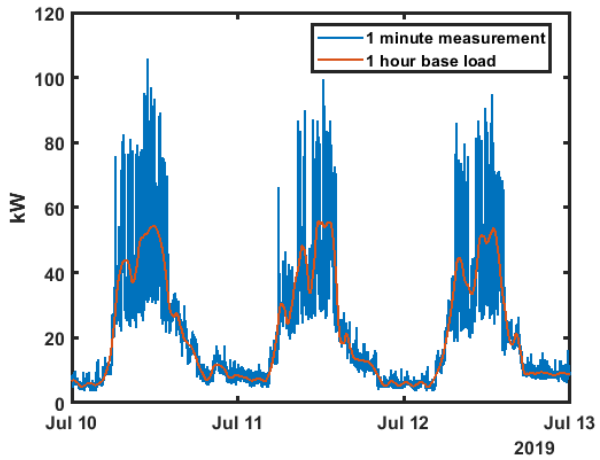


FIGURE 1. One-minute (blue) and one-hour (red) load profiles measured in an actual building for a three-day period.

To address this lack of variability in building load profiles, this research builds upon ongoing work, by proposing a novel data-driven approach to investigate the dependency between characteristics of missing variability and base load information, and predict characteristics of the missing variability in a low-resolution load profile. It is applicable to both measured and modeled time-series data. To predict this information, this research first hypothesizes that various statistics of this variability may be correlated to the base load, as well as to other characteristics of the signal. Therefore, if the function describing this relationship is known, statistics of the missing variability in a low-resolution signal can be predicted.

The major contribution and novelty of this work focuses on the development of a framework methodology to train regression models for the prediction of missing variability characteristics in both modeled and measured time series data.

The remainder of this work is organized as follows: In **section II**, the research methodology, decomposition, quantification methods, and regression model are discussed. Then in **section III**, results from different models are presented and analyzed, and **section IV** concludes this work.

II. PROBLEM FORMULATION

The proposed framework in this research is part of ongoing work and it is divided into two phases of discovery and implementation (shown in Figure 2). The discovery phase begins with a decomposition process previously discussed in [1] and [26], applied to a subset of available high-resolution measured data, to separate the variability signal from the base load. Next, the variability signal is analyzed and characterized by various metrics such as root-mean-square-variability (RMSV) [1] that summarize the typical deviation of the base load over a given period of time, or by other metrics that are related to its statistical distribution. This represents an important distinction from other types of high-resolution analysis, in that the variability is not directly modeled as a time-series signal, but rather characterized by these scalar metrics. Using the base load metrics as input features, and variability metrics as response variables, the functional relationship between them is determined using multivariate regression.

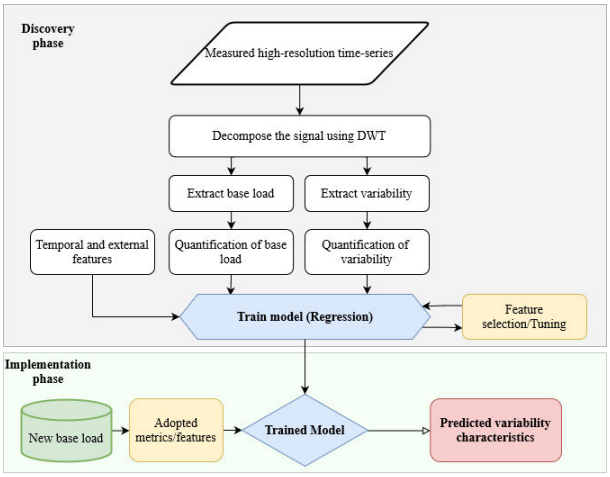


FIGURE 2. Research framework.

In the implementation phase, it is assumed that only a low-resolution load profile is available. By using the generated model with this load profile as the input, the missing variability characteristics can be predicted. The following subsections describe the process in more detail, using an actual building load profile as a case study.

A. CASE STUDY AND DATA DESCRIPTION

The example in this paper focuses on the prediction of variability characteristics in a single building, by training a model over a different time period in that same building. The data set used is a 1-minute time-resolution load profile of a cafeteria

building in the U.S. Midwest collected over two consecutive years, 2018 and 2019. The data set includes the measured load profile for the entire building consumption as well as load profiles of five end-use categories that form the full building consumption: mechanical, "cooking 1", "cooking 2", lighting, and miscellaneous electrical loads. The instantaneous power demand is recorded once per minute, and for the purposes of this analysis the power readings are treated as equivalent to the average demand over that minute. In reality, this results in an overestimation of the energy variability at the one-minute timescale, as the variance of instantaneous power is larger than that of power averaged over longer timeframes. Conversely, it is an underestimation of the variability in instantaneous power, due only to being sampled once per minute. Thus, the magnitude of variability in this case study lies between the variability in instantaneous power and one-time energy demand. It is noted that while the relationship between the base load and the actual one-minute energy variability may be different from that obtained through this assumption, the framework for generating a model at any timescale remains the same. Data preprocessing was performed to replace missing data with representative data using an interpolation method. Less than 0.5% of the data required such replacement, and for no contiguous periods longer than 1 hour.

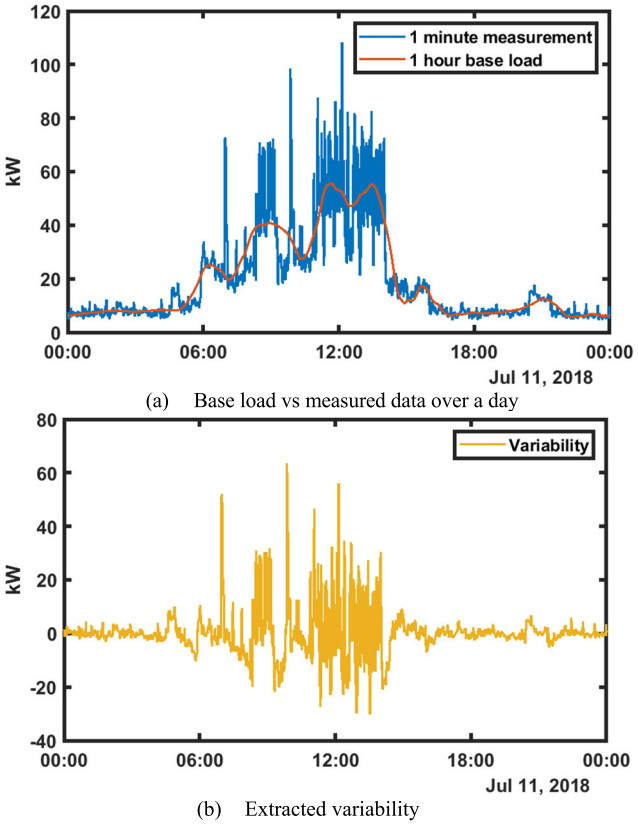
To validate the research framework using this case study, it is supposed that only low-resolution load profile data is available for 2019. Specifically, the 1-minute data is aggregated to hourly intervals, a common resolution for both commercial and residential metering. The model is trained on 2018 data to predict hourly variability characteristics from the hourly load profile, following the discovery framework in Figure 2. The resulting model is then applied to the hourly 2019 data (following the implementation framework) to make hourly predictions of 2019 variability characteristics. These predictions are tested against the actual high-resolution 2019 data. Table 1 presents a table of the various variables used in the following sections.

B. SIGNAL DECOMPOSITION

The first step in developing a model that relates high-resolution load profile characteristics to low-resolution characteristics is to separate these two components of the signal. Previous research by the authors in [1] proposes the use of the Discrete Wavelet Transform (DWT). The DWT process separates a signal into an approximation signal and a detail signal, corresponding to the base load and variability, respectively. For the present case study, a level 6 Daubechies-4 (db4) wavelet function is applied to the 1-minute time-resolution measured data. The choice of level 6 corresponds to a frequency cutoff between base load and variability with a period of 64 minutes. Figure 3 shows a sample of separated variability and base load signals over a day after applying DWT. As can be seen from this figure, variability is the difference between the high-resolution measured data and the base load.

TABLE 1. Table of variables.

P	High-resolution load profile, $P = [P_1, P_2, .. P_m]$ at minutes $m=1...M$, used for training
P_{LR}	Low-resolution load profile, $P_{LR} = [P_{LR,1}, P_{LR,2}, \ldots P_{LR,H}]$ at hours $h=1...H$, used for testing
w	DWT wavelet ('db4' in this case study)
l	DWT decomposition level, where L is the maximum level (6 in this case study)
D	Time-domain detail components of DWT decomposition
A	Time-domain approximation component of DWT decomposition
V	Variability signal for a given w and L
SM	Matrix of hourly statistical metrics of variability, $SM = [SM_1, SM_2, .. SM_H]$ at hours $h=1...H$
B	Matrix of base load metrics used as model inputs
G	Matrix of additional load information used as model inputs
f	The functional relationship established by a machine learning model



(a) Base load vs measured data over a day

(b) Extracted variability

FIGURE 3. Decomposition output as a base load and variability.

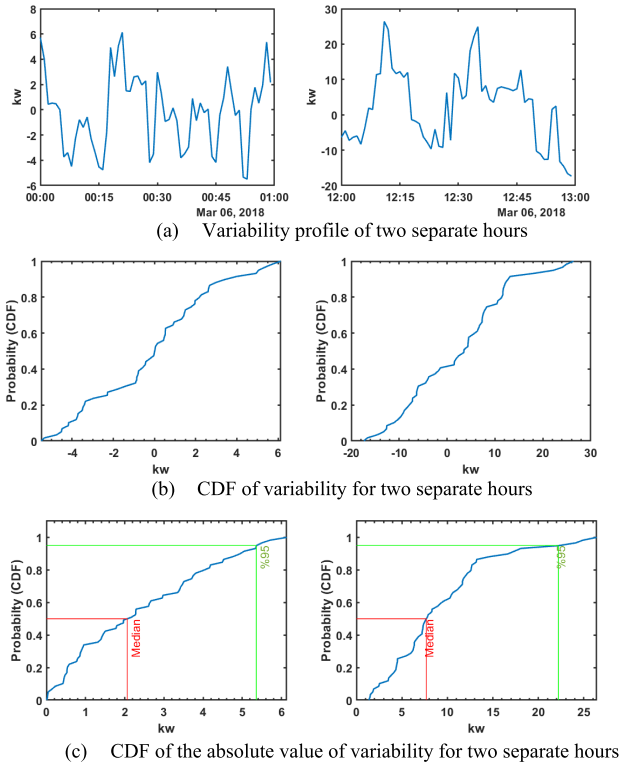
C. FEATURE EXTRACTION AND STATISTICAL QUANTIFICATION

This framework does not seek to predict an actual variability signal in the time domain. Rather, the variability is considered

to be a random signal with certain quantifiable metrics. Several metrics can be used to describe the statistical or frequency content of such a signal, and the importance of various metrics is wholly dependent on the application.

In this case study, the variability signal is separated into hourly segments, corresponding to the resolution of the hourly base load. Figure 4a shows the variability signal over two such hours on the same day - one during off-peak hours during which variability is low, and one during peak hours during which variability is high. Figure 4b shows the Cumulative Distribution Function (CDF) of the variability over each of these hours. The CDF allows for the visualization of not only the minimum and maximum values, but any percentiles.

Because high-resolution energy measurements must sum to low-resolution measurements, the mean of the variability signal is near zero by design, with slight deviations due to the smoothing of the DWT filter. To visualize the magnitude of variability without respect to sign, the CDF of the absolute variability over each hour is shown in Figure 4c, with the median absolute variability and 95th percentile marked for each hour.



(a) Variability profile of two separate hours

(b) CDF of variability for two separate hours

(c) CDF of the absolute value of variability for two separate hours

FIGURE 4. Daily variability and statistical characteristics of two separate hours of March 6th of 2019.

In the subsequent analysis, the variability signal over each hour is characterized by six metrics. The median absolute variability, the mean absolute variability, and the Root-Mean-Square Variability (RMSV) all describe the ''typical'' deviation of the high-resolution signal from the base load, with an increasing sensitivity to skewness and outliers. The RMSV is almost identical to the standard deviation of the variability, except that it measures the RMS distance from the base load, rather than the variability signal's mean. The maximum and minimum variability are simply the extremes of the variability over each hour, which may be especially relevant to applications involving peak demand or solar integration. Finally, the 95th percentile of the absolute variability disregards outliers, and constitutes a 95% confidence bound for the absolute deviation of the high-resolution signal from the base load. Table 2 displays these metrics for the two example hours in Figure 4. These six metrics constitute the response variables in the regression models and are calculated for every hour.

TABLE 2. Variability metrics for two different hours.

Variability Metric (kW)	12am (00:00-00:59)	12pm (12:00-12:59)
Median absolute	2.06	7.68
Mean absolute	2.34	9.1
RMS	2.92	10.76
Maximum	6.13	26.4
Minimum	-5.49	-17.36
95th percentile absolute	5.21	23.14

D. REGRESSION INPUT FEATURES

While the base load is the primary input for the proposed predictive model, it is not the only information that can be leveraged for the prediction of variability. Variability could have a relationship to temporal variables that is at least partially distinct from its relationship to the base load. In addition, when constructing a model to encompass a large set of possible buildings, various known categorical and numerical building characteristics could serve as model inputs, as well as external data such as weather information. Finally, additional features may be engineered from available information, which could correlate more strongly with variability than the raw data itself. In the present case study for a single building, several possible inputs are considered, falling into 3 categories:

1. Temporal - These categorical variables describe the time of day, day of week, etc.
2. Base Load Metrics - in addition to the base load for a given hour, statistics of the base load over the day, week, and month are considered as well.
3. Engineered Features - These are mathematical combinations of other features, such as normalizing the hourly base load by the daily mean.
4. External - Relevant data obtained from other sources, specifically an hourly temperature profile.

Table 3 shows all of the input features tested in this case study. In total, 15 features are considered.

E. RESEARCH ALGORITHMS

Algorithms 1 and 2 formally summarize the research framework outlined above.

TABLE 3. Regression input features.

Temporal (Categorical)	Day of week, Month, Week, Hour of day, Weekend/Weekday binary, Business Day binary
Base load metrics (Numerical)	Hourly Mean, Monthly Mean, Daily Mean, Weekly Mean
Engineered features (Numerical)	Abs Difference hourly Mean, Hourly Mean scaled by daily Mean, Hourly Mean scaled by monthly Mean, Hourly Mean scaled by weekly Mean
External information (Numerical)	Hourly Temperature

Algorithm 1 Discovery Phase

1: Given a high-resolution load profile P, measured each minute:

2: Decompose the signal using wavelet w (w = db4) at maximum level L ($L = 6$)

3: Let D and A denote the time-domain detail and approximation components of P, such that for maximum level L, wavelet w, and time t

$$P_t = \sum_{l=1}^{L} D_{l,w,t} + A_{L,w,t}$$

4: Define variability V at level L as the sum of details up to level L

$$V_{L,w,t} = \sum_{l=1}^{L} D_{l,w,t}$$

5: Quantify the variability V using the statistical metrics SM for each hour

$$SM_h = \begin{bmatrix} V_{L,w,h}^{|Median|}, V_{L,w,h}^{|Mean|}, V_{L,w,h}^{RMS}, \\ V_{L,w,h}^{Max}, V_{L,w,h}^{Min}, V_{L,w,h}^{|95\%|} \end{bmatrix}$$

6: Quantify the base load metrics B by calculating statistical metrics of A over various time frames

7: Train a model f using each statistical metric SM, base load metrics B, and other available information G

$$SM = f(B, G)$$

8: Repeat step 7 as needed using alternative models or input features, to determine the best-performing models

Algorithm 2 Implementation Phase

1: Given a new low-resolution load profile P_{LR}, measured each hour:

2: Quantify the base load metrics B_{LR} by calculating statistical metrics of P_{LR} over various time frames

3: Predict the statistical metrics of variability $\hat{SM}$ using the trained model f for each metric

$\hat{SM} = f(B_{LR}, G_{LR})$

4: Measure the model's performance error using R^2, RMSE, and MAE

$$error_h = \hat{SM}_h - SM_h^{Actual}$$

$$R^2 = 1 - \frac{\sum_h (error_h)^2}{\sum_h \left(SM_h^{Actual} - \bar{SM}_h^{Actual}\right)^2}$$

$$\text{RMSE} = \sqrt{\frac{\sum_h (error_h)^2}{H}}$$

$$\text{MAE} = \frac{\sum_h |error_h|}{H}$$

III. ANALYSIS OF RESULTS

A. LINEAR REGRESSION MODEL

As discussed in the introduction, the underlying hypothesis in this work is that a functional relationship exists between various characteristics of the variability and the base load. Preliminary analysis examined one of the simplest tests of this hypothesis - a univariate linear regression for a single variability metric as a function of the base load. Figure 5 shows the hourly mean absolute variability vs. the base load for all of 2018. The black line represents the best fit using linear regression on the full data set, with an R^2 of 0.88 and an RMSE of 1.64 kW.

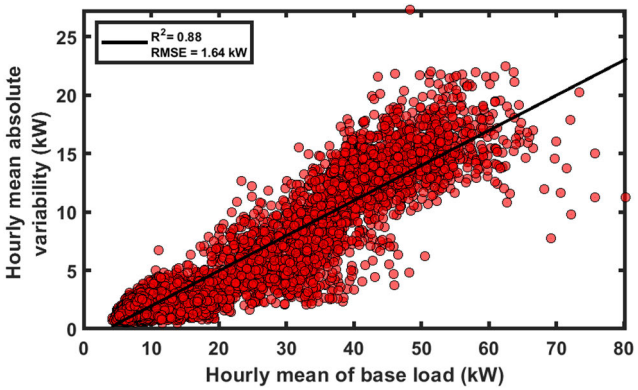


FIGURE 5. Linear regression of hourly mean absolute variability as a function of hourly mean base load.

This linear model trained on individual building data is not a robust predictive model for variability, but it offers a tentative confirmation of the hypothesis that variability metrics are indeed strongly correlated with the base load. At the same time, the non-zero intercept of the fit line contradicts even simpler assumptions about variability, particularly that it can be assumed to be a fixed percentage of the base load.

B. SUMMARY OF MODEL PERFORMANCE

To address the potentially nonlinear dependence of variability on the full space of numerical and categorical inputs, a broad range of machine learning regressions are evaluated, including Decision Trees, Support Vector Regression, Tree-based

TABLE 4. Performance accuracy of multivariate regression models.

Model / Response Variables	*Testing Metric*	Linear Regression	Decision Tree	SVM-Quadratic	SVM-Cubic	Ensemble-Boosted Tree	Ensemble-Bagged Tree	Neural Network
|Median|	R^2	0.83	0.79	0.83	0.80	0.87	0.87	0.75
	RMSE	1.86	2.06	1.84	2.03	1.65	1.62	2.26
	MAE	3.47	1.16	1.12	1.15	0.94	0.92	1.24
|Mean|	R^2	0.88	0.88	0.88	0.87	0.91	0.92	0.87
	RMSE	1.65	1.69	1.65	1.73	1.42	1.38	1.74
	MAE	1.18	1.02	1.06	2.99	0.85	0.82	1.02
RMSV	R^2	0.88	0.84	0.87	0.86	0.90	0.91	0.88
	RMSE	2.08	2.35	2.15	2.18	1.88	1.79	2.09
	MAE	1.46	1.49	1.32	1.28	1.07	1.04	1.26
MAX	R^2	0.74	0.70	0.73	0.71	0.78	0.79	0.74
	RMSE	7.69	8.27	7.93	8.16	7.13	6.96	7.66
	MAE	4.92	4.61	4.44	4.56	3.85	3.82	4.17
MIN	R^2	0.90	0.87	0.90	0.89	0.92	0.92	0.89
	RMSE	2.71	3.00	2.65	2.75	2.35	2.34	2.80
	MAE	1.91	1.88	1.75	1.79	1.51	1.48	1.81
|0.95|	R^2	0.79	0.75	0.77	0.76	0.82	0.82	0.79
	RMSE	5.44	5.95	5.67	5.80	5.02	4.99	5.51
	MAE	3.05	3.09	3.09	3.16	2.62	2.59	2.86

Ensemble methods like Bagged Tree, and Neural Networks. The 2018 data set is decomposed into variability and base load signals, and the input and response variables outlined in Section II are calculated for each hour of the year. These features are used to train and validate the machine learning models, using 5-fold cross validation for hyperparameter tuning where applicable.

To test the models (implementation phase), the hourly aggregated 2019 data is used as the model input, and predictions of each variability metric are made for every hour of the year. These are compared to the actual variability metrics obtained through decomposition of the high-resolution 2019 data.

Table 4 presents a summary of the performance of each machine learning algorithm, as well as a multivariate linear model (which excludes the categorical inputs), for each variability metric. The performance of each model is quantified through the three error metrics R^2, RMSE, and MAE, as defined in Algorithm 2.

Neither the chosen algorithms nor the evaluation metrics constitute an exhaustive set, but have been selected to present a range of bias and variance trade-offs, as well as sensitivity to outliers. Overall, the Ensemble Bagged Tree algorithm produced the lowest error metrics for the 2019 test case, and the following subsections examine the performance of this model in more detail.

C. PERFORMANCE ACROSS VARIABILITY METRICS

The R^2 values for the prediction of each variability metric are shown in Figure 6. The maximum of variability exhibits the lowest predictive accuracy, followed by the 95% confidence bound on absolute variability. These results indicate that the upper extremes of this building's variability are more unpredictable than its long-term averages. Notably, the minimum of variability exhibits a similar predictive accuracy to the mean and RMS value, as for this particular load profile, short-term drops in demand are less extreme and more predictable than short-term increases. Overall, these R^2 (and RMSE) values still indicate a relatively strong functional relationship between these variability metrics and the input features.

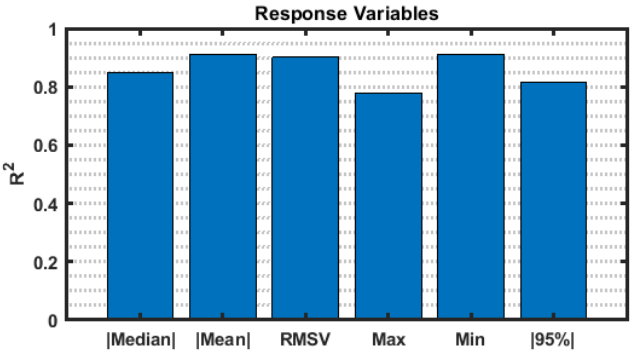


FIGURE 6. Reported R^2 of various response variables including hourly mean absolute, hourly max, hourly absolute 95%, hourly standard deviation (root mean squared), hourly median absolute, and hourly min.

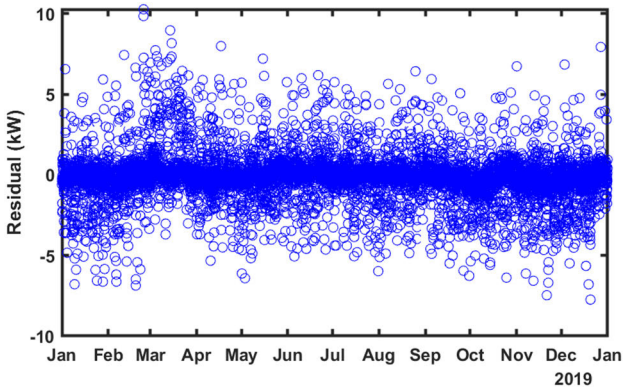
Subsequent analysis will focus on prediction of mean absolute variability, though similar patterns exist for each of the variability metrics.

D. RESIDUAL ANALYSIS

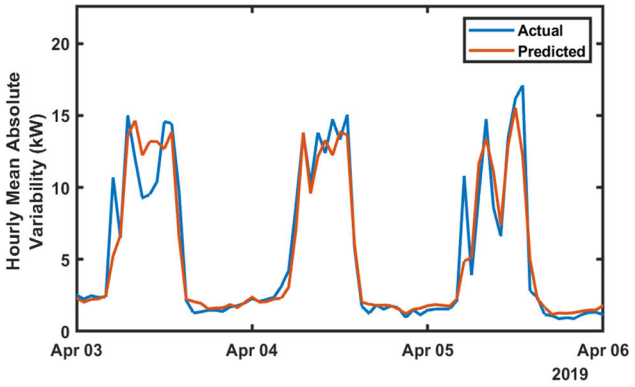
Figure 7a shows the residuals of the model predictions for mean absolute variability over the full year of 2019, while Figure 7b shows a sample and the actual and predicted values vs. time, over three days. Overall, there are no obvious undesired patterns in the residuals over the course of the year, with the exception of a visible positive skewness to the highest residuals in the month of March. This is likely due to a change in the building's load profile during March of 2018, in which the "cooking2" end use category was largely unused. By including the month as a categorical input, the model is trained to associate this anomalous behavior with the month of March, and incorrectly predicts it to repeat in 2019. Aside from this skewness, the overall accuracy in the month of march is nearly identical to the rest of the year.

To analyze patterns with respect to the magnitude of the base load, Figure 8 shows the predicted and actual mean absolute variability for each hour, plotted vs. the hourly base load, analogous to the linear regression in Figure 5. With an R^2 of 0.92 and a RMSE of 1.38 kW, this model shows a

quantifiable improvement over the linear regression, visibly following the slope changes in the relationship between the variability and base load.



(a) Residual plot



(b) Predicted vs. actual hourly mean absolute variability over three days

FIGURE 7. Results for an Ensemble Tree model trained on 2018 data, predicting hourly mean absolute variability in 2019. a) Residuals vs. time and b) predicted vs. actual values over three days in April.

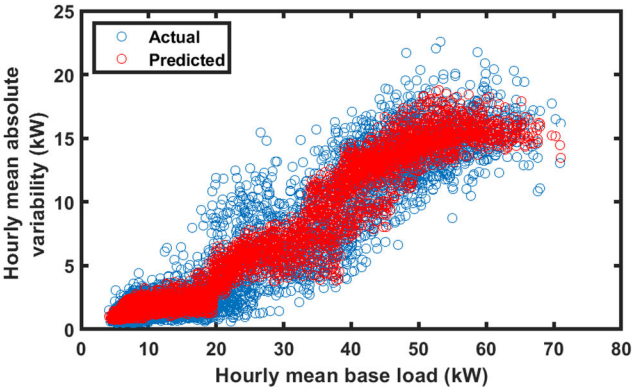


FIGURE 8. Predictions vs. actual results of hourly mean absolute variability vs hourly mean base load using the Ensemble Bagged Tree regression method (Trained on 2018 and tested on 2019 data).

E. IMPACT OF TEMPORAL FEATURES

The impact of the various temporal features on predictive accuracy can be examined by calculating a performance

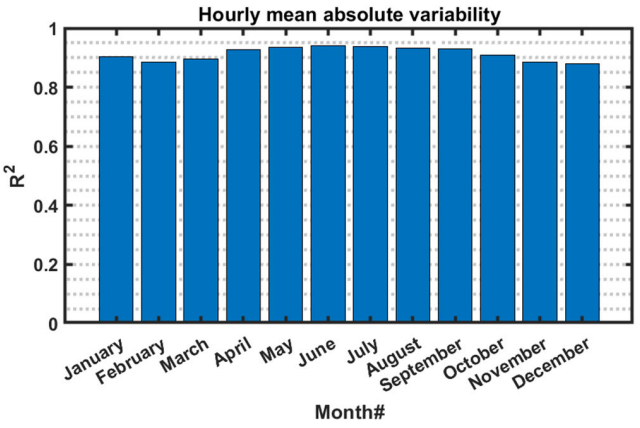


FIGURE 9. R^2 of predicted mean absolute variability, categorized by Month.

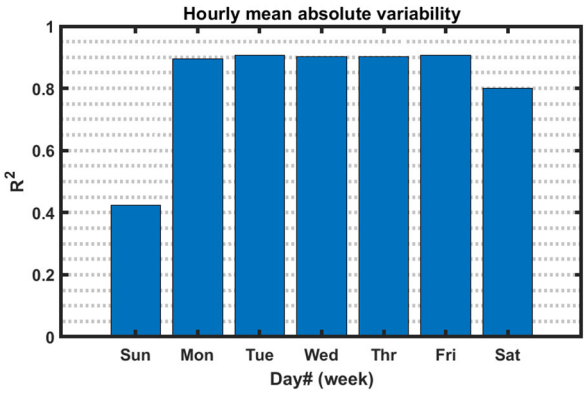
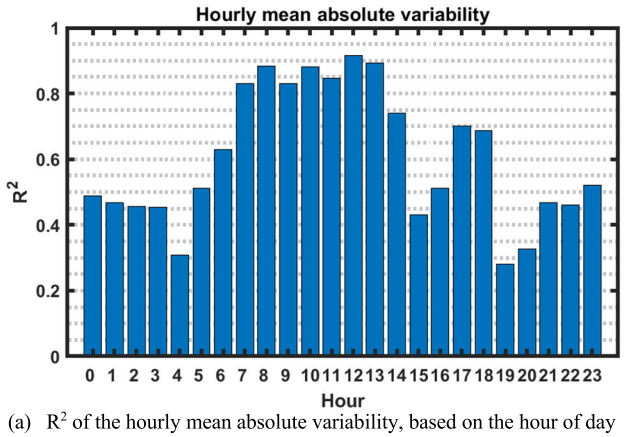


FIGURE 10. R^2 of predicted mean absolute variability categorized by Day of Week.

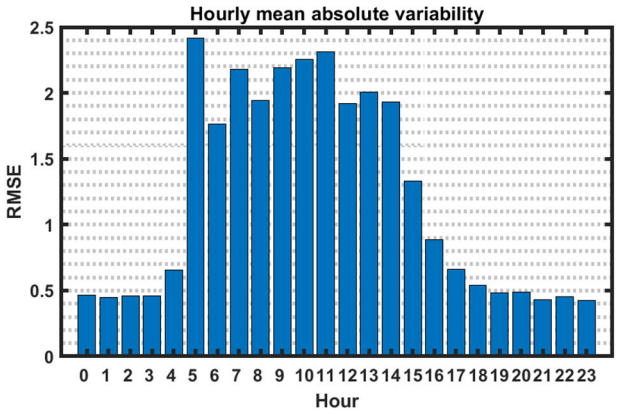
metric such as R^2 separately over various temporal categories in the full data set. Figure 9 shows the R^2 of mean absolute variability for each month. Similar to examination of the residuals in Figure 9, there are no clear monthly patterns, and accuracy seems to be similar in each month of the year, including the anomalous month of March.

Figure 10 shows the R^2 of mean absolute variability when the data is sorted by day of week, starting with Sunday. Here, all weekdays exhibit a similar predictive accuracy, with a small decrease on Saturdays, and a large decrease on Sundays. This corresponds with the fact that for this building load profile, while there is some level of activity on some Saturdays, Sundays are universally days of both low base load and low variability.

This relationship between normalized accuracy metrics (such as R^2) and periods of high and low demand can be illustrated by examining the most significant temporal variable, the time of day. Building load profiles are characterized by large changes in demand over the course of the day, and the hourly variability metrics exhibit similar patterns. Figure 11a shows the R^2 of each hour of the day (calculated from that hour's residuals on every day of the year), while Figure 11b shows the RMSE for each hour.

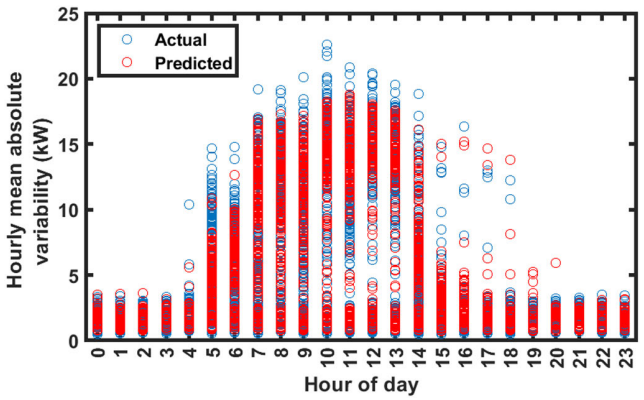


(a) R^2 of the hourly mean absolute variability, based on the hour of day

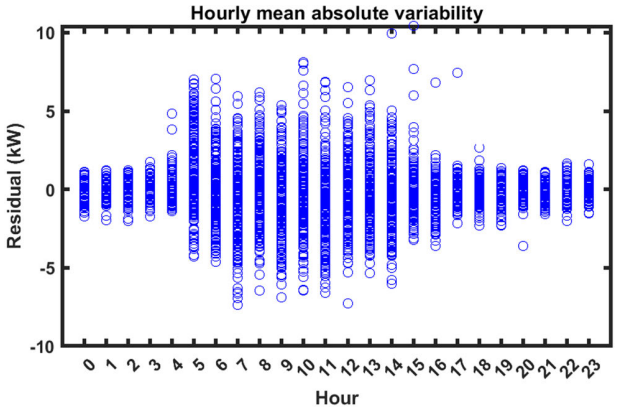


(b) RMSE of the hourly mean absolute variability, base on hour of day

FIGURE 11. Comparison of normalized and absolute error metrics (R^2 and RMSE) based on hour of the day.



(a) Predictions vs. actual results of hourly mean absolute variability based on the hour of day



(b) Residuals vs. hour of day

FIGURE 12. Results for an Ensemble Tree model Trained on 2018 data, predicting hourly mean absolute variability in 2019.

It can be seen in Figure 11b that the absolute error - the RMSE - is generally higher during daytime hours, and lower during nighttime hours. Conversely, in Figure 11a, the R^2 indicates a lower error (higher R^2) during daytime hours, and a higher error during nighttime hours.

This discrepancy arises from the large differences in the magnitude of the target variable (mean absolute variability) over each timeframe. Figure 12a shows the actual and predicted values for the entire year, binned by hour of the day, and Figure 12b shows the residuals for each hour. During off-peak hours, the absolute model error is low, but the target variable is low as well. During these time periods, a higher percentage of the variability's variance is still "random", or at least uncorrelated to the input features used in the model. As the absolute magnitude of the variability increases during peak hours, a higher percentage of its variance is correlated to the input features, resulting in a lower R^2.

As a final note, there is a sharp decrease in R^2 during the transition hours between peak and off-peak - specifically 4-6am in the morning, and 6-7 pm in the evening. This may reflect an actual increase in the volatility of high-frequency signal characteristics during this time period, but it is also in part an artifact of separating high and low frequency components of the load profile during periods of rapid base load change. When the load is much higher at the end of an hour than the beginning of the hour, the one-minute loads are often very far from the hourly average load, resulting in artificially large values for variability metrics. Alternate approaches to interpolating base loads between hourly measurements may partially alleviate this discrepancy.

IV. CONCLUSION

In this ongoing research, a novel data-driven approach to predict characteristics of the missing variability in a base load signal is proposed. Based on the introduced framework, the relationship between various statistics of variability and base load is investigated, and it is found that a stable relationship may exist between them for a given building. For the building load profile in this case study, the relationship between the response variables and the most significant inputs is primarily linear. However, an Ensemble Bagged Tree regression model exhibits a notable increase in accuracy. Across the six variability metrics examined, R^2 values range from 0.79 to 0.92, and RMSE values are typically less than 20% of the response variable mean.

While several input features are considered in this case study as examples of available data that may affect variability, many are highly correlated with the base load and are

largely redundant. Feature impact, selection, and engineering are far more important for the more difficult task of training generalized models applicable to multiple building types, where some of these normalized variables may be useful in distinguishing buildings. Further research will apply this framework to larger datasets comprising disparate buildings and load profiles, to quantify the dependence of variability on building characteristics. The specific machine learning algorithms included in this research are not presented prescriptively, but are representative of the many options available for regression. With a robust set of high-resolution data, meaningful conclusions can be drawn about the accuracy and trade-offs of various models. Future studies will assess the ultimate efficacy of this approach in comparison to alternative methods such as compressed sensing or uncertainty propagation.

Overall, despite the many factors that might cause short-term metrics of day-to-day load behavior to differ between years of building operation, these preliminary results speak to the tractability of predicting variability characteristics by leveraging base load information in machine learning models.

REFERENCES

[1] A. Parker, S. Moayedi, K. James, D. Peng, and M. A. Alahmad, "A case study to quantify variability in building load profiles," *IEEE Access*, vol. 9, pp. 127799–127813, 2021, doi: 10.1109/ACCESS.2021.3112103.

[2] X. Zhu and B. Mather, "DWT-based aggregated load modeling and evaluation for quasi-static time-series simulation on distribution feeders," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2018, pp. 1–5, doi: 10.1109/PESGM.2018.8585535.

[3] E. Proedrou, "A comprehensive review of residential electricity load profile models," *IEEE Access*, vol. 9, pp. 12114–12133, 2021, doi: 10.1109/ACCESS.2021.3050074.

[4] A. Elkholy, A. E.-S.-A. Nafeh, and F. H. Fahmy, "Impact of time resolution averaging analysis on integrated photovoltaic with office buildings and grid interaction metrics: Case study," *Energy Buildings*, vol. 257, Feb. 2022, Art. no. 111818, doi: 10.1016/J.ENBUILD.2021.111818.

[5] S. Cao and K. Sirén, "Impact of simulation time-resolution on the matching of PV production and household electric demand," *Appl. Energy*, vol. 128, pp. 192–208, Sep. 2014, doi: 10.1016/j.apenergy.2014.04.075.

[6] F. Bu, K. Dehghanpour, and Z. Wang, "Enriching load data using micro-PMUs and smart meters," *IEEE Trans. Smart Grid*, vol. 12, no. 6, pp. 5084–5094, Nov. 2021, doi: 10.1109/TSG.2021.3101685.

[7] A. Pinceti, O. Kosut, and L. Sankar, "Data-driven generation of synthetic load datasets preserving spatio-temporal features," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2019, pp. 1–5, doi: 10.1109/PESGM40551.2019.8973532.

[8] A. J. R. Reis and A. P. A. D. Silva, "Feature extraction via multiresolution analysis for short-term load forecasting," *IEEE Trans. Power Syst.*, vol. 20, no. 1, pp. 189–198, Feb. 2005, doi: 10.1109/TPWRS.2004.840380.

[9] C. Bianchi, L. Zhang, D. Goldwasser, A. Parker, and H. Horsey, "Modeling occupancy-driven building loads for large and diversified building stocks through the use of parametric schedules," *Appl. Energy*, vol. 276, Oct. 2020, Art. no. 115470, doi: 10.1016/J.APENERGY.2020.115470.

[10] K. Ashok, M. J. Reno, L. Blakely, and D. Divan, "Systematic study of data requirements and AMI capabilities for smart meter analytics," in *Proc. IEEE 7th Int. Conf. Smart Energy Grid Eng. (SEGE)*, Aug. 2019, pp. 53–58, doi: 10.1109/SEGE.2019.8859916.

[11] U. S. Energy Information Administration (EIA). *How Many Smart Meters are Installed in the United States, and Who Has Them?* Accessed: May 10, 2022. [Online]. Available: https://www.eia.gov/tools/faqs/faq.php?id=108&t=3

[12] J. Peppanen, M. Hernandez, J. Deboever, and M. Rylander, "Enhanced load modeling—Leveraging expanded monitoring and metering," EPRI, Palo Alto, CA, USA, Tech. Rep. 3002015283, 2019.

[13] L. Blakely, M. J. Reno, and K. Ashok, "AMI data quality and collection method considerations for improving the accuracy of distribution models," in *Proc. IEEE 46th Photovoltaic Spec. Conf. (PVSC)*, Jun. 2019, pp. 2045–2052, doi: 10.1109/PVSC40753.2019.8981211.

[14] J. A. Azzolini and M. J. Reno, "Impact of load allocation and high penetration PV modeling on QSTS-based curtailment studies," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Jul. 2021, pp. 1–5, doi: 10.1109/PESGM46819.2021.9638101.

[15] EnergyPlus. *EnergyPlus*. Accessed: May 10, 2022. [Online]. Available: https://energyplus.net/

[16] H. Software. *Generating Synthetic Load Data*. Accessed: Jan. 25, 2022. [Online]. Available: https://www.homerenergy.com/products/pro/docs/latest/generating_synthetic_load_data.html

[17] DOE2. *DOE2.com Home Page*. Accessed: Aug. 15, 2022. [Online]. Available: https://www.doe2.com/

[18] ESP-r. *Welcome to ESP-r*. Accessed: Aug. 15, 2022. [Online]. Available: https://www.esru.strath.ac.uk//Courseware/ESP-r/tour/

[19] J. Inga-Ortega, E. Inga-Ortega, C. Gomez, and R. Hincapie, "Electrical load curve reconstruction required for demand response using compressed sensing techniques," in *Proc. IEEE PES Innov. Smart Grid Technol. Conf.-Latin Amer. (ISGT Latin America)*, Sep. 2017, pp. 1–6, doi: 10.1109/ISGT-LA.2017.8126731.

[20] D. Pinos-Mendez and J. Inga, "Compressed sensing and P1P in electrical consumption prediction," in *Proc. Int. Conf. Inf. Syst. Comput. Sci. (INCISCOS)*, Nov. 2018, pp. 158–164, doi: 10.1109/INCISCOS.2018.00030.

[21] G. Zhou, M. Bai, X. Zhao, J. Li, Q. Li, J. Liu, and D. Yu, "Study on the distribution characteristics and uncertainty of multiple energy load patterns for building group to enhance demand side management," *Energy Buildings*, vol. 263, May 2022, Art. no. 112038, doi: 10.1016/j.enbuild.2022.112038.

[22] X. Zhu and B. Mather, "Data-driven load diversity and variability modeling for quasi-static time-series simulation on distribution feeders," in *Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM)*, Aug. 2019, pp. 1–5, doi: 10.1109/PESGM40551.2019.8973929.

[23] S. E. Kababji and P. Srikantha, "A data-driven approach for generating synthetic load patterns and usage habits," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 4984–4995, Nov. 2020, doi: 10.1109/TSG.2020.3007984.

[24] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Trans. Smart Grid*, vol. 10, no. 1, pp. 841–851, Jan. 2019, doi: 10.1109/TSG.2017.2753802.

[25] Z. Dai and J. E. Tate, "A data-driven load fluctuation model for multi-region power systems," *IEEE Trans. Power Syst.*, vol. 34, no. 3, pp. 2152–2159, May 2019, doi: 10.1109/TPWRS.2018.2882560.

[26] A. Parker, K. James, D. Peng, and M. A. Alahmad, "Framework for extracting and characterizing load profile variability based on a comparative study of different wavelet functions," *IEEE Access*, vol. 8, pp. 217483–217498, 2020, doi: 10.1109/ACCESS.2020.3042125.

SAM MOAYEDI received the B.Sc. degree in electrical engineering from Azad University, in 2009, the M.Sc. degree in electrical engineering from Wichita State University, in 2014, and the Ph.D. degree in architectural engineering from the University of Nebraska–Lincoln, in 2022. His research interests include power distribution system analysis/control, real-time remote power monitoring in buildings' electrical systems, and building's electrical load profile analysis and prediction.

ANDREW PARKER is currently working as a Mechanical Engineer with the Commercial Buildings Controls and Analytics Group, National Renewable Energy Laboratory. Since joining the laboratory in 2010, he has been focused on making high-efficiency building design commonplace by making existing energy analysis tools more accessible. His past projects include the development of a now-commercialized mobile energy auditing tool, the development of the OpenStudio energy modeling platform, and the development of various software and analysis tools in support of electric and gas utilities. Recently, he has focused on using OpenStudio to develop a nationwide model of the commercial building stock in the USA, and is particularly interested in the integration of building energy modeling with grid simulation.

KEVIN JAMES received the bachelor's degree in mechanical and electrical engineering from Iowa State University, in 2003, and the master's and Ph.D. degrees in mechanical engineering from the University of Michigan, in 2005 and 2009, respectively. As a Student and a Postdoctoral Researcher with the University of Michigan, he published several articles on underwater acoustic modeling in uncertain environments. He is currently a Part Time Employee with the Durham School of Architectural Engineering and Construction, University of Nebraska–Lincoln. His research interests include electric vehicle usage and variability in electrical grids.

XIAOYUE CHENG received the B.S. and M.S. degrees in statistics from the Renmin University of China, and the Ph.D. degree in statistics from Iowa State University. She is an Associate Professor at the Department of Mathematical and Statistical Sciences, University of Nebraska, Omaha. She has extensive interdisciplinary research experience in a variety of fields, including civil engineering, aviation, agronomy, public administration, K-12 education, psychology, public health, medical clinics, and business marketing. She is the author or the coauthor of several peer-reviewed journal articles, statistical software packages, and visualization dashboard web applications. Her research interests include statistical modeling and simulation, machine learning, data visualization, interactive graphics, image recognition, and exploratory data analysis. She is a member of the American Statistical Association. She has served as an Associate Editor for *The R Journal* and a reviewer for several journals.

MICHAEL HEMPEL (Member, IEEE) received the Ph.D. degree in computer engineering from the University of Nebraska–Lincoln, Nebraska. He is currently working as a Research Assistant Professor at the Advanced Telecommunication Engineering Laboratory (TEL), University of Nebraska–Lincoln. For his research in networking, he has been developing various network simulation and analysis solutions for streaming media and WiFi/WiMAX technologies. He has authored or coauthored more than 150 publications in major international journals and conferences. His research interests include wireless communication protocol design and performance analysis, wireless multimedia services, and distributed computing. He served as a TPC member on numerous international conferences.

HAMID SHARIF (Fellow, IEEE) is the Charles J. Vranek Distinguished Professor with the Department of Electrical and Computer Engineering, University of Nebraska–Lincoln (UNL). He is also the Director of the Advanced Telecommunication Engineering Laboratory (TEL), UNL. He has over 35 years of academic and industrial experience. He has published over 400 research articles in national and international journals and conferences. His research interests include surface transportation communications, including freight rails, cybersecurity, mobile communications security, cyber physical systems, wireless network security, wireless sensors, the IoT networks, vehicular communications, and intelligent transportation. He has served as a PI/a Co-PI for different research projects funded by DoE, DoT, NSF, DoD, and local and national industries, such as Union Pacific. His research has been recognized through several research awards and best papers. He is currently a Distinguished Lecturer of the IEEE Vehicular Technology Society. He has served on many IEEE and other international journal editorial boards.

MAHMOUD A. ALAHMAD (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Idaho, Moscow, in 1989, 1991, and 2005, respectively. He is currently an Associate Professor with the Durham School of Architectural Engineering and Construction, University of Nebraska–Lincoln. His previous publications include several articles in the area of electrified transportation, power management, microbattery testing, and system design implementation. His industry experience includes electrical distribution system infrastructure planning and design for new and renovated facilities. Since 1996, he has been holding a Professional Engineering Licensure. His research interests include electrified transportation, load profile analysis and real-time power monitoring in the built environment, battery power management, and renewable energy alternatives. He is a Peer Reviewer of several journals and conferences, including the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS.