PERFORMANCE OF EXISTING BASELINE MODELS IN QUANTIFYING THE EFFECTS OF SHORT-TERM LOAD SHIFTING OF CAMPUS BUILDINGS

SHUNBO LEI, UNIVERSITY OF MICHIGAN
JOHANNA MATHIEU, UNIVERSITY OF MICHIGAN
RISHEE JAIN, SLAC/STANFORD UNIVERSITY

SLAC-R-1131
SEPTEMBER 2019

PREPARED FOR THE DEPARTMENT OF ENERGY
UNDER CONTRACT NUMBER DE-AC02-76SF00515

Published by

SLAC National Accelerator Laboratory
2575 Sand Hill Road
Menlo Park, CA 94025

Acknowledgments

This work was supported by the U.S. Department of Energy Building Technologies Office under the project *I-DREEM: Impact of Demand Response on short and long-term building Energy Efficiency Metrics,* contract number DE-AC02-76SF00515.

Abstract

Owing to the large thermal inertia of commercial buildings, their heating, ventilation, and air conditioning (HVAC) systems have been extensively studied for providing grid ancillary services by demand response. For example, the power consumption of HVAC system fans can be controlled to track a frequency regulation signal. One critical issue is to estimate the counterfactual baseline power consumption that would have occurred without demand response. Baseline methods have been proposed and analyzed based on whole-building electric load profiles. This report evaluates those methods on the total HVAC fan power profiles, which have different characteristics than whole-building electric load profiles. Specifically, we evaluate the following baseline methods: 1) averaging methods including Y-day simple average, MidXofY, and NearestXofY; 2) regression methods that relate the electric demand to weather and/or other relevant parameters; and 3) a simple linear interpolation method that uses least squares to fit a linear baseline to the fan power data over the 5-minute period just before the demand response event and the 5-minute period immediately after the settling time. We also explore the performance of an additive adjustment applied to the averaging methods. We use minutely power consumption data of HVAC system supply and return fans that were sub-metered in three campus buildings of the University of Michigan during the summers of 2017 and 2018. A quantitative assessment of the evaluated methods' applicability for estimating the baseline fan power is provided, including results of the mean relative absolute error, mean relative bias error, and relative error ratio that quantify the accuracy, bias and variability, respectively. From the numerical results, we conclude that the linear interpolation method can be used to analyze our future demand response experiments, although some improvements may be necessary to further enhance its performance in baselining the fan power during the morning time.

Keywords: Commercial building, Demand response, Load shifting, Baseline model, Fan power, Smart grid.

Please cite this report as follows:

S. Lei, J.L. Mathieu, and R.K. Jain. 2019. *Performance of existing baseline models in quantifying the effects of short-term load shifting of campus buildings.* SLAC-R-XXXX.

Contents

List of Figures

List of Tables

List of Acronyms

AEC	Additional Energy Consumption
BBB	Bob and Betty Beyster building
FERC	Federal Energy Regulatory Commission
HVAC	Heating, Ventilation, and Air Conditioning
MBE	Mean Bias Error
MRAE	Mean Relative Absolute Error
MRBE	Mean Relative Bias Error
RAC	Rackham building
RER	Relative Error Ratio
RTE	Round-Trip Efficiency
WH	Weill Hall

1.0 Introduction

The Federal Energy Regulatory Commission (FERC) defines demand response as "changes in electric usage by demand-side resources from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized" [1]. It is one of the most flexible and effective solutions to reduce power system operational costs and displace generation and network reinforcement [2]. It is also capable of mitigating renewable fluctuations, and enhancing system reliability during periods of high demand [3].

Baseline estimation is a critical task to assess the benefits from demand response programs. Specifically, it estimates the counterfactual power profile that would have occurred without demand response. Accurate baseline estimation methods are necessary to assess the performance of demand response programs, to compensate participants for changing their behavior, and to provide utilities and system operators with a prediction of how much demand-side flexibility is expected, which is useful for different operational and planning problems [4]. It is challenging to measure or calculate what would have occurred without demand response and thus, fundamentally, baselines are imperfect [5]. Specifically, for smaller devices and for more granular end-uses with irregular or unpredictable power consumption, establishing a robust and accurate baseline can be more difficult [6-8]. When a load is dependent upon consumer behavior, it is also typically more difficult to establish a baseline [9].

A variety of baseline methods have been proposed, generally based on whole-building electric load (power) profiles. Those baseline methods can be classified into three categories: averaging methods, regression methods, and control group methods. Averaging methods use the average load of several days selected from recent days without demand response events to estimate the baseline [10, 11]. Regression methods fit a linear or non-linear function to describe the relationship between the load and explanatory variables such as temperature and humidity, and then the baseline can be estimated using this function [12]. Control group methods estimate the baseline by using the load data of non-responsive customers which exhibit the most similar load patterns to the demand response participants [13]. Some baseline methods further incorporate multiplicative or additive adjustments [10].

Commercial buildings account for roughly 20% of the energy consumed in the United States [14]. They are also attractive loads for demand response for two main reasons. First, commercial buildings are normally equipped with relatively sophisticated control and communications architectures, reducing the incremental capital cost of enabling demand response capability [15]. Second, commercial buildings generally have high thermal inertia that can be utilized as energy reservoir for short periods of time, while respecting to the occupants' comfort [16]. Some equipment within a building is more responsive and has a higher demand response potential, while some other equipment is less responsive or even non-responsive. Specifically, the Heating, Ventilation, and Air Conditioning (HVAC) systems of commercial buildings are one of the largest resources of the demand response [17]. A variety of demand response strategies have been developed to directly or indirectly control the supply and return fans of HVAC systems. For example, [18] introduces a feedforward architecture to control the fans in commercial building HVAC systems to provide frequency regulation to the grid.

It may be possible to establish more accurate baseline estimates if, rather than using whole-building electric load data, we use sub-metered load data from the equipment providing demand response. For example, for short-term load shifting (less than one hour) via room temperature setpoint control, we expect a response primarily from the supply and return fans, and secondarily from the chiller(s). In our recent experiments to assess the energy efficiency impacts of short-term load shifting [19, 20], we sub-metered all supply and return fans and applied a simple baseline model to assess the changes in fan power

consumption during demand response events. Not only does sub-metering have the potential to improve baselining but it also allows us to attain a more granular understanding of and insights into the demand response actions. A highly-related problem, short-term load forecasting can attain improved accuracy with sub-metering data, e.g., by distinguishing between weather-sensitive loads and the total loads [21].

In this report, we explore the performance of existing baseline models in estimating total fan power baselines. Since these models were designed for baselining whole-building electric load profiles, their applicability for baselining total fan power profiles is unknown. We evaluate averaging methods and regression methods, and compare their performance to our simple method based on linear interpolation. Control group methods are not evaluated, as we do not have a large enough data set (see Section 2.3.).

The remainder of this report is organized as follows. Section 2 introduces the baseline methods evaluated in this report. Section 3 introduces the evaluation methodology. Section 4 provides the numerical results of our evaluation. Section 5 and 6 discuss the results and conclude this report, respectively.

2.0 Baseline Methods

In this section, we introduce the baseline methods evaluated in this report.

Before introducing the baseline methods, we define demand response days as the days when demand response events occur, and the other days as non-demand response days. Note that weekdays and weekends normally have different load patterns. Our current data set only includes a limited number of weekends. Therefore, we only consider weekdays in this report. Nevertheless, the methods for weekday baseline estimation also apply to weekends, and the same process of baseline method performance evaluation can be conducted on the load data from weekends in the future with a large enough data set.

2.1. Averaging Methods

2.1.1. Y-Day Simple Average Method

The Y-day simple average method uses the average load over the Y most recent non-demand response days before a demand response event to predict the load on the demand response day. In this report, we evaluate the 5-day simple average method, and the 10-day simple average method [10].

2.1.2. HighXofY Method

The HighXofY method considers Y non-demand response days preceding the demand response event. The baseline is estimated using the average load of the X days with the highest electricity consumption within those Y days. In this report, we evaluate the High4of5 (for a weekday in PJM Economic) [22] and High5of10 (for a weekday in NYISO) [23] baseline models.

Note that the demand response program considered in this report is not specifically for peak load shaving on peak days. It is designed for ancillary services that are necessary at any time. The HighXofY method is useful for baselining peak days with high electricity consumption. Therefore, it may not work well here, as we will use it to baseline every non-demand response day. Still, we include this method in our evaluation, since it is frequently used in the industry.

2.1.3. MidXofY Method

The MidXofY method calculates the baseline using X of Y non-demand response days preceding the demand response event. However, it drops (Y-X)/2 days with the lowest electricity consumption and (Y-X)/2 days with the highest electricity consumption, retaining only the X days with median levels of electricity consumption. In this report, we evaluate the Mid4of6 baseline model [22].

2.1.4. LowXofY Method

The LowXofY method calculates the baseline using the X lowest consumption days of Y non-demand response days preceding the demand response event. In this report, we evaluate the Low4of5 and Low5of10 baseline models [24].

Again, as the demand response program considered in this report is designed for ancillary services that are necessary at any time, the LowXofY method can be biased when it is used on non-demand response days with high electricity consumption. Nevertheless, as stated in [25], although this method can have large negative bias, it can also have a high accuracy. Therefore, we also include this method in our evaluation.

2.1.5. NearestXofY Method

The total electricity consumption during a day outside the event window is calculated and used to determine nearness. Thus, among the Y non-demand response days preceding a demand response day, we select X days having the total energy consumption that are closest to the demand response day. The NearestXofY method then uses the average load of these X days as the baseline estimation. In this report, we evaluate the Nearest3of6 and Nearest5of10 baseline methods. We wish to emphasize that, the HighXofY, MidXofY and LowXofY averaging baseline methods calculate and rank the electricity consumption level of a non-demand response day by summing the load over a day. The NearestXofY method only considers the total electricity consumption outside of the demand response event window.

2.1.6. Adjustment Method

Adjustments, including additive adjustments and multiplicative adjustments, are frequently used in baseline estimation to improve accuracy and reduce bias. An additive adjustment adds a fixed amount to the baseline load in each time slot, such that the adjusted baseline is equal to the observed load at a time shortly before the start of the demand response event. A multiplicative adjustment multiplies the baseline load at each time slot by a fixed amount, such that the adjusted baseline is equal to the observed load on average during a time window shortly before the start of the demand response event. Note that additive adjustments are generally preferred to multiplicative adjustments, as the resulting baseline can become volatile under a multiplicative adjustment [26].

In this report, we adopt an additive adjustment. That is, the baseline estimated by the above averaging methods is vertically shifted, so that the baseline load in the first time step of the time window of the demand response event is equal to the measured load. To evaluate the effectiveness of the adjustment method, results of both with and without the additive adjustment are presented.

2.2. Regression Methods

Regression methods are conventionally used for load forecasting, a problem highly related to baseline estimation. Regression methods fit a linear or non-linear function to describe the relationship between the load and explanatory variables including weather variables such as the outdoor temperature, humidity, and wind speed. This function is then used to forecast the load (using predictions of the explanatory variables) or estimate the baseline power (using measured explanatory variables) [12].

Regression methods can work well for whole-building power baseline estimation [21]. Specifically, time of day/week and outdoor ambient conditions (in particular outdoor air temperature) are often strong predictors of whole-building electric load in commercial buildings. However, in this report, we do not evaluate regression methods, as we find that total fan power is not strongly correlated to measured outdoor ambient conditions like outdoor air temperature (see Section 4.2.).

We wish to emphasize that, although the fan power and outdoor air temperature are not strongly correlated, it is possible that the fan power is strongly correlated with some other exogenous variable, for example, occupancy, which impacts internal gains of the building. However, we do not have occupancy data.

2.3. Control Group Methods

To estimate a customer's baseline power consumption on a demand response day, control group methods take advantage of the load curves of other non-responsive customers from the same day [27, 28]. For example, the load curves are first clustered, and the customer's load curve is matched to one of the clusters. Then, the customer's baseline power is calculated by averaging the load curves in the cluster.

More complex fitting methods can also be adopted to attain a weighted combination of the load curves in the cluster [4]. Some control group methods may also use the load curves of the same customer on non-demand response days, and/or the load curves of other non-responsive customers in other days, and/or the load curves of other responsive customers on non-demand response days [11].

Control group methods can work well when the data set involves a large amount of buildings and/or includes historical data over a long period. We only have fan power data from three buildings and their fan power profiles show significantly different patterns. Thus, we do not explore control group methods in this report. It is worth mentioning that, the averaging baseline methods, especially the NearestXofY method, is relatively similar to the control group methods that utilize the load curves of the same customer on preceding non-demand response days.

2.4. Simple Linear Interpolation Method

This baseline method was first proposed in [15]. It estimates the baseline by a simple linear interpolation, with the ends of the linear baseline estimation given by the fan power data just prior to and some settling time after the demand response event. That is, the fan power data is interpolated within a short time window that is immediately preceding the demand response event and after the fans return to their normal operation following the demand response event. This method was also used and improved in [19] and [20]. Specifically, [20] used least squares to fit a linear baseline to the fan power data over the 5-minute period just before the demand response event and the 5-minute period immediately after the settling time. In those papers, this baseline method seems to work reasonably well. This report evaluates it on a much larger data set, and compares it with the aforementioned baseline methods.

3.0 Performance Evaluation Methodology

In this section, we introduce the methodology used to evaluate the performance of the selected baseline methods. We first summarize the data and then explain the evaluation process. After that, the metrics for assessing baseline method errors are introduced. At last, two remarks regarding the evaluation methodology are provided.

3.1. Fan Power Data and Outdoor Temperature Data

Current sensors have been installed in three buildings on the University of Michigan campus to sub-meter HVAC supply and return fans. The three buildings are the Bob and Betty Beyster Building (BBB), the Rackham Building (RAC), and Weill Hall (WH). We use their available fan power data from the summers of 2017 and 2018. Specifically, for BBB, we use non-demand response day data from June 21 to October 31, 2017 and October 3 to October 31, 2018. For RAC, we use non-demand response day data from July 31 to October 29, 2017 and May 1 to October 11, 2018. For WH, we use non-demand response day data from June 10 to October 29, 2017.

The sub-metering data are minutely single-phase current of each HVAC system fan in each building. We assume constant power factors (0.95 for supply fans and 0.99 for return fans) and voltage (275.8 V), which were determined using one week of measured voltage and power factor data, and use these values to estimate the three-phase fan power.

Hourly outdoor air temperature data was downloaded from NOAA's database (Station: ANN ARBOR MUNICIPAL ARPT). The readings are generally at the 53-minute mark of each hour. The data were interpolated to convert to the same resolution as the fan power data.

3.2. Evaluation Process

We use fan power data on non-demand response days, i.e., days without demand response events, to evaluate the baseline methods. If a baseline method is perfectly accurate, the estimated fan power should be exactly the same as the measured fan power data on non-demand response days. By comparing the baseline with the measured fan power data, we can calculate and evaluate baseline model errors. Specifically, we use the baseline methods to estimate the fan power from 9:00 am to 11:00 am and 13:00 pm to 15:00 pm, which correspond to the demand response events on experiment days. (Specifically, on an experiment day, we conducted two demand response events each lasting for 1 hour, i.e., 9:00 am – 10:00 am and 13:00 pm – 14:00 pm, and assume the HVAC system settles back to its baseline operation in the hour after the event.)

Note that, the effectiveness of regression methods depends on the assumption that the electric load is strongly correlated with some exogenous variable(s). In Section 4, we first investigate the correlation between HVAC system fan power and outdoor air temperature. It is shown that the fan power and outdoor air temperature are weakly correlated. This weak correlation limits the effectiveness of outdoor air temperature-based regression methods for baselining fan power. Thus, we do not estimate fan power baselines using regression methods.

3.3. Evaluation Metrics

The performance of the baseline methods is quantified by the following error metrics that evaluate the accuracy, bias and variability.

3.3.1. Mean Relative Absolute Error (MRAE)

The MRAE is calculated by averaging the relative absolute errors between the estimated baseline fan power and the measured fan power over the timeslots of a demand response event. In this report, the minutely relative absolute differences between the baseline and the measurement from 9:00 am to 11:00 am (or 13:00 pm to 15:00 pm) are calculated first. The MRAE is then obtained by taking the mean. (Note that we quantify the errors for the time windows of 9:00 am – 11:00 am and 13:00 pm – 15:00 pm separately, so that we can determine the performance of the baseline methods in these two different time windows.) We can calculate the MRAE for each non-demand response day considered in the evaluation data set, and then take the average to obtain the average MRAE. The average MRAE is then used to quantify the accuracy of a baseline method. It generally represents the absolute value of the difference between the estimation of the baseline method and the actual load. Lower average MRAE values indicate more accurate baseline estimates.

3.3.2. Mean Relative Bias Error (MRBE)

The MRBE is similar to the MRAE except that the absolute value is not taken. Averaging the relative errors between the estimated fan power and the measured fan power over the timeslots of a demand response event attains the MRBE. For the evaluation in this report, we first calculate the minutely relative differences between the baseline and the measurement from 9:00 am to 11:00 am (or 13:00 pm to 15:00 pm), and then take the mean to obtain the MRBE. The average MRBE is calculated by taking the average of MRBEs over all non-demand response days in the evaluation data set. The average MRBE can also be simply referred to as the relative bias. A positive bias indicates over-estimation while a negative bias indicates under-estimation. When the bias is close to zero and the MRAE is larger than zero, it indicates that the baseline method sometimes over-estimates and sometimes under-estimates the baseline, but overall, the over- and under-estimations balance each other out. The bias is more relevant than the accuracy in determining the compensation given to demand response participants. Thus, generally, the closer the MRBE to zero, the better the baseline method.

3.3.3. Relative Error Ratio (RER)

The RER is defined as the standard deviation of the baseline estimation errors expressed as a fraction of the average load during the event period of time. For the evaluation in this report, the minutely relative errors of the estimated baseline fan power with respect to the average measured fan power from 9:00 am to 11:00 am (or 13:00 pm to 15:00 pm) are calculated first, and then used for calculating their standard deviation to obtain the RER. The average RER is attained by taking the mean of RERs over all non-demand response days in the evaluation data set.

The average RER is then used to quantify the variability of a baseline method. Specifically, variability evaluates the robustness of a baseline method under various conditions (e.g., different demand response days). The smaller the average RER, the more stable the baseline method's error.

In Section 4, in addition to providing the average MRAE, MRBE, and RER for each baseline method, we also show box plots of these metrics enabling visualization of the error performance statistics of each method.

3.3.4. Errors Associated with Additional Energy Consumption (AEC) and Round-Trip Efficiency (RTE)

In previous works [15, 19, 20], we used AEC (kWh) and RTE (%) to quantify the building energy efficiency in demand response events. Specifically, AEC is the energy consumed by the building above its baseline minus the energy consumed by the building below its baseline. RTE is the ratio of the energy consumed by the building below its baseline to the energy consumed by the building above its baseline.

For a particular building, with all else equal, a higher AEC indicates a lower RTE, i.e., lower energy efficiency. In this report, to make our results more general, we have adopted metrics including MRAE, MRBE, and RER, which are often used to evaluate baseline method errors. However, in our future work, we will continue using AEC and RTE to analyze the energy efficiency impacts of demand response events and so it is important to understand how MRAE, MRBE, and RER relate to errors in AEC and RTE.

From the definitions of MRBE and AEC, it is clear that the MRBE is closely related to the error associated with the AEC. Specifically, the AEC error is linear to the mean bias error (MBE). With t-minutely load data and a T-minute time window of the event, the ratio between the AEC error and the MBE is $t*T/60$. For example, in this report, the AEC error is twice the MBE. Although the AEC error and MRBE do not have such a strictly linear relationship, the MRBE generally indicates the percentage value of the AEC error.

In presenting the numerical results, we also provide the standard deviation of the MRAE, MRBE, and RER across different days. We also calculate the 95% confidence interval of MRBE, which is related to the confidence interval on the AEC (reported in [23]). The 95% confidence interval is calculated by mean(MRBE) ± 1.96*std(MRBE)/ $\sqrt{n}$, where n is the number of days that the baseline method is tested on and function std takes the standard deviation.

3.4. Remarks

3.4.1. Pearson Correlation Coefficient

As mentioned, in evaluating the regression methods for baseline estimation, we only investigate the correlation between the HVAC fan power and the outdoor temperature. Specifically, the Pearson correlation coefficient, denoted as R, is used. It is a statistical measure that calculates the strength of the linear relationship between the relative movements of two variables. It is determined by dividing the covariance by the product of the two variables' standard deviations. The value of the Pearson correlation coefficient can range between -1.00 and 1.00. A correlation of -1.00 indicates a perfect negative correlation, while a correlation of 1.00 indicates a perfect positive correlation. A correlation of zero indicates no relationship between the movements of the two variables. Experts generally do not consider the correlation significant until its value surpasses at least 0.80.

3.4.2. Under-Estimation of the Errors

We need to emphasize that the error evaluation methods in this report may under-estimate the true baseline model errors. Specifically, we have assumed that the HVAC system settles back to its baseline operation an hour after a demand response event and use that time frame for our error assessment. However, in practice, the settling time of the HVAC system is uncertain and unknown. Therefore, our assumption will affect the accuracy of our error assessment. Moreover, the practical application of some baseline methods including the linear interpolation method depends on an estimate of the settling time, which introduces additional errors. Our error assessment does not capture this error.

4.0 Results

In this section, numerical results for the performance evaluation of baseline methods are provided. Note that the data is separated into five building-years, i.e., BBB-2017, RAC-2017, WH-2017, BBB-2018 and RAC-2018. For example, BBB-2017 uses the data of the BBB building in the summer of 2017, and likewise for other building-years. The baseline methods are evaluated on the data of each building-year. The results are also reported for each building-year.

4.1. Example Time Series Plots

As an example, Figure 1 shows time series plots of the actual load (total fan power) of the RAC building on August 30, 2017, and baselines estimated by the evaluated methods. The additive adjustment is not applied here. Specifically, the upper-left and upper-right plots include the actual and estimated fan power curves during the morning 8:00 a.m.-12:00 p.m. For clarity, each plot includes the actual fan power curve and five estimated baselines. The lower-left and lower-right plots show similar results for the afternoon 12:00 p.m.-16:00 p.m.

Note that the High4of5 and Mid4of6 methods produce the same baseline in this example, as both methods are averaging the same 4 days to obtain the baseline. Thus, their time series curves overlap in the upper-left and lower-left plots of Figure 1. It can also be seen that the High4of5 and High5of10 baseline fan power curves are above the 5-day average and 10-day average baseline curves, respectively, which makes sense given the definitions of these baseline methods. Similar relationships can be found among other baseline fan power curves. In this example, the linear interpolation method (Lin. intrpl.) has the best performance; none of the averaging baseline methods work well.

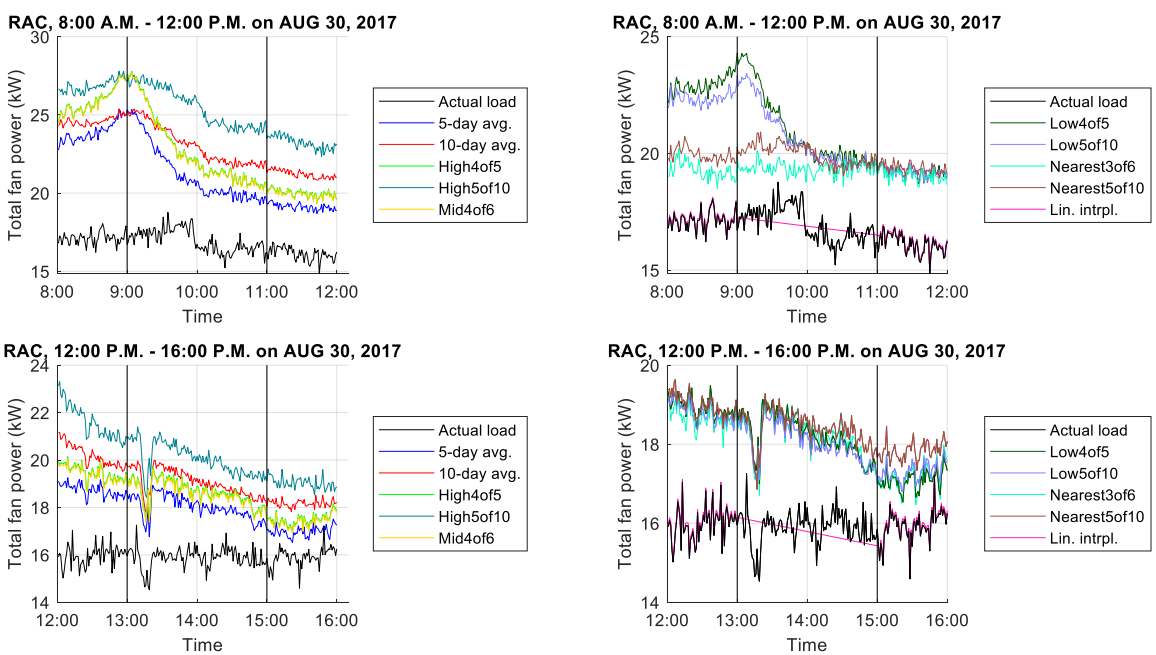

Fig. 1. Example time series plots of the actual load (total fan power) and baselines estimated by the evaluated methods without adjustments. (Upper-left and upper-right: 8:00 a.m.-12:00 p.m. Lower-left and lower-right: 12:00 p.m.-16:00 p.m.)

For the same example, Figure 2 shows the baselines estimated with the averaging methods using the additive adjustment introduced in Section 2.1.6. By aligning the estimated baselines with the actual fan power just prior to the demand response event time window (i.e., at 9:00 a.m. or 13:00 p.m.), some methods perform much better. Specifically, for the morning event time window, the NearestXofY baselines are close to the actual load curve, while the other averaging baselines are still far from the actual fan power curve. As for the afternoon event time window, most baselines are close to the actual load curve, except the 10-day average and High5of10 baselines, which have relatively larger errors.

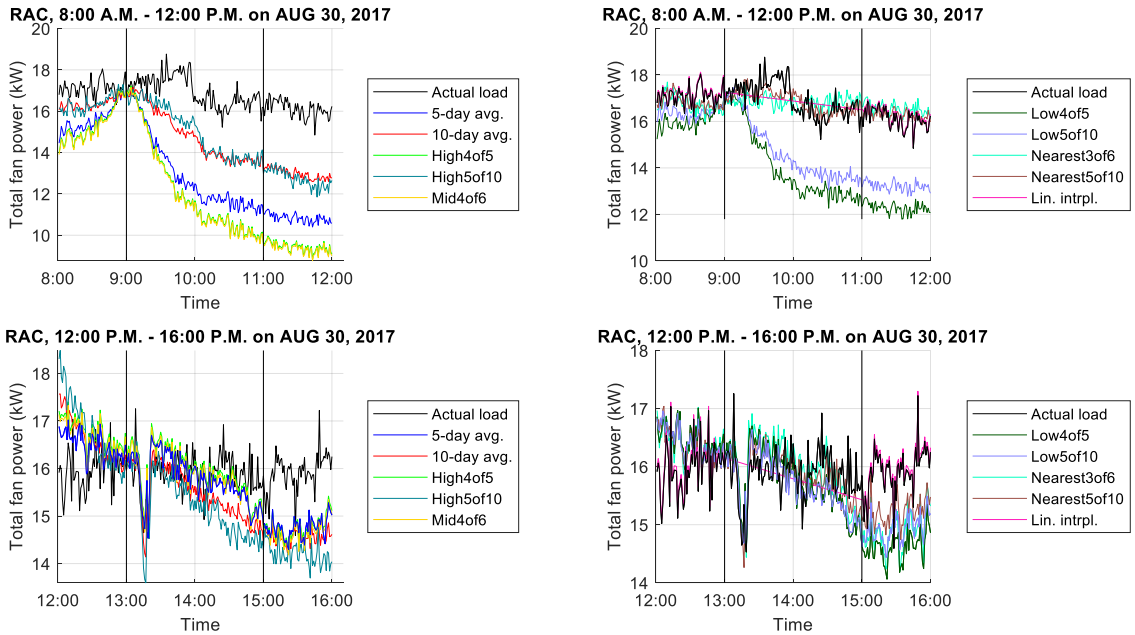

Fig. 2 Example time series plots of the actual load (total fan power) and baselines estimated by the evaluated methods with an additive adjustment. (Upper-left and upper-right: 8:00 a.m. -12:00 p.m. Lower-left and lower-right: 12:00 p.m.-16:00 p.m.)

For the same example, Figure 3 shows the time series errors of the three best baseline methods, i.e., the linear interpolation method, the Nearest3of6 method, and the 5-day simple average method. (Note that these methods are the best across all days, but not necessarily the best for this specific day.) Except for the 5-day average method applied to the morning event time window, the three best baseline methods have small MRBE values (less than 1%) and MRAE values (around 3%).

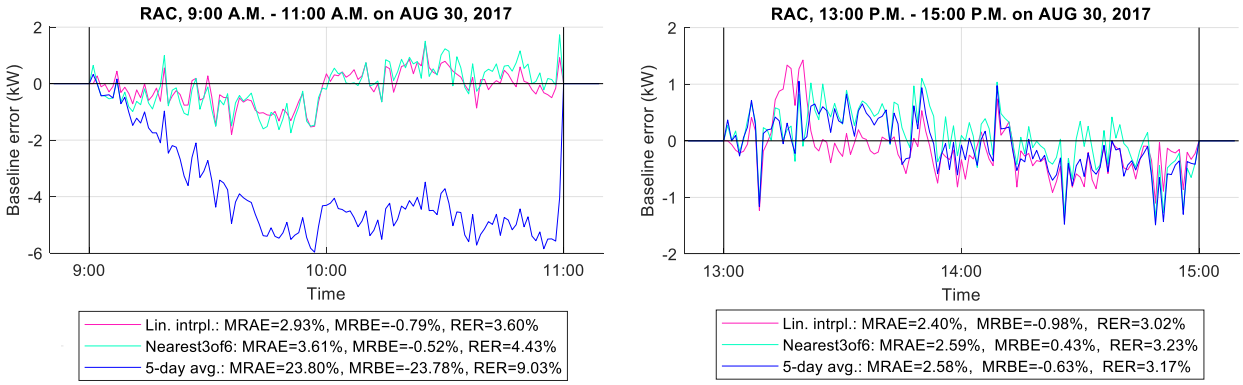

Fig. 3. Example time series plots of the errors of the three best baseline methods. (Upper: 9:00 a.m.-11:00 a.m. Lower: 13:00 p.m.-15:00 p.m.)

All time series plots are shown in Appendix A.

4.2. Fan Power and Outdoor Air Temperature Correlations

Figures 4-8 show scatter plots of the total fan power against the outdoor air temperature for the five building-years. The data within a 4-hour morning time window and a 4-hour afternoon time window are plotted separately, to attain the correlations between the total fan power and outdoor temperature in these two time windows. Values of the Pearson correlation coefficient are also marked in the titles of scatter plots.

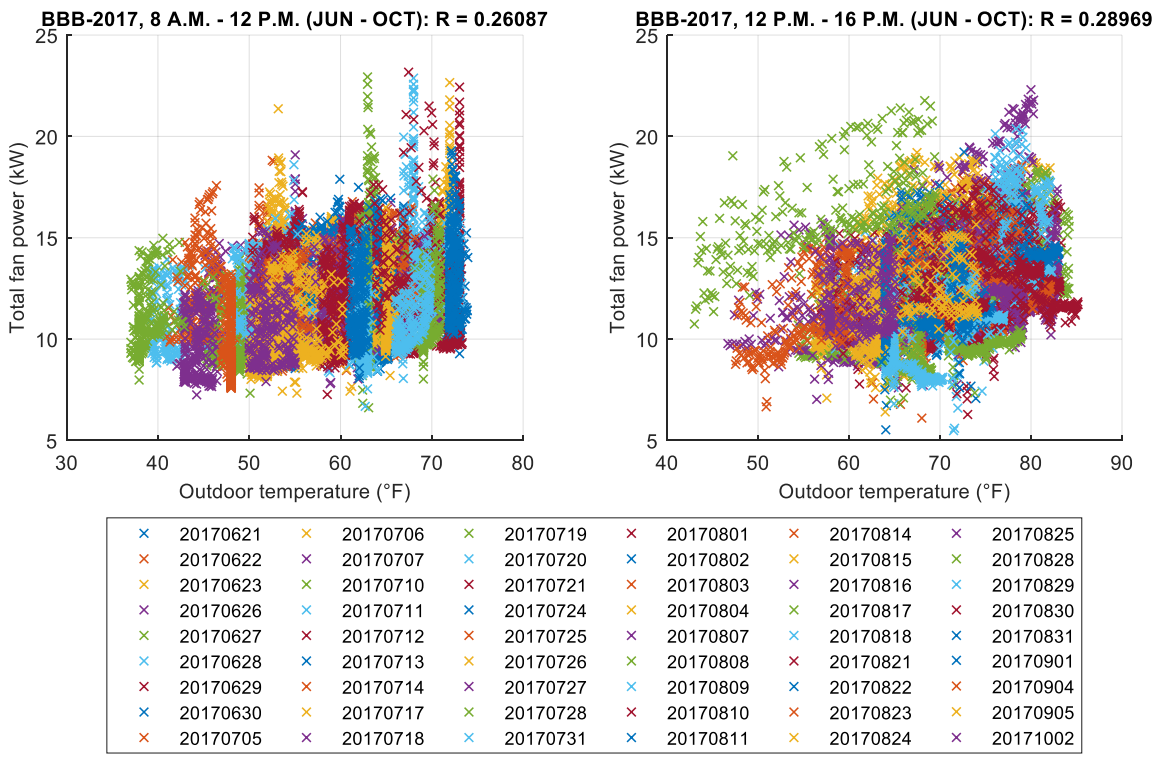

Fig. 4. Scatter plots of the total fan power from BBB-2017 against the outdoor air temperature. (Left: 8:00 a.m.-12:00 p.m. Right: 12:00 p.m.-16:00 p.m.)

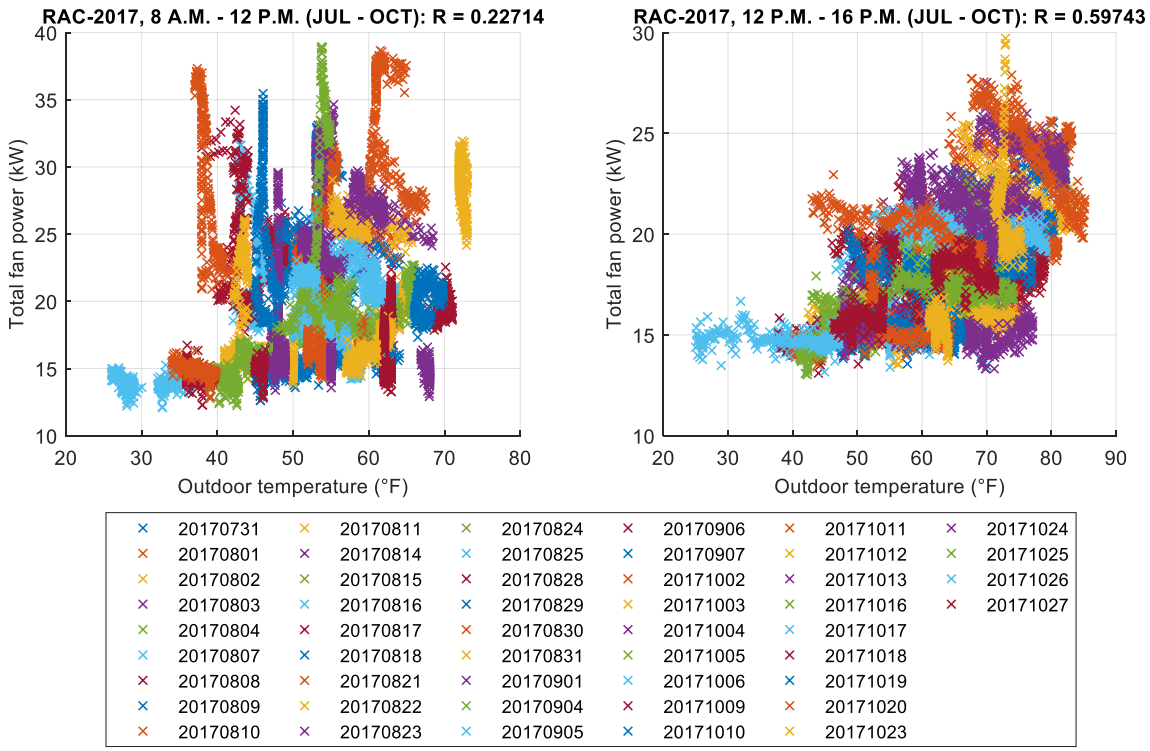

Fig. 5. Scatter plots of the total fan power from RAC-2017 against the outdoor air temperature. (Left: 8:00 a.m.-12:00 p.m. Right: 12:00 p.m.-16:00 p.m.)

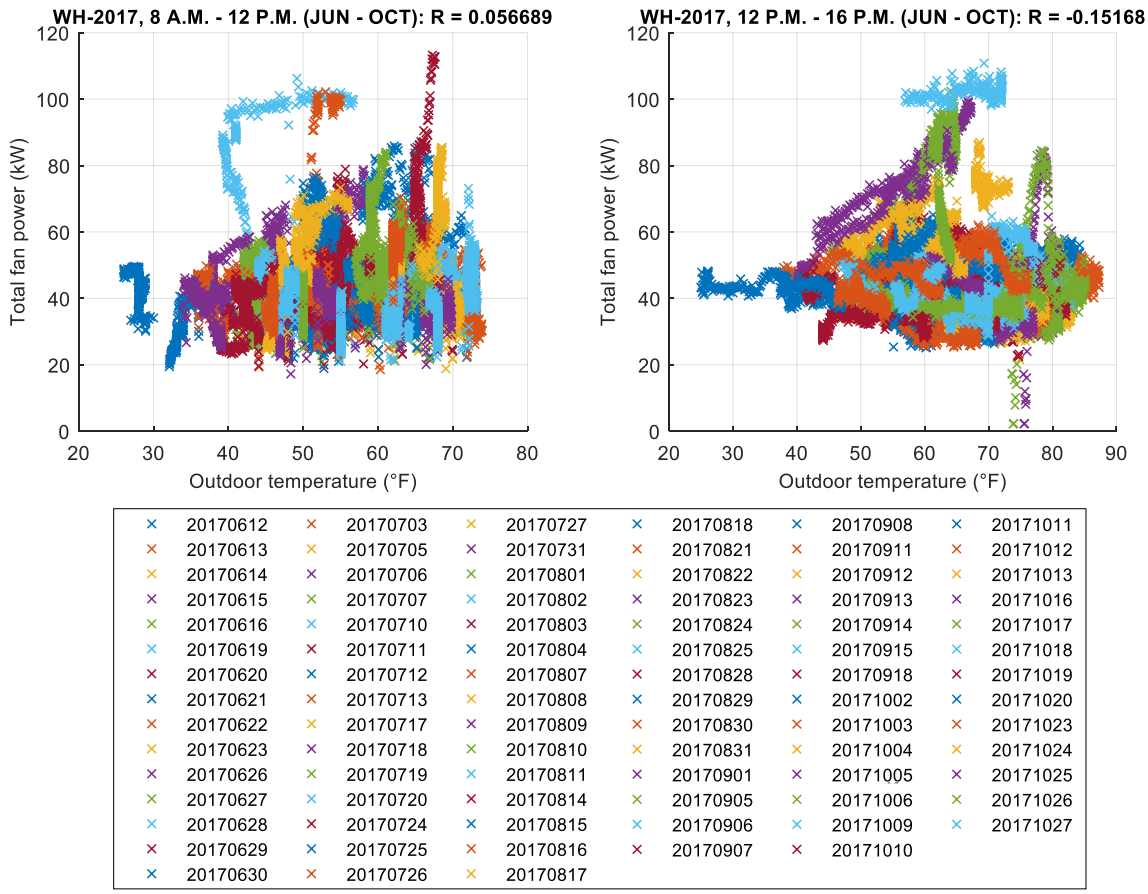

Fig. 6. Scatter plots of the total fan power from WH-2017 against the outdoor air temperature. (Left: 8:00 a.m.-12:00 p.m. Right: 12:00 p.m.-16:00 p.m.)

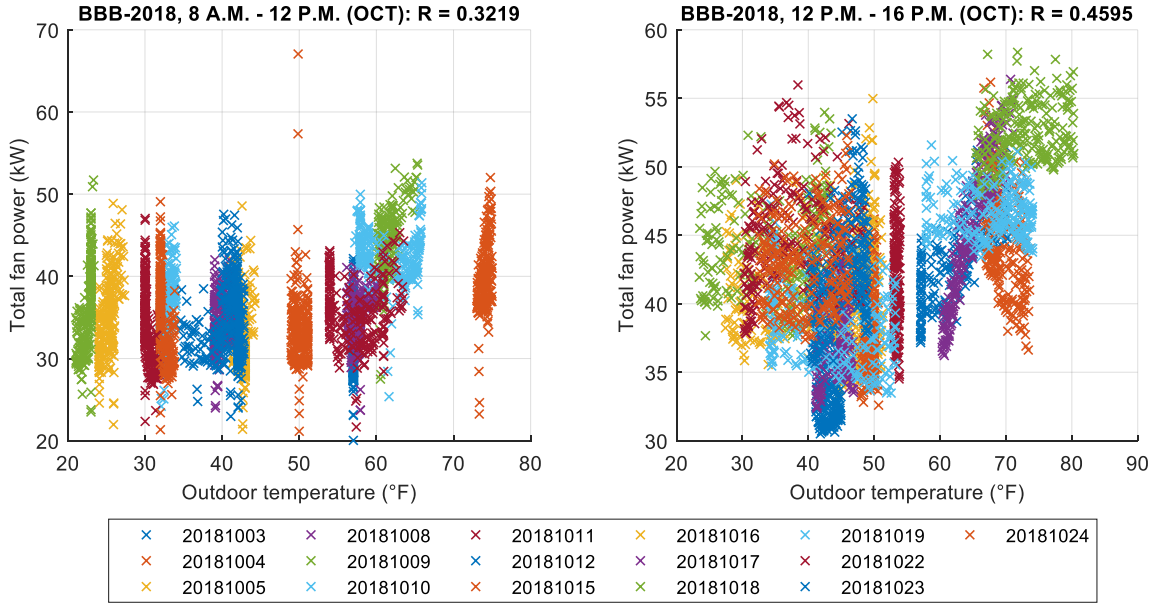

Fig. 7. Scatter plots of the total fan power from BBB-2018 against the outdoor air temperature. (Left: 8:00 a.m.-12:00 p.m. Right: 12:00 p.m.-16:00 p.m.)

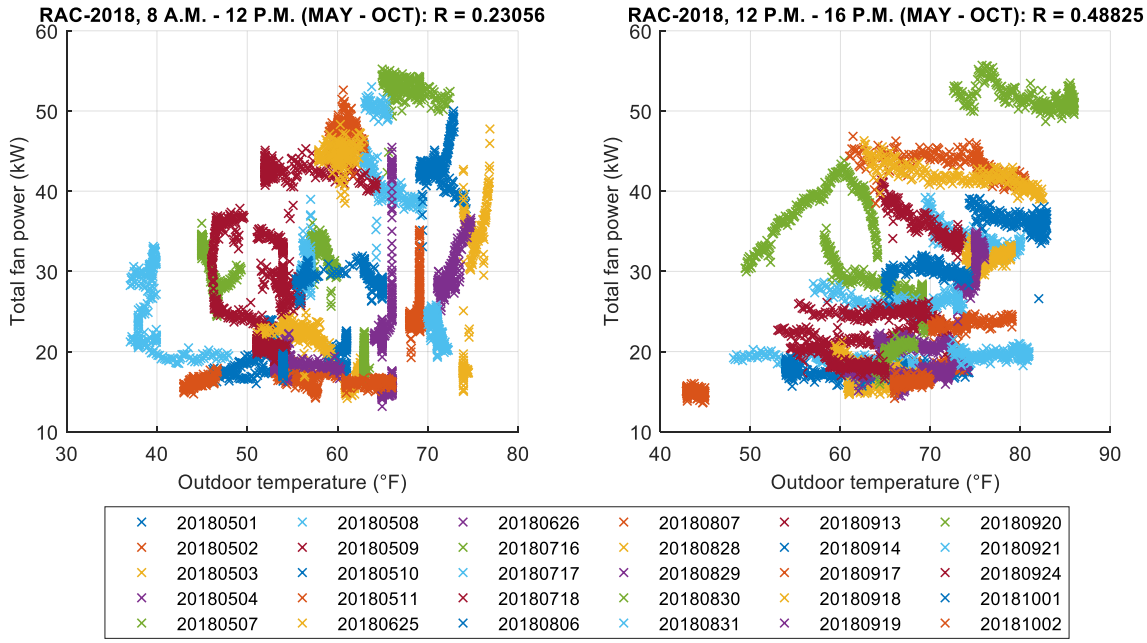

Fig. 8. Scatter plots of the total fan power from RAC-2018 against the outdoor air temperature. (Left: 8:00 a.m.-12:00 p.m. Right: 12:00 p.m.-16:00 p.m.)

As shown in the figures, the correlation between total fan power and outdoor air temperature is generally low. Thus, baseline methods regressing the total fan power against the outdoor air temperature would not perform well. In an extreme case, the fan power from WH-2017 and outdoor air temperature have a negative correlation between 12:00 p.m. and 16:00 p.m. The reason for this negative correlation might be the high volatility of WH fan power data.

The figures show that the total fan power and outdoor air temperature generally have a higher correlation in the afternoon. One possible reason is that the operation of the HVAC fans is more sensitive to outdoor temperature when the outdoor temperature is within a higher range. That is, the HVAC fans might have different operational patterns corresponding to different ranges of the outdoor temperature. This motivates the use of a change-point linear regression model here, which is widely used in correlating building energy consumption with outdoor temperature [29]. However, the scatter plots do not show any apparent change points. Another possible reason is that, the operation of HVAC system in the morning is driven by achieving a thermal balance in the building, while in the afternoon it is driven by maintaining the thermal balance. In this regard, the HVAC fan power could be more sensitive to outdoor temperature in the afternoon, as it is a major factor impacting the external gains of the building.

4.3. Evaluation Metrics and Visualization via Boxplots

Next, we report the average value and standard deviation of the evaluation metrics in Tables 1-6. (The values in the brackets are standard deviations.) Note that those results correspond to the baseline methods without adjustments. The evaluation is conducted and reported for the morning time window and afternoon time window separately. In Table 3 and Table 4, we also provide the 95% confidence interval of the MRBE. In Figures 9-13, we provide five figures showing box plots enabling visualization of the evaluation metrics.

Table 1. Average Value and Standard Deviation of MRAE (%) for each Baseline Method (9:00 A.M. – 11:00 A.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	15.02 (4.37)	17.98 (16.52)	18.52 (11.35)	10.35 (2.99)	34.53 (24.64)
10-day average	14.37 (3.47)	21.16 (13.80)	17.31 (10.36)	8.95 (1.20)	35.50 (22.23)
High4of5	15.76 (4.67)	20.07 (19.16)	20.54 (12.90)	11.40 (4.08)	39.20 (30.36)
High5of10	17.20 (4.92)	31.15 (22.80)	24.04 (16.09)	12.72 (3.12)	51.37 (33.43)
Mid4of6	14.75 (4.01)	17.54 (15.44)	17.18 (10.64)	9.59 (2.11)	36.12 (20.23)
Low4of5	14.60 (3.96)	15.97 (13.95)	16.15 (10.20)	9.74 (2.91)	31.38 (21.43)
Low5of10	13.42 (2.32)	13.70 (11.53)	15.34 (10.06)	8.42 (1.06)	31.23 (18.16)
Nearest3of6	14.67 (2.91)	14.03 (11.94)	15.08 (9.70)	8.44 (1.12)	20.86 (24.79)
Nearest5of10	13.69 (2.49)	12.51 (8.57)	14.40 (9.29)	8.09 (0.81)	24.21 (21.51)
Linear interpolation	12.65 (2.03)	4.86 (4.11)	10.42 (6.33)	7.74 (1.41)	5.68 (11.44)

Table 2. Average Value and Standard Deviation of MRAE (%) for each Baseline Method (13:00 P.M. – 15:00 P.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	14.48 (6.31)	8.71 (6.34)	22.71 (18.16)	12.18 (8.70)	30.01 (20.54)
10-day average	13.79 (5.83)	8.95 (5.90)	23.13 (16.10)	9.45 (3.27)	31.17 (19.68)
High4of5	15.08 (7.11)	9.32 (6.88)	24.74 (21.69)	13.76 (10.51)	34.75 (24.86)
High5of10	17.06 (7.66)	12.25 (7.48)	32.19 (26.92)	13.06 (7.32)	44.98 (31.07)
Mid4of6	14.56 (6.24)	8.46 (5.96)	22.88 (16.62)	13.13 (8.12)	29.89 (17.16)
Low4of5	14.52 (5.80)	8.16 (5.96)	21.16 (15.89)	11.39 (6.90)	24.97 (18.36)
Low5of10	14.15 (5.98)	7.59 (4.91)	18.71 (14.65)	11.36 (3.85)	29.71 (17.48)
Nearest3of6	12.86 (4.53)	8.04 (5.20)	17.04 (15.63)	11.08 (6.04)	13.33 (13.50)
Nearest5of10	12.56 (4.98)	7.66 (4.89)	16.25 (12.27)	9.08 (2.79)	17.02 (12.04)
Linear interpolation	9.09 (2.11)	2.99 (1.11)	8.01 (8.58)	5.96 (1.55)	2.67 (1.34)

Table 3. Average Value ± Confidence Interval and Standard Deviation of MRBE (%) for each Baseline Method (9:00 A.M. – 11:00 A.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	3.46±2.85 (10.30)	7.66±6.82 (23.09)	4.74±4.51 (20.70)	3.28±5.09 (8.61)	13.52±15.93 (40.63)
10-day average	4.86±2.41 (8.25)	11.27±7.05 (22.46)	4.35±4.32 (19.21)	3.90±2.82 (3.52)	12.74±17.64 (40.25)
High4of5	4.95±3.01 (10.86)	11.03±7.50 (25.37)	8.41±4.86 (22.30)	5.16±5.64 (9.54)	21.68±17.64 (45.00)
High5of10	10.72±2.64 (9.04)	24.15±9.43 (30.04)	16.07±5.33 (23.70)	10.63±3.48 (4.34)	38.00±21.35 (48.72)
Mid4of6	2.12±2.74 (9.79)	6.96±6.64 (22.23)	1.17±4.29 (19.58)	3.58±3.83 (6.18)	13.30±15.89 (39.73)
Low4of5	1.03±2.70 (9.76)	3.53±6.11 (20.69)	-1.94±3.99 (18.31)	0.41±4.77 (8.08)	5.26±14.90 (38.02)
Low5of10	-1.00±2.24 (7.67)	-1.60±5.52 (17.59)	-7.37±3.61 (16.04)	-2.82±2.58 (3.23)	-12.51±14.89 (33.98)
Nearest3of6	2.17±2.38 (8.49)	3.80±5.29 (17.68)	0.63±3.72 (16.95)	1.28±2.25 (3.63)	5.76±12.83 (32.07)
Nearest5of10	2.75±2.01 (6.87)	2.71±4.50 (14.34)	0.57±3.67 (16.31)	-0.31±2.31 (2.88)	7.71±13.93 (31.79)
Linear interpolation	3.34±1.18 (4.45)	-0.98±1.58 (5.64)	-3.59±2.11 (9.98)	3.14±1.46 (2.99)	2.22±4.35 (12.15)

Table 4. Average Value ± Confidence Interval and Standard Deviation of MRBE (%) for each Baseline Method (13:00 P.M. – 15:00 P.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	1.41±3.77 (13.60)	3.57±2.99 (10.11)	6.28±5.67 (26.03)	3.90±8.26 (13.97)	13.34±13.43 (34.25)
10-day average	3.05±3.59 (12.28)	4.88±2.96 (9.44)	5.53±5.66 (25.17)	1.33±7.34 (9.17)	8.92±15.96 (36.41)
High4of5	3.72±3.91 (14.11)	4.91±3.08 (10.44)	10.63±6.33 (29.07)	7.42±9.02 (15.27)	20.94±14.77 (37.67)
High5of10	11.00±3.78 (12.95)	9.63±3.31 (10.55)	19.93±7.91 (35.18)	10.37±7.75 (9.69)	33.92±19.02 (43.39)
Mid4of6	0.82±3.79 (13.54)	2.42±2.99 (10.01)	5.70±5.47 (24.96)	4.97±8.77 (14.16)	6.79±13.73 (34.33)
Low4of5	-1.40±3.68 (13.29)	1.23±2.95 (9.97)	1.59±5.10 (23.40)	0.51±7.49 (12.68)	0.80±12.30 (31.38)
Low5of10	-4.90±3.58 (12.24)	0.12±2.80 (8.93)	-8.87±4.12 (18.34)	-7.70±6.99 (8.73)	-16.08±13.59 (31.02)
Nearest3of6	0.69±2.74 (9.79)	2.29±2.75 (9.20)	2.03±4.52 (20.64)	1.68±7.10 (11.45)	1.84±7.60 (19.00)
Nearest5of10	1.98±2.82 (9.64)	2.55±2.70 (8.60)	0.06±3.79 (16.85)	-1.65±6.23 (7.79)	0.97±9.25 (21.11)
Linear interpolation	-1.96±1.05 (3.98)	-0.03±0.54 (1.91)	-0.99±1.03 (4.85)	0.35±1.09 (2.23)	-0.14±0.93 (2.60)

Table 5. Average Value and Standard Deviation of RER (%) for each Baseline Method (9:00 A.M. – 11:00 A.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	14.99 (1.47)	7.08 (4.90)	9.14 (3.52)	8.77 (1.59)	7.44 (6.19)
10-day average	14.36 (1.47)	7.01 (4.34)	8.75 (3.30)	9.71 (0.93)	6.34 (6.62)
High4of5	15.35 (1.43)	7.67 (5.33)	9.46 (3.59)	8.92 (1.64)	8.03 (6.32)
High5of10	15.07 (1.46)	8.54 (4.82)	9.44 (3.51)	9.72 (0.93)	7.70 (6.49)
Mid4of6	15.31 (1.43)	7.10 (4.52)	9.34 (3.43)	9.57 (1.10)	7.28 (6.35)
Low4of5	15.33 (1.51)	7.00 (5.09)	9.24 (3.45)	9.06 (1.68)	7.13 (6.19)
Low5of10	14.99 (1.48)	6.77 (4.64)	9.09 (3.25)	10.38 (1.06)	6.22 (6.65)
Nearest3of6	15.88 (1.60)	6.80 (4.74)	9.39 (3.70)	9.95 (0.98)	6.31 (6.20)
Nearest5of10	15.02 (1.43)	6.62 (4.69)	8.91 (3.37)	10.09 (1.08)	6.40 (6.57)
Linear interpolation	13.72 (1.53)	4.23 (2.46)	8.99 (2.42)	8.99 (1.25)	4.91 (5.82)

Table 6. Average Value and Standard Deviation of RER (%) for each Baseline Method (13:00 P.M. – 15:00 P.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	12.41 (2.01)	3.57 (1.32)	9.69 (5.25)	8.16 (1.36)	3.34 (1.35)
10-day average	12.08 (1.91)	3.54 (1.39)	9.09 (5.05)	8.19 (0.36)	2.87 (1.24)
High4of5	12.63 (2.04)	3.65 (1.33)	10.14 (5.72)	8.34 (1.34)	3.58 (1.52)
High5of10	12.63 (1.94)	3.88 (1.48)	10.66 (6.54)	8.59 (0.44)	3.28 (1.28)
Mid4of6	12.62 (1.95)	3.63 (1.36)	10.40 (5.12)	8.81 (0.61)	3.27 (1.26)
Low4of5	12.65 (1.98)	3.62 (1.31)	9.96 (4.89)	8.34 (1.24)	3.12 (1.36)
Low5of10	12.52 (1.94)	3.58 (1.36)	8.78 (4.70)	8.34 (0.30)	2.81 (1.25)
Nearest3of6	12.99 (2.00)	3.75 (1.33)	9.84 (5.02)	9.11 (0.73)	3.21 (1.22)
Nearest5of10	12.49 (1.89)	3.59 (1.39)	9.24 (4.89)	8.60 (0.53)	3.08 (1.39)
Linear interpolation	11.36 (1.87)	3.53 (0.88)	7.57 (4.59)	6.91 (1.73)	2.38 (0.72)

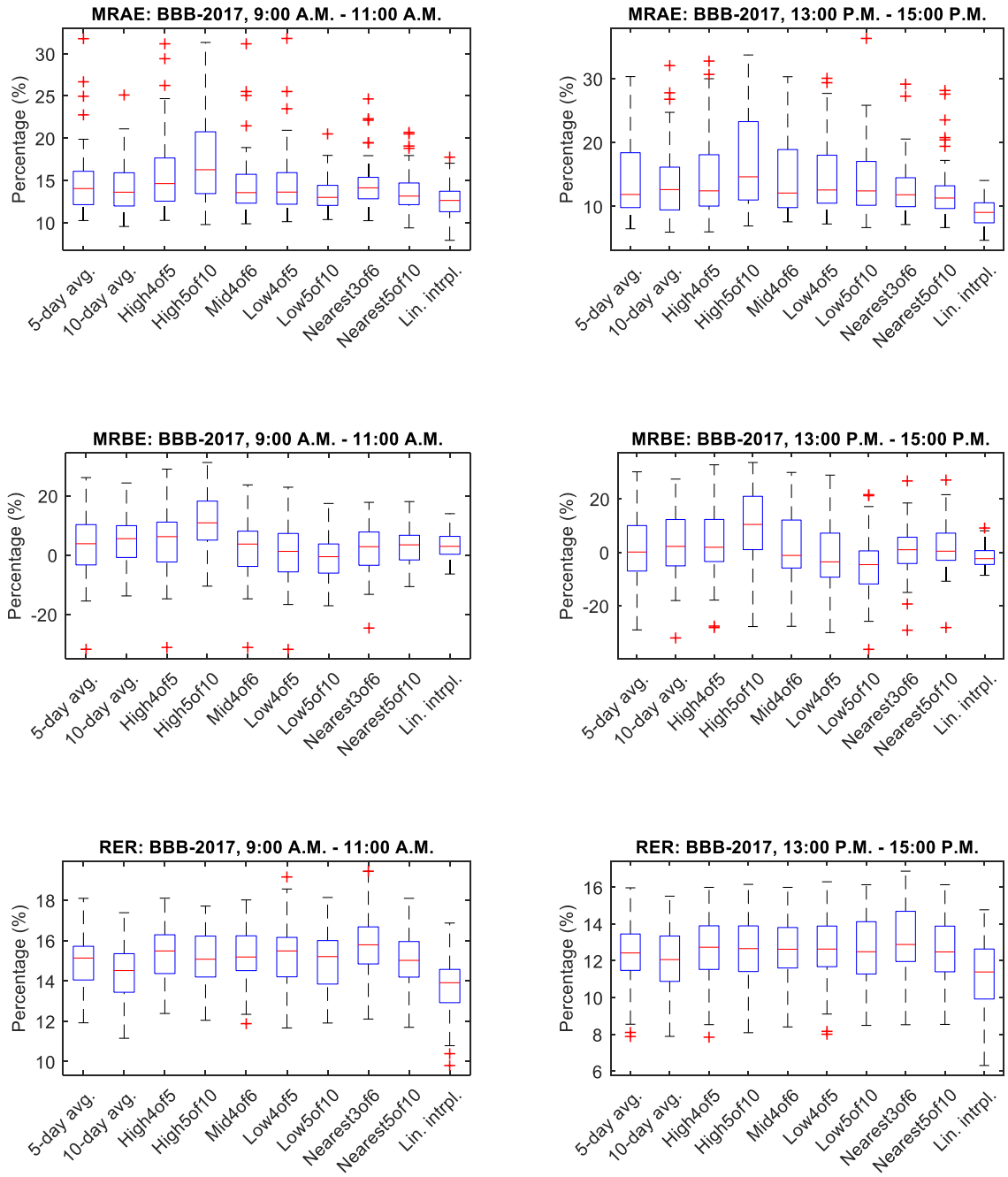

Fig. 9. Boxplots of the evaluation metrics for BBB-2017.

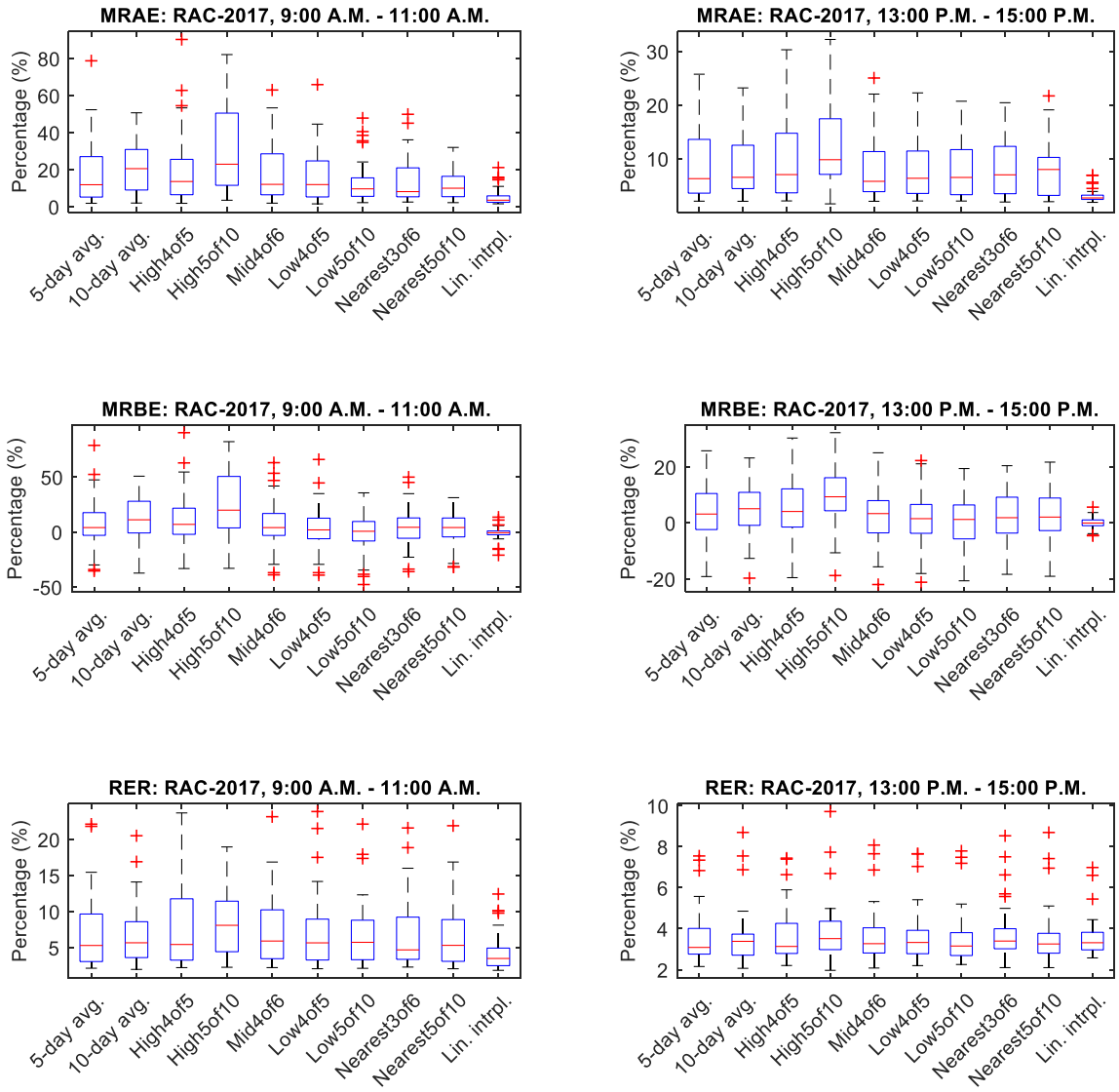

Fig. 10. Boxplots of the evaluation metrics for RAC-2017.

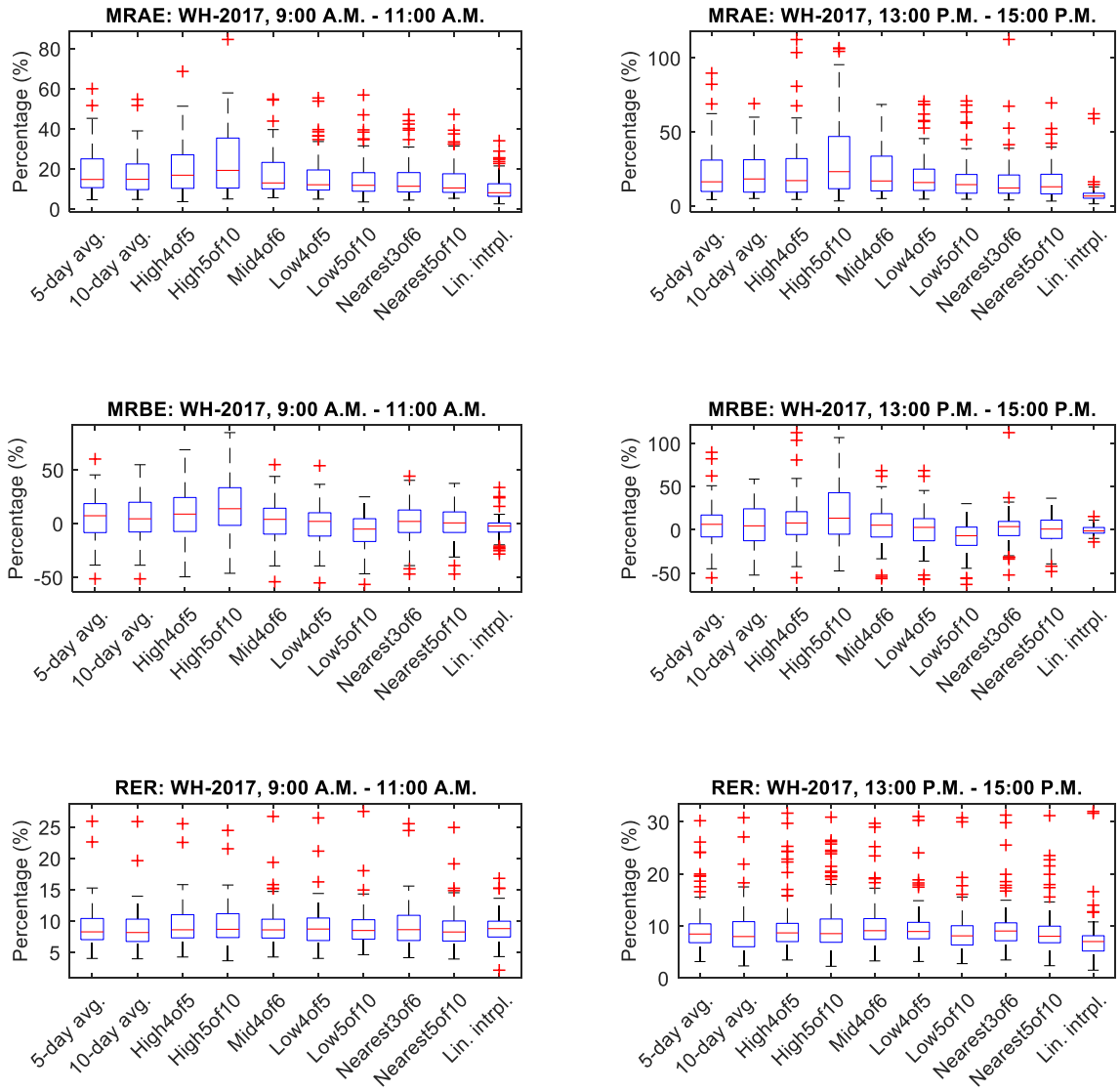

Fig. 11. Boxplots of the evaluation metrics for WH-2017.

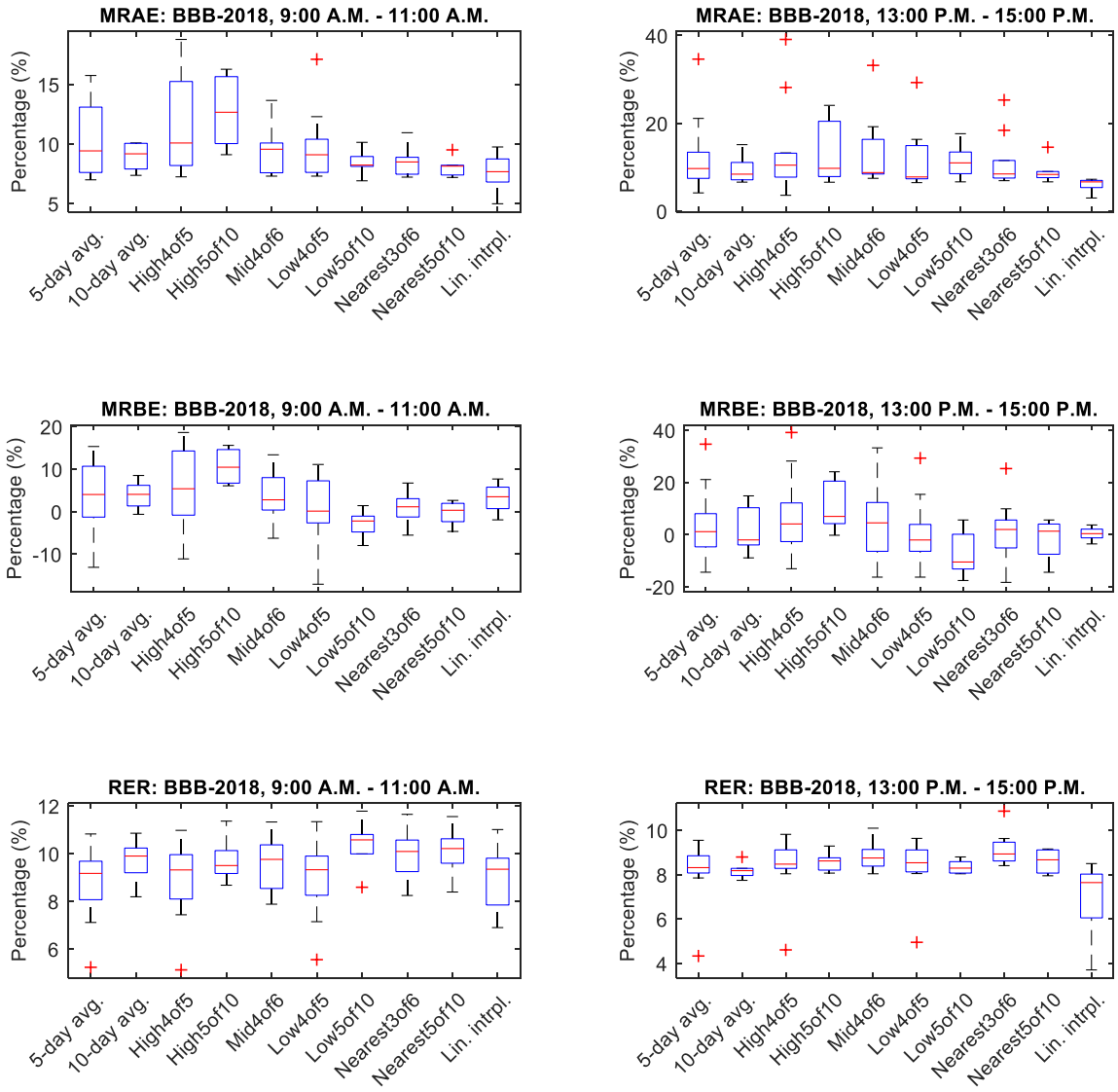

Fig. 12. Boxplots of the evaluation metrics for BBB-2018.

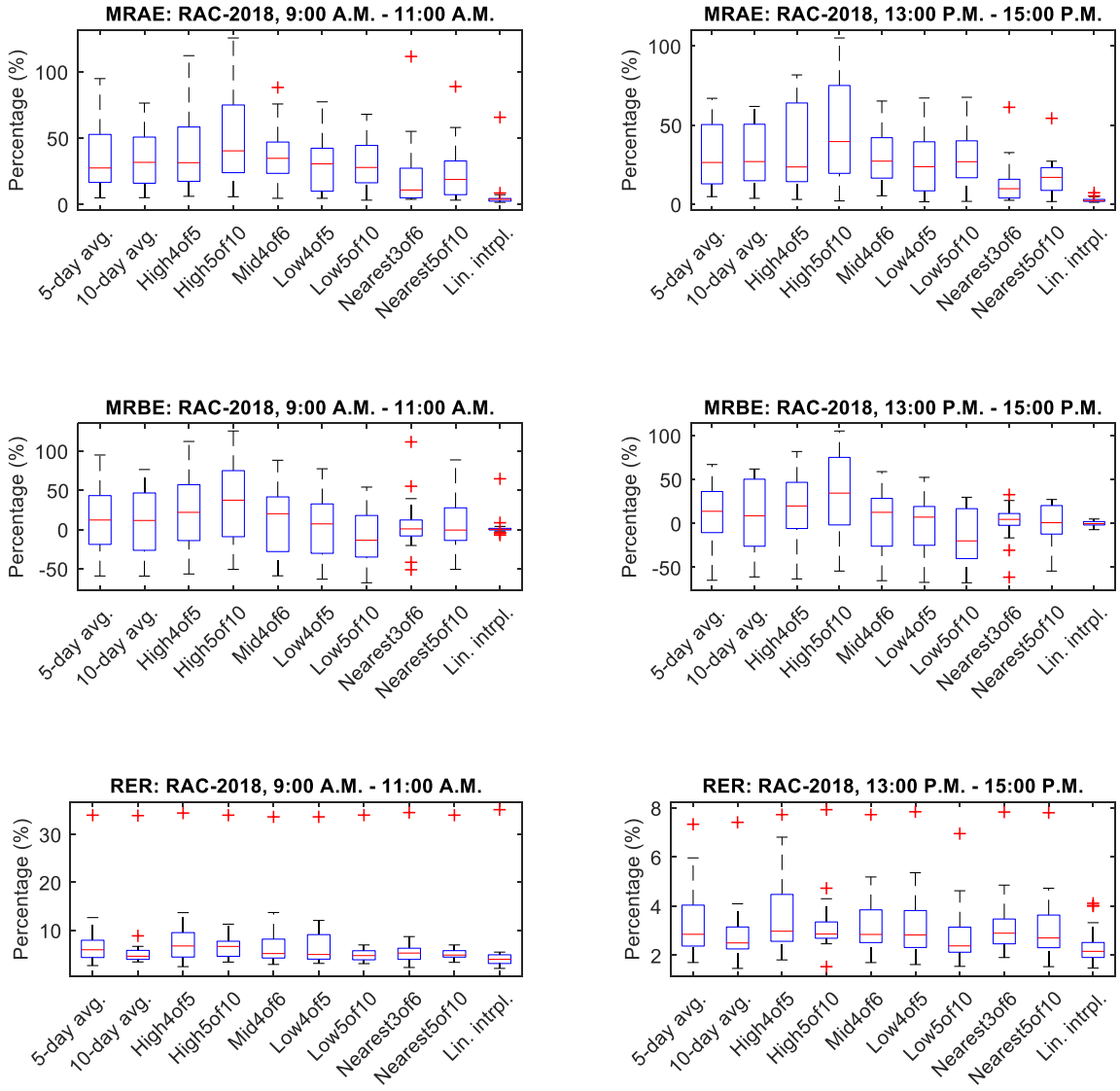

Fig. 13. Boxplots of the evaluation metrics for RAC-2018.

As shown in the tables and figures, the simple linear interpolation method generally works the best; on average, it has the lowest values and standard deviations of MRAE, MRBE, and RER. All methods have high MRAE values (though the linear interpolation method attains values of MRAE less than 3% in two cases) and so new baseline methods are required if a low MRAE is desired. However, in general, bias is more relevant than accuracy for analyzing building energy efficiency and determining the compensation given to demand response participants, so MRBE is a more important metric. As shown in Table 3 and Table 4, several methods have very small MRBE values, and even beat the linear interpolation method, in some cases. However, on average, the linear interpolation method is still the best. As shown in Table 4, the linear interpolation method has especially low values of MRBE when evaluated on afternoon demand response events. The reason may be that the total fan power profile is more stable in the afternoon, while it is more volatile in the morning.

All baseline methods have large values of RER in most cases meaning that error results are volatile across different time slots during the event time window. The linear interpolation method is the most stable.

The box plots show that the baseline methods have variable performance across different days. Again, the linear interpolation method is less variable and, in some cases, it is fairly stable across different days.

4.4. Impact of the Additive Adjustment

To make the results more comparable, we further apply the additive adjustment mentioned to the baseline methods. That is, the baseline estimated by the above averaging methods is vertically shifted, so that the baseline load at the start of the time window of the demand response event is equal to the measured load. Tables 7-12 and Figures 14-18 show the results.

Table 7. Average Value and Standard Deviation of MRAE (%) for each Baseline Method with an Additive Adjustment (9:00 A.M. – 11:00 A.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	17.26 (5.88)	11.33 (11.15)	15.08 (13.57)	7.90 (1.94)	10.62 (11.35)
10-day average	15.58 (5.01)	11.04 (10.26)	13.69 (12.45)	8.99 (2.37)	8.85 (11.78)
High4of5	17.90 (5.86)	12.41 (12.17)	15.85 (14.22)	8.23 (1.84)	11.57 (11.65)
High5of10	17.21 (4.82)	14.83 (11.96)	14.98 (12.28)	9.15 (3.07)	9.27 (11.83)
Mid4of6	17.45 (6.09)	11.36 (10.91)	14.50 (13.86)	8.91 (1.65)	10.34 (10.99)
Low4of5	17.68 (6.55)	11.05 (11.22)	14.58 (13.65)	8.24 (2.06)	10.27 (10.81)
Low5of10	15.93 (5.81)	10.00 (9.85)	13.69 (13.73)	9.54 (1.96)	9.41 (11.78)
Nearest3of6	18.53 (6.28)	9.77 (9.71)	14.88 (14.11)	9.02 (1.93)	9.58 (11.36)
Nearest5of10	16.22 (5.56)	9.92 (10.56)	13.36 (13.05)	10.14 (3.49)	9.45 (11.75)
Linear interpolation	12.65 (2.03)	4.86 (4.11)	10.42 (6.33)	7.74 (1.41)	5.68 (11.44)

Table 8. Average Value and Standard Deviation of MRAE (%) for each Baseline Method with an Additive Adjustment (13:00 P.M. – 15:00 P.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	14.03 (5.98)	4.55 (2.42)	14.31 (13.72)	8.86 (3.08)	4.66 (2.80)
10-day average	13.33 (5.64)	4.42 (2.43)	13.71 (12.35)	10.05 (4.67)	4.20 (2.40)
High4of5	14.05 (5.75)	4.63 (2.51)	15.37 (15.07)	8.82 (2.96)	4.89 (2.77)
High5of10	13.87 (6.04)	4.98 (2.68)	17.27 (15.88)	9.96 (4.25)	5.00 (2.36)
Mid4of6	14.26 (6.14)	4.55 (2.25)	15.89 (13.61)	9.81 (3.17)	4.47 (2.52)
Low4of5	14.67 (6.38)	4.52 (2.31)	14.64 (12.94)	9.26 (3.51)	4.24 (2.73)
Low5of10	14.24 (5.47)	4.51 (2.38)	12.16 (12.14)	10.53 (5.16)	3.75 (2.67)
Nearest3of6	15.04 (6.61)	4.78 (2.03)	14.51 (12.67)	9.93 (3.54)	4.34 (2.71)
Nearest5of10	14.12 (5.75)	4.49 (2.18)	13.40 (12.41)	10.47 (5.17)	4.32 (2.91)
Linear interpolation	9.09 (2.11)	2.99 (1.11)	8.01 (8.58)	5.96 (1.55)	2.67 (1.34)

Table 9. Average Value ± Confidence Interval and Standard Deviation of MRBE (%) for each Baseline Method with an Additive Adjustment (9:00 A.M. – 11:00 A.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	1.44±4.42 (15.95)	0.39±4.56 (15.45)	-0.01±4.32 (19.82)	-1.64±2.89 (4.90)	1.56±5.78 (14.75)
10-day average	0.22±4.10 (14.03)	0.65±4.59 (14.64)	-0.63±4.04 (17.97)	-3.70±4.93 (6.16)	0.99±6.09 (13.89)
High4of5	0.12±4.61 (16.63)	-0.17±5.01 (16.96)	-0.56±4.54 (20.83)	-2.05±3.13 (5.29)	1.31±6.13 (15.65)
High5of10	-2.87±4.53 (15.49)	1.15±5.89 (18.78)	-2.60±4.18 (18.61)	-3.22±5.72 (7.15)	0.33±6.13 (13.99)
Mid4of6	1.57±4.50 (16.06)	-0.04±4.58 (15.32)	0.61±4.27 (19.46)	-3.96±3.15 (5.09)	0.44±5.72 (14.30)
Low4of5	3.33±4.46 (16.10)	0.68±4.51 (15.26)	1.38±4.22 (19.36)	-3.01±2.80 (4.73)	1.73±5.52 (14.09)
Low5of10	3.31±3.91 (13.37)	0.15±4.22 (13.46)	1.35±4.22 (18.76)	-4.18±4.68 (5.85)	1.65±6.23 (14.22)
Nearest3of6	1.24±4.74 (16.93)	-0.59±3.97 (13.29)	-0.30±4.37 (19.94)	-2.77±3.58 (5.78)	0.99±5.68 (14.19)
Nearest5of10	1.22±4.17 (14.26)	-0.98±4.42 (14.07)	-0.33±4.07 (18.08)	-3.91±6.51 (8.13)	0.71±6.27 (14.31)
Linear interpolation	3.34±1.18 (4.45)	-0.53±1.58 (5.64)	-3.59±2.11 (9.98)	3.14±1.46 (2.99)	2.22±4.35 (12.15)

Table 10. Average Value ± Confidence Interval and Standard Deviation of MRBE (%) for each Baseline Method with an Additive Adjustment (13:00 P.M. – 15:00 P.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	1.22±3.67 (13.25)	-0.33±1.39 (4.71)	2.28±3.71 (17.03)	3.15±4.15 (7.02)	1.07±2.00 (5.11)
10-day average	-0.04±3.69 (12.62)	-0.39±1.43 (4.57)	2.36±3.38 (15.02)	5.31±6.92 (8.65)	0.37±2.06 (4.70)
High4of5	0.76±3.62 (13.07)	-0.73±1.40 (4.75)	3.28±4.11 (18.89)	2.87±4.10 (6.94)	0.89±2.08 (5.32)
High5of10	0.60±3.78 (12.93)	-1.83±1.53 (4.86)	5.90±4.50 (20.01)	4.25±6.94 (8.67)	-0.43±2.38 (5.42)
Mid4of6	-0.81±3.82 (13.65)	-0.56±1.35 (4.53)	3.64±3.88 (17.71)	4.40±4.62 (7.46)	0.28±1.96 (4.90)
Low4of5	0.43±3.92 (14.14)	-0.17±1.35 (4.58)	2.29±3.53 (16.20)	4.40±4.14 (7.00)	0.55±1.87 (4.77)
Low5of10	-0.68±3.92 (13.42)	1.06±1.41 (4.49)	-1.18±3.01 (13.38)	6.38±6.98 (8.73)	1.17±1.84 (4.21)
Nearest3of6	0.82±4.09 (14.61)	-0.08±1.40 (4.68)	0.80±3.68 (16.79)	5.35±4.19 (6.76)	0.47±1.94 (4.85)
Nearest5of10	0.25±3.90 (13.35)	0.43±1.40 (4.46)	0.84±3.41 (15.19)	5.54±7.32 (9.15)	0.68±2.19 (5.01)
Linear interpolation	-1.96±1.05 (3.98)	-0.03±0.54 (1.91)	-0.99±1.03 (4.85)	0.35±1.09 (2.23)	-0.14±0.93 (2.60)

Table 11. Average Value and Standard Deviation of RER (%) for each Baseline Method with an Additive Adjustment (9:00 A.M. – 11:00 A.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	14.99 (1.47)	7.08 (4.90)	9.14 (3.52)	8.77 (1.59)	7.44 (6.19)
10-day average	14.36 (1.47)	7.01 (4.34)	8.75 (3.30)	9.71 (0.93)	6.34 (6.62)
High4of5	15.35 (1.43)	7.67 (5.33)	9.46 (3.59)	8.92 (1.64)	8.03 (6.32)
High5of10	15.07 (1.46)	8.54 (4.82)	9.44 (3.51)	9.72 (0.93)	7.70 (6.49)
Mid4of6	15.31 (1.43)	7.10 (4.52)	9.34 (3.43)	9.57 (1.10)	7.28 (6.35)
Low4of5	15.33 (1.51)	7.00 (5.09)	9.24 (3.45)	9.06 (1.68)	7.13 (6.19)
Low5of10	14.99 (1.48)	6.77 (4.64)	9.09 (3.25)	10.38 (1.06)	6.22 (6.65)
Nearest3of6	15.87 (1.60)	6.80 (4.74)	9.39 (3.70)	9.95 (0.98)	6.31 (6.20)
Nearest5of10	15.02 (1.43)	6.62 (4.69)	8.91 (3.37)	10.09 (1.08)	6.40 (6.57)
Linear interpolation	13.72 (1.53)	4.23 (2.46)	8.99 (2.42)	8.99 (1.25)	4.91 (5.82)

Table 12. Average Value and Standard Deviation of RER (%) for each Baseline Method with an Additive Adjustment (13:00 P.M. – 15:00 P.M.)

Building-year	BBB-2017	RAC-2017	WH-2017	BBB-2018	RAC-2018
5-day average	12.41 (2.01)	3.57 (1.32)	9.69 (5.25)	8.16 (1.36)	3.34 (1.35)
10-day average	12.08 (1.91)	3.54 (1.39)	9.09 (5.05)	8.19 (0.36)	2.87 (1.24)
High4of5	12.63 (2.04)	3.65 (1.33)	10.14 (5.72)	8.34 (1.34)	3.58 (1.52)
High5of10	12.63 (1.94)	3.88 (1.48)	10.66 (6.54)	8.59 (0.44)	3.28 (1.28)
Mid4of6	12.62 (1.95)	3.63 (1.36)	10.40 (5.12)	8.81 (0.61)	3.27 (1.26)
Low4of5	12.65 (1.98)	3.62 (1.31)	9.96 (4.89)	8.34 (1.24)	3.12 (1.36)
Low5of10	12.52 (1.94)	3.58 (1.36)	8.78 (4.70)	8.34 (0.30)	2.81 (1.25)
Nearest3of6	12.99 (2.00)	3.75 (1.33)	9.84 (5.02)	9.11 (0.73)	3.21 (1.22)
Nearest5of10	12.49 (1.89)	3.59 (1.39)	9.24 (4.89)	8.60 (0.53)	3.08 (1.39)
Linear interpolation	11.36 (1.87)	3.52 (0.88)	7.57 (4.59)	6.91 (1.73)	2.38 (0.72)

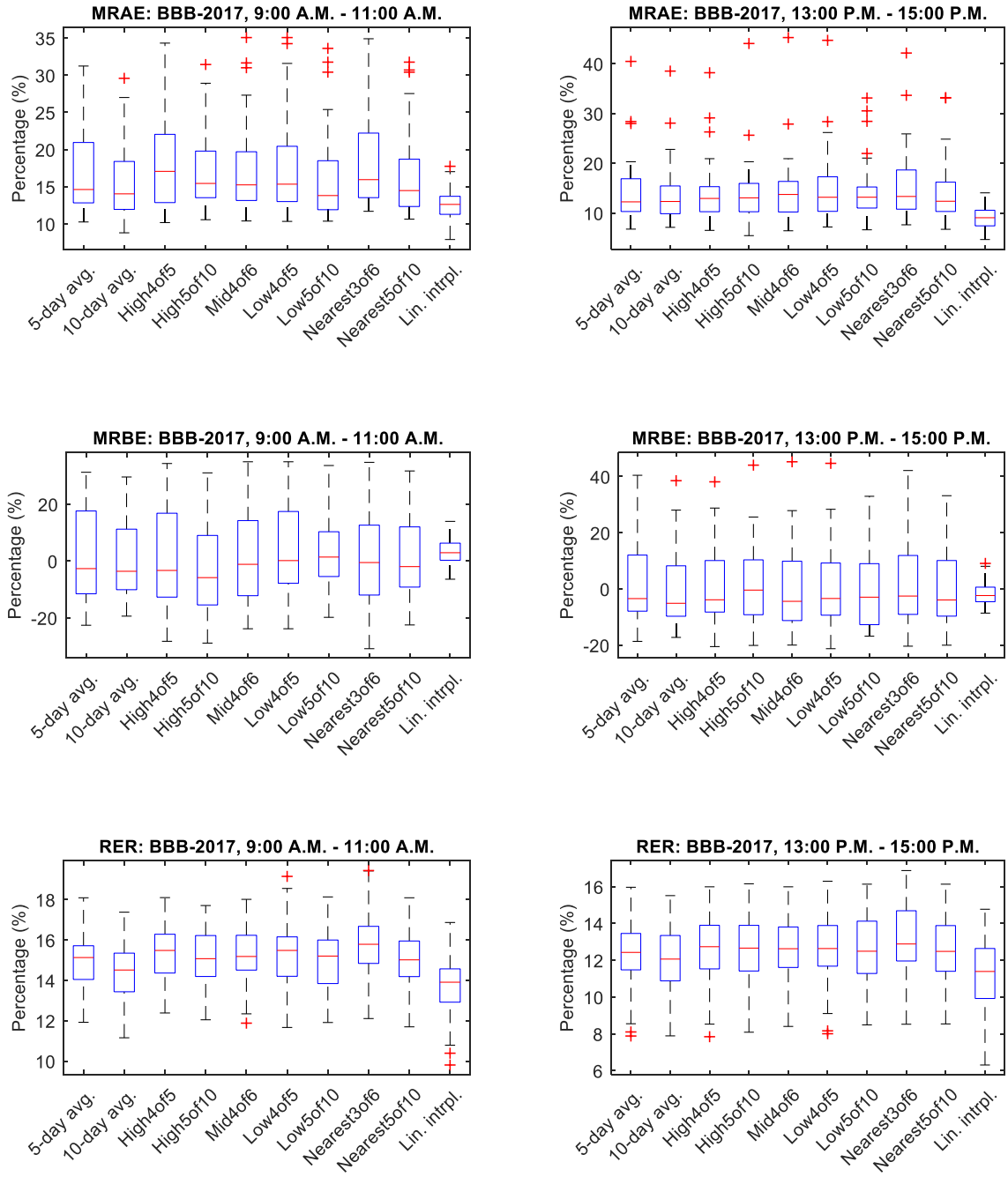

Fig. 14. Boxplots of the evaluation metrics for BBB-2017 (with an additive adjustment for the baseline methods).

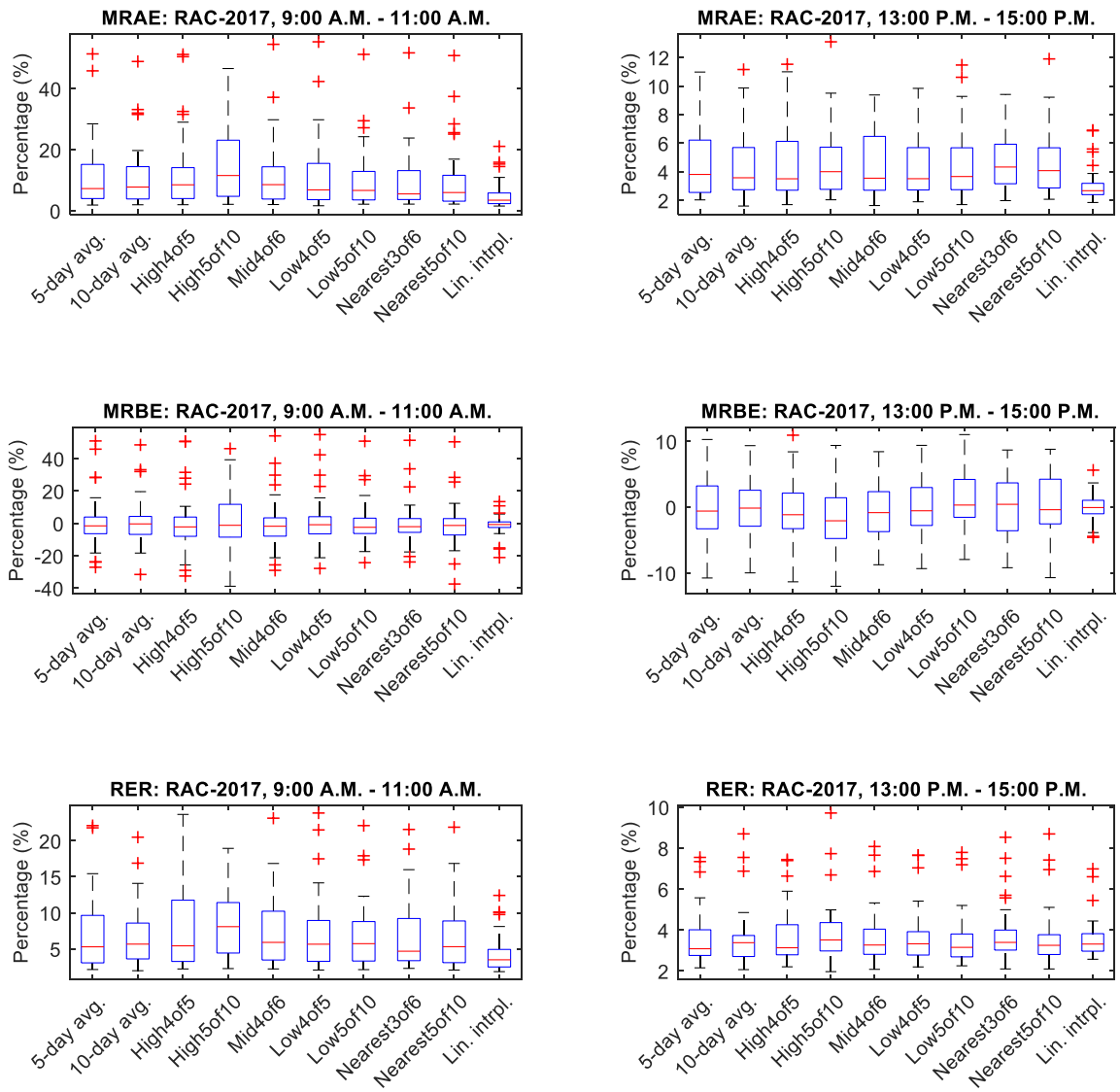

Fig. 15. Boxplots of the evaluation metrics for RAC-2017 (with an additive adjustment for the baseline methods).

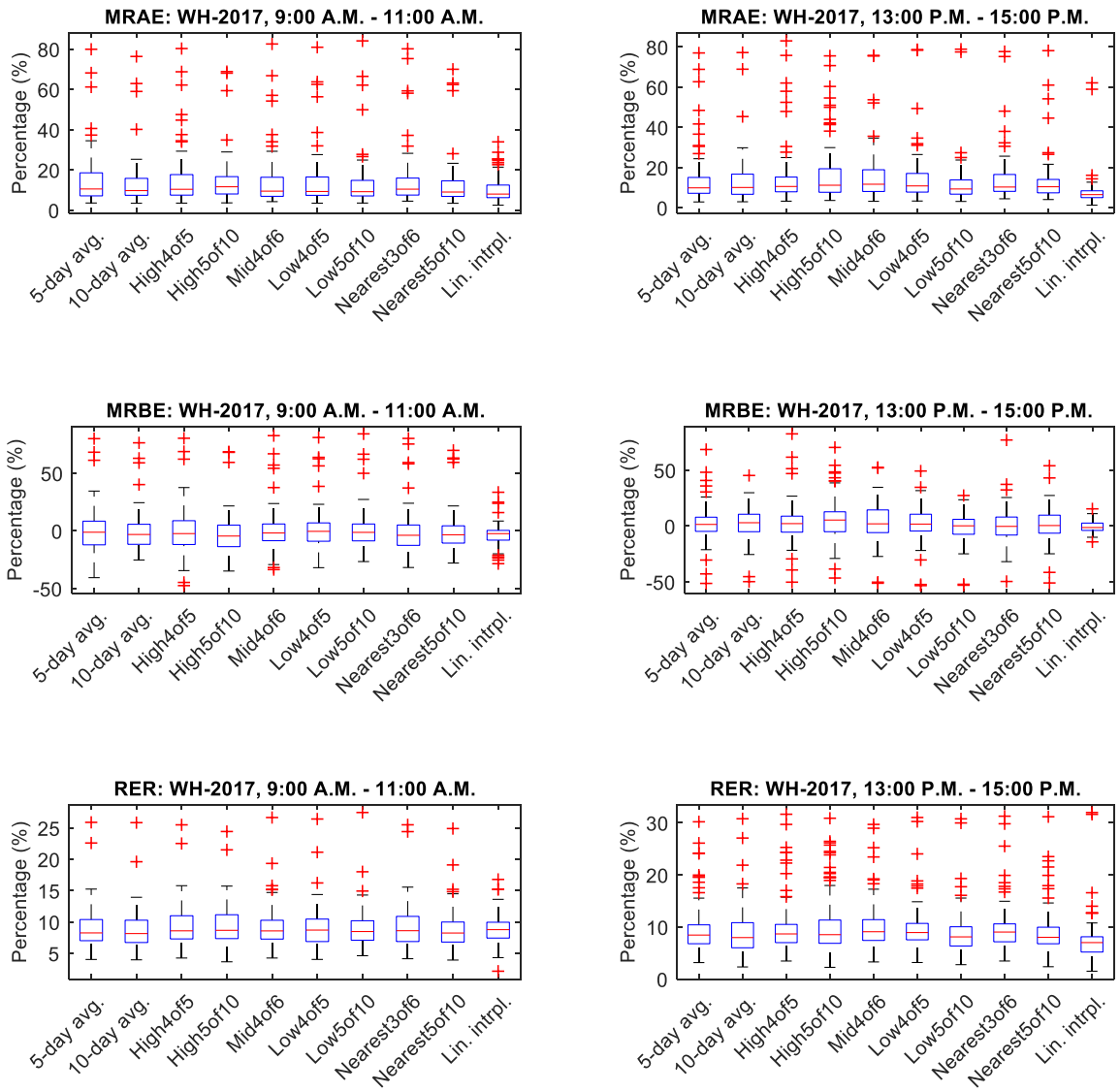

Fig. 16. Boxplots of the evaluation metrics for WH-2017 (with an additive adjustment for the baseline methods).

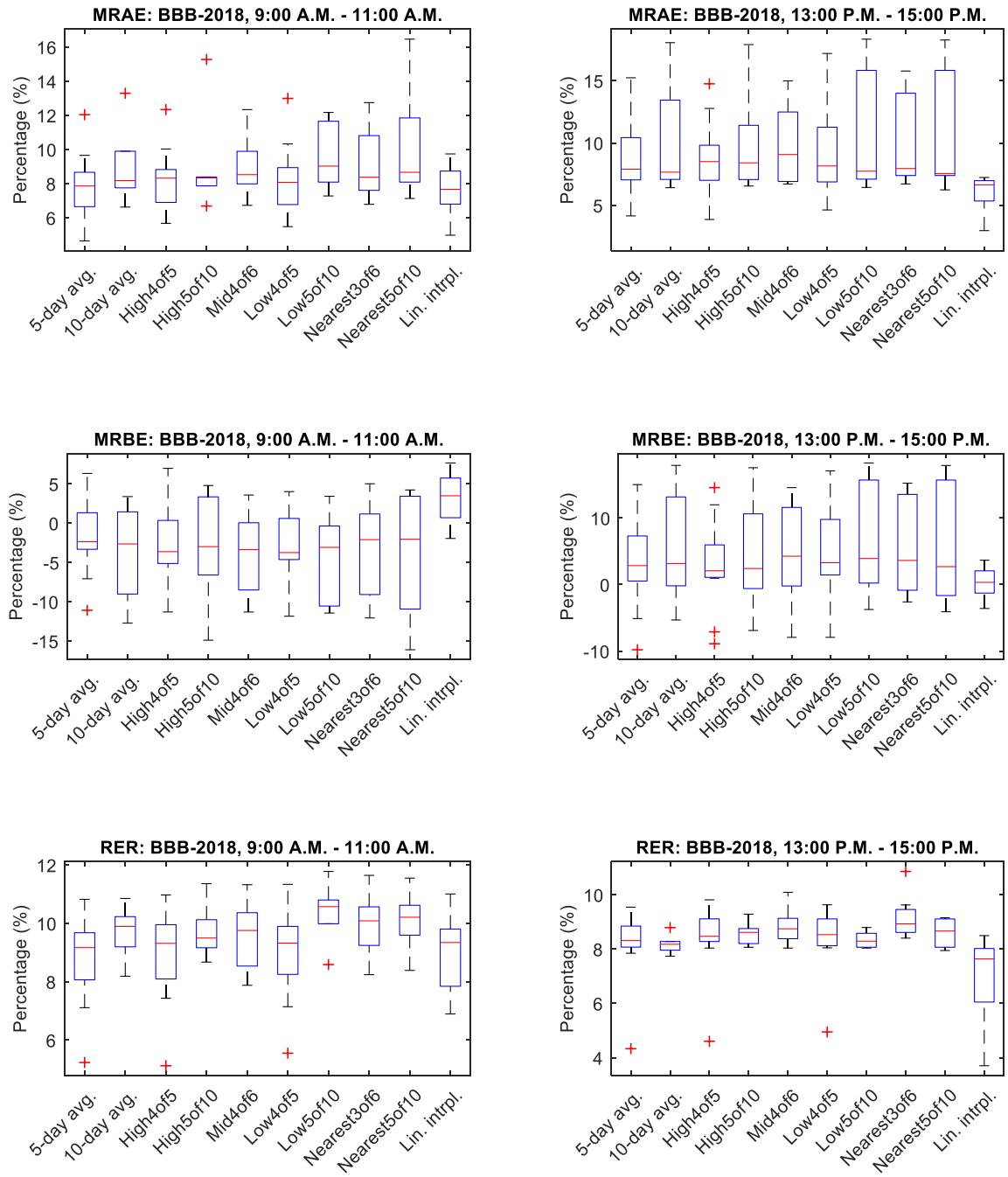

Fig. 17. Boxplots of the evaluation metrics for BBB-2018 (with an additive adjustment for the baseline methods).

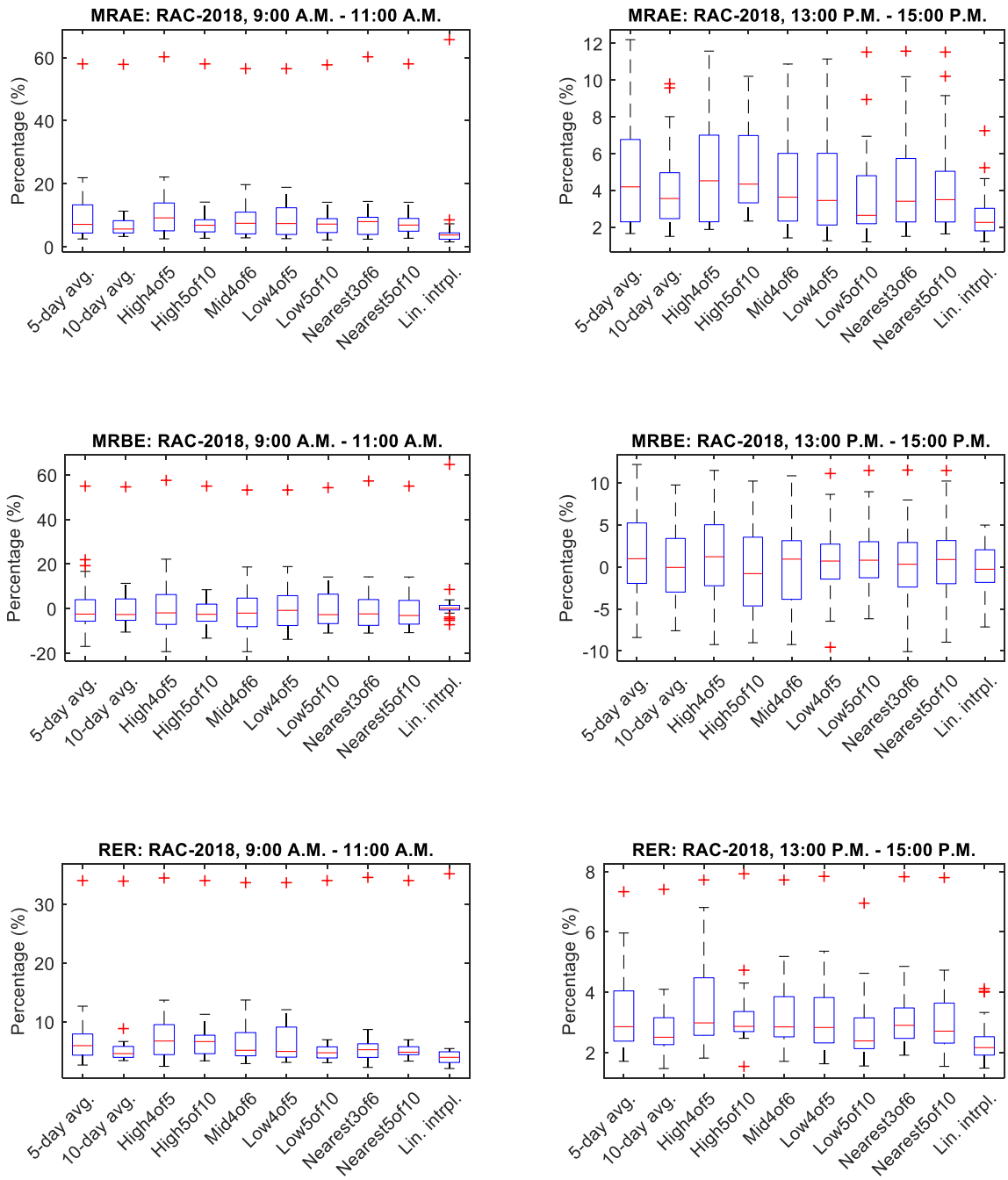

Fig. 18. Box plots of the evaluation metrics for RAC-2018 (with an additive adjustment for the baseline methods).

The additive adjustment greatly improves the performance of averaging baseline methods. Specifically, the average values and standard deviations of MRAE and MRBE of all averaging methods are decreased for all or almost all building-years. In this regard, the performance of averaging methods and the linear interpolation method are closer. For some building-years, some averaging methods attain lower average values of MRBE compared with that of the linear interpolation method, especially for morning demand response events. Still, the linear interpolation method always attains the lowest standard deviation of MRBE, indicating that it is the most stable method in terms of bias variability. Moreover, it has both the smallest average value and the smallest standard deviation of MRBE in most cases. Therefore, it is still the best baseline method we have evaluated. The other observations and conclusions based on the results of baseline methods without adjustments are also valid here.

5.0 Discussion

The baseline method evaluation results indicate that the simple linear interpolation method has the best performance, even compared with averaging methods greatly improved by an additive adjustment. In most cases, the linear interpolation method has the smallest average value and standard deviation of the evaluation metrics, i.e., MRAE, MRBE, and RER that assess the accuracy, bias and variability of the method, respectively. For the analysis of our future demand response experiments, we need a method with lower bias, as the bias of the baseline is the major error source when analyzing the energy efficiency impacts of demand response actions. Thus, we plan to use the linear interpolation method for baselining our future demand response experiment data. As reported in the previous section, the linear interpolation method generally has the lowest average value and the smallest standard deviation of MRBE, especially when it is used to baseline total fan power during afternoon demand response events. In addition to computing results like AEC and RTE using the linear interpolation method for baselining, we will also compute results with the second and third-best methods we have identified: nearest3of6 average method and 5-day simple average method. This will enable us to compare estimates across methods, giving us additional insights into building response and baseline performance.

Load forecasting, especially short-term load forecasting, is a problem highly related to baseline estimation, except that baseline estimation can take advantage of a posteriori knowledge of exogenous variables rather than forecasted variables. In this regard, some load forecasting methods may be modified for baseline estimation. Load forecasting methods can be generally classified as regression methods, time-series methods, and artificial intelligence methods. While we found that regression methods are inappropriate for baseline estimation of total fan power data, time-series methods and artificial intelligence methods are potentially applicable. In [30], the historical hourly loads are processed with a sequential least-squares estimator to identify an autoregressive moving average model for load forecasting. In [31], general exponential smoothing is used to develop an adaptive forecasting system based on observed values of integrated hourly demand. References [32] and [33] construct artificial neural network models for electric load forecasting and building energy use prediction, respectively. In [34], an expert system based algorithm is developed to take advantage of the expertise of system operators for short-term load forecasting. It also may be possible to take advantage of the per-fan power data instead of the total fan power data. It is possible that such more-granular data can be utilized to obtain fan power profiles that are consistent among different fans and over different days. Such patterns can be useful for estimating the fan power baseline. In this regard, we are exploring methods such as tensor decomposition [35, 36], which are capable of high-dimensional data mining and analysis. While the simple linear interpolation method may satisfy our need for baselining our demand response experimental data, it will be beneficial to also investigate these new methods, some of which are effective in load forecasting but have not yet been studied for baseline estimation. One critical research topic is how these methods can be leveraged to improve the current linear interpolation method, especially to enhance its performance in baselining the fan power from morning demand response events.

6.0 Conclusion

In this report, a group of baseline methods were evaluated based on building HVAC fan power data. Our correlation analysis shows that regression baseline methods that regress the fan power against outdoor air temperature are not applicable here. Our numerical results show that averaging baseline methods, especially with an additive adjustment, may work well for baselining the fan power in some cases. However, their performance is not consistent across all cases. A simple linear interpolation method generally has the best performance compared with all other evaluated baseline methods. In particular, it has a small and stable bias. It is selected as the baseline method to be used in our future analysis of demand response experimental data. It may be possible to improve the linear interpolation method using advanced techniques applied in short-term load forecasting. In the future, we will explore the use of these techniques for baselining the fan power from morning time demand response events.

References

[1] "*Assessment of Demand Response and Advanced Metering,*" Federal Energy Regulatory Commission (FERC), Washington, DC, USA, Staff Report, Nov. 2018. [Online]. Available: https://www.ferc.gov/legal/staff-reports/2018/DR-AM-Report2018.pdf

[2] G. Strbac, "Demand Side Management: Benefits and Challenges," *Energy Policy,* vol. 36, no. 12, pp. 4419-4426, Dec. 2008.

[3] J. Aghaei, and M.-I. Alizadeh, "Demand Response in Smart Electricity Grids Equipped with Renewable Energy Sources: A Review," *Renewable and Sustainable Energy Reviews,* vol. 18, pp. 64-72, Feb. 2013.

[4] Y. Zhang, W. Chen, R. Xu, and J. Black, "A Cluster-Based Method for Calculating Baselines for Residential Loads," *IEEE Transactions on Smart Grid,* vol. 7, no. 5, pp. 2368-2377, Sep. 2016.

[5] "*The Demand Response Baseline,*" EnerNOC, Boston, MA, USA, White Paper, 2011. [Online]. Available: https://www.naesb.org/pdf4/dsmee_group3_100809w3.pdf

[6] G. Thanos, M. Minou, T. Ganu, V. Arya, D. Chakraborty, J. V. Deventer, and G. D. Stamoulis, "Evaluating Demand Response Programs by Means of Key Performance Indicators," in *Proceedings of Fifth International Conference on Communication Systems and Networks (COMSNETS)*, Bangalore, India, Jan. 2013, pp. 1-6.

[7] R. K. Jain, K. M. Smith, P. J. Culligan, and J. E. Taylor, "Forecasting Energy Consumption of Multi-Family Residential Buildings Using Support Vector Regression: Investigating the Impact of Temporal and Spatial Monitoring Granularity on Performance Accuracy," *Applied Energy,* vol. 123, pp. 168-178, Jun. 2014.

[8] R. Sevlian, and R. Rajagopal, "A Scaling Law for Short Term Load Forecasting on Varying Levels of Aggregation," *International Journal of Electrical Power & Energy Systems,* vol. 98, pp. 350-361, Jun. 2018.

[9] S. Nolan, and M. O'Malley, "Challenges and Barriers to Demand Response Deployment and Evaluation," *Applied Energy,* vol. 152, pp. 1-10, Aug. 2015.

[10] 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.

[11] F. Wang, K. Li, C. Liu, Z. Mi, M. Shafie-Khah, and J. P. S. Catalão, "Synchronous Pattern Matching Principle-Based Residential Demand Response Baseline Estimation: Mechanism Analysis and Approach Description," *IEEE Transactions on Smart Grid,* vol. 9, no. 6, pp. 6972-6985, Nov. 2018.

[12] J. L. Mathieu, P. N. Price, S. Kiliccote, and M. A. Piette, "Quantifying Changes in Building Electricity Use, With Application to Demand Response," *IEEE Transactions on Smart Grid,* vol. 2, no. 3, pp. 507-518, Sep. 2011.

[13] T. Song, Y. Li, X.-P. Zhang, J. Li, C. Wu, Q. Wu, and B. Wang, "A Cluster-Based Baseline Load Calculation Approach for Individual Industrial and Commercial Customer," *Energies,* vol. 12, no. 1, pp. 64, Jan. 2018.

[14] U.S. Energy Information Administration (EIA), "Use of Energy in the United States Explained," accessed on May 29, 2018. [Online]. Available: https://www.eia.gov/energyexplained/index.php?page=us_energy_use

[15] I. Beil, I. A. Hiskens, and S. Backhaus, "Round-Trip Efficiency of Fast Demand Response in a Large Commercial Air Conditioner," *Energy and Buildings,* vol. 97, pp. 47-55, Jun. 2015.

[16] K. O. Aduda, T. Labeodan, and W. Zeiler, "Towards Critical Performance Considerations for Using Office Buildings as a Power Flexibility Resource-A Survey," *Energy and Buildings,* vol. 159, pp. 164-178, Jan. 2018.

[17] H. Hao, T. Middelkoop, P. Barooah, and S. Meyn, "How Demand Response from Commercial Buildings will Provide the Regulation Needs of the Grid," in *Proceedings of 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, USA, Mar. 2012, pp. 1908-1913.

[18] H. Hao, Y. Lin, A. S. Kowli, P. Barooah, and S. Meyn, "Ancillary Service to the Grid Through Control of Fans in Commercial Building HVAC Systems," *IEEE Transactions on Smart Grid,* vol. 5, no. 4, pp. 2066-2074, Jul. 2014.

[19] A. Keskar, D. Anderson, J. X. Johnson, I. A. Hiskens, and J. L. Mathieu, "Experimental Investigation of the Additional Energy Consumed by Building HVAC Systems Providing Grid Ancillary Services," in *Proceedings of ACEEE Summer Study on Energy Efficiency in Buildings*, Pacific Grove, CA, USA, Aug. 2018., pp. 1-12.

[20] A. Keskar, D. Anderson, J. X. Johnson, I. A. Hiskens, and J. L. Mathieu, "Do Commercial Buildings Become Less Efficient When They Provide Grid Ancillary Services?" *Energy Efficiency*, pp. 1–15, Apr. 2019.

[21] Y. Chen, P. Xu, Y. Chu, W. Li, Y. Wu, L. Ni, Y. Bao, and K. Wang, "Short-Term Electrical Load Forecasting Using the Support Vector Regression (SVR) Model to Calculate the Demand Response Baseline for Office Buildings," *Applied Energy,* vol. 195, pp. 659-670, Jun. 2017.

[22] "*PJM Empirical Analysis of Demand Response Baseline Methods,*" KEMA, Inc., Clark Lake, MI, USA, White Paper, Apr. 2011. [Online]. Available: http://pjm.com/markets-and-operations/demand-response/~/media/markets-ops/dsr/pjm-analysis-of-dr-baseline-methods-full-report.ashx

[23] C. Grimm, "Evaluating Baselines for Demand Response Rrograms," in *Proceedings of AEIC Load Research Workshop*, San Antonio, TX, USA, Feb. 2008., pp. 1-31.

[24] S. Mohajeryami, M. Doostan, A. Asadinejad, and P. Schwarz, "Error Analysis of Customer Baseline Load (CBL) Calculation Methods for Residential Customers," *IEEE Transactions on Industry Applications,* vol. 53, no. 1, pp. 5-14, Jan.-Feb. 2017.

[25] T. K. Wijaya, M. Vasirani, and K. Aberer, "When Bias Matters: An Economic Assessment of Demand Response Baselines for Residential Customers," *IEEE Transactions on Smart Grid,* vol. 5, no. 4, pp. 1755-1763, Jul. 2014.

[26] M. L. Goldberg, and G. K. Agnew, "*Measurement and Verification for Demand Response,*" U.S. Department of Energy (DOE) and the Federal Energy Regulatory Commission (FERC), Washington, DC, USA, Technical Report, Feb. 2013. [Online]. Available: https://www.ferc.gov/industries/electric/indus-act/demand-response/dr-potential/napdr-mv.pdf

[27] G. Raman, and J. C. Peng, "A Hybrid Customer Baseline Load Estimator for Small and Medium Enterprises," in *Proceedings of 44th Annual Conference of the IEEE Industrial Electronics Society (IECON)*, Washington, DC, USA, Oct. 2018, pp. 59-64.

[28] "*Demand Response Measurement and Verification,*" Association of Edison Illuinating Companies (AEIC) Load Research Committee, Birmingham, AL, USA, White Paper, Mar. 2009. [Online]. Available: https://www.smartgrid.gov/files/demand_response.pdf

[29] "*ASHRAE Guideline 14–2014, Measurement of Energy, Demand, and Water Savings,*" American Society of Heating, Ventilating, and Air Conditioning Engineers, Atlanta, GA, USA, Dec. 2014. [Online]. Available: https://www.ashrae.org/technical-resources/standards-and-guidelines/titles-purposes-and-scopes

[30] S. Vemuri, W. L. Huang, and D. J. Nelson, "On-Line Algorithms for Forecasting Hourly Loads of an Electric Utility," *IEEE Transactions on Power Apparatus and Systems,* vol. PAS-100, no. 8, pp. 3775-3784, Aug. 1981.

[31] W. R. Christiaanse, "Short-Term Load Forecasting Using General Exponential Smoothing," *IEEE Transactions on Power Apparatus and Systems,* vol. PAS-90, no. 2, pp. 900-911, Mar. 1971.

[32] D. C. Park, M. A. El-Sharkawi, R. J. Marks, L. E. Atlas, and M. J. Damborg, "Electric Load Forecasting Using an Artificial Neural Network," *IEEE Transactions on Power Systems,* vol. 6, no. 2, pp. 442-449, May 1991.

[33] S. Karatasou, M. Santamouris, and V. Geros, "Modeling and Predicting Building's Energy Use With Artificial Neural Networks: Methods and Results," *Energy and Buildings,* vol. 38, no. 8, pp. 949-958, Aug. 2006.
[34] S. Rahman, and R. Bhatnagar, "An Expert System Based Algorithm for Short Term Load Forecast," *IEEE Transactions on Power Systems,* vol. 3, no. 2, pp. 392-399, May 1988.
[35] T. G. Kolda, and B. W. Bader, "Tensor Decompositions and Applications," *SIAM Review,* vol. 51, no. 3, pp. 455-500, Aug. 2009.
[36] D. Hong, S. Lei, J. L. Mathieu, and L. Balzano, "Exploration of Tensor Decomposition Applied to Commercial Building Baseline Estimation", submitted to 2019 7th IEEE Global Conference on Signal and Information Processing (GlobalSIP).

Appendix A

The remainder of this document provides all time series plots of the actual load (total fan power), the baselines estimated by the evaluated methods with and without the additive adjustment, and the errors of the three best baseline methods. Please see the following pages.