Reliability Demonstration of A Novel Randomized Measurement and Verification Method for Switchable Control Retrofit Using Large-scale Public Dataset

Aoyu Zou [[1]](#footnote-20) [[2]](#footnote-21), Paul Raftery 1, Stefano Schiavon 1, Carlos Durate 1

Conventional measurement and verification (M&V) methods for whole-building energy savings estimation are both time-consuming and unreliable, especially when non-routine events occur during the M&V process. Those events are unrelated to the proposed intervention strategy but have substaintial impacts on the building energy consumption. In this study, we argue that for switchable interventions (e.g. most of the control retrofits) can benefit from random sampling where the analyst randomly decide which strategy (i.e. baseline or intervention) to implement each day. We tested the novel randomized M&V method on a large public dataset which covers multiple climate zones and types of commercial buildings. We applied a virtual chilled water supply temperature reset based on outdoor weather as a control retrofit intervention. Our study shows that the new M&V method can estimate the savings accurately much quicker than the conventional method and most importantly, the estimation results are much more robust compared to the conventional method when non-routine events are present.

# 1 Introduction

## 1.1 Background

### 1.1.1 Conventional M&V

The conventional M&V process can be found in ASHRAE Guideline 14 , which are summarized in Figure 2.5. The process normally starting with baseline measurement for a whole year before retrofit. After deploying the intervention, building performance measurement continues for another year. At the end of the two-year period,

### 1.1.2 Randomized M&V

### 1.1.3 Non-routine events

A common non-routine event in an energy-saving M&V project is a change in occupancy or a substantial change in occupant behavior. Those changes have a significant impact on measured building energy consumption and are typically not caused by the intervention strategy. If the M&V analyst are not aware of the change and has no reasonable approach to make adjustment on the energy measurements, the saving estimation result is largely biased. For example, most commercial buildings are unoccupied during the pandemic in 2020 and thus building managers have observed a drastically decrease in monthly energy bills despite no energy-efficient measures were implemented. In addition, common adjustments by an building analyst only consider the variations of outdoor weather conditions such as by fitting a regression model (e.g. Time-Of-Week Temperature model).

### 1.1.4 BDG2 dataset

## 1.2 Literature review

### 1.2.1 Measurement and verification

### 1.2.2 Randomized experimental design

## 1.3 Objectives

As mentioned, the goal of a M&V project is to determine the effect (normally savings) of an energy-efficient intervention. And in this study, we further limit the study scope to switchable interventions, which mostly encapsulates control retrofits. An example intervention of such type can be a control retrofit developed by a software-as-a-service company that enables the chilled water plant to reset its supply water temperature based on outdoor weather condition. Therefore, we defined the M&V scenario as follows:

*“A company wishes to sell their supply temperature reset control software package to a building owner and guarantees its effect in reducing building electricity usage. If the building owner decides to purchase the service, the company agrees to charge the service fee based on a fraction of the measured savings.”*

As required by the M&V scenario, we assessed the performance of both the conventional and the novel randomized M&V methods by estimating the intervention energy savings for all valid buildings in the dataset. By conducting such analysis, we hope to:

1. Demonstrate the implementation of the proposed randomized M&V method using a public available dataset. We ensured the reproducibility of the M&V method by making the analysis code open source including randomized schedule generation, sequential statistical analysis, energy modeling and normalized saving calculation. Using the available open resources, building analysts should be able to seamlessly integrate and apply them in their own M&V projects.
2. Compare the energy saving estimation accuracy between the conventional and the randomized method. In particular, this study extends the comparison to large samples of buildings of various types and across multiple climate zones.
3. Verify the superior robustness of the randomized method over the conventional method. By using the realistic measurements from real-world buildings, which contains various sources of noises, could largely reflect the challenges that a building analyst would be facing in any real project. Particularly, as we will demonstrate in the following sections, non-routine events (i.e. ‘noises’ in the measurements) have less impact in the energy saving estimation when the analyst uses the randomized approach.

# 2 Method

As mentioned, this study leverages a large public dataset to demonstrate the energy saving estimation results of a novel M&V method inspired by other scientific research fields. We outlined the methodology of the study in Figure 2.1 and outlined several key components in this section.

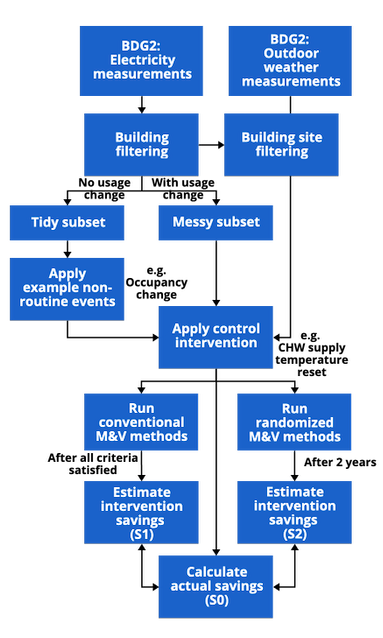


Figure 2.1: Workflow summary of the methodology of this paper

## 2.1 Building filtering

### 2.1.1 ‘Tidy’ subset

In this study, we extracted the electricity measurements from the BDG2 dataset. On a first pass, we filtered out buildings with less noise based on the following criteria:

1. Missing values < 1000: given the hourly resolution of all measurements, this is equivalent to more than a month of missing days.
2. Mean electricity usage > 0 kWh: target buildings should have active electricity usage.
3. No statistical significant difference (P-value > 0.05) between the two-year electricity usage: target buildings should have no change in the electricity usage between the two years.
4. Target buildings should have known site location: buildings with anonymous location can not request typical meteorological weather.
5. Warehouse and parking types are excluded: target buildings that are likely to have non-regular electricity and chilled water usage should be excluded.

As a summary, the resulting subset contains all buildings with ‘tidy’ measurements. Figure 2.2 shows in total, the subset contains 66 buildings in 6 types from 7 different climate zones.

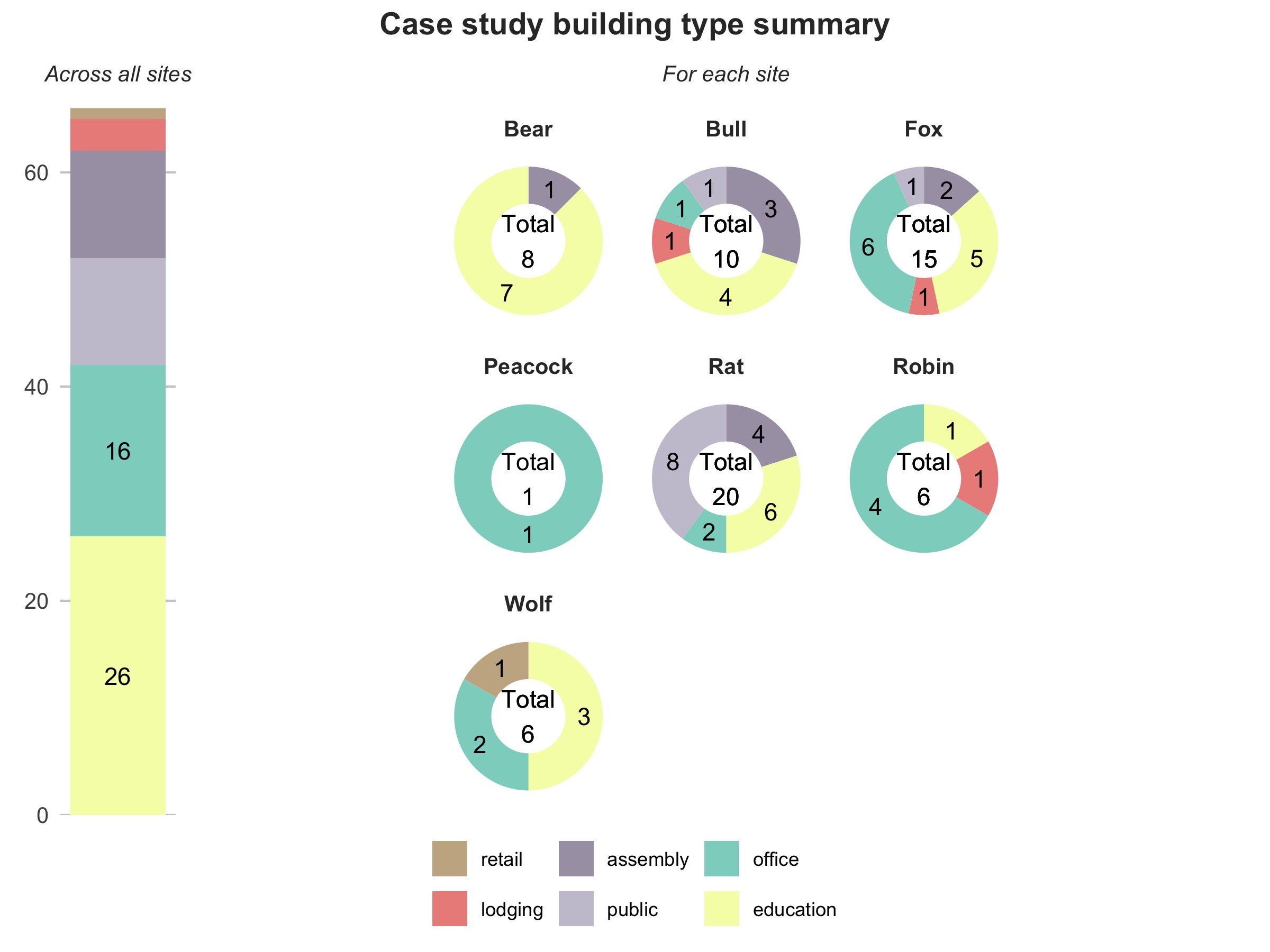


Figure 2.2: Site summary of the tidy building subset

### 2.1.2 ‘Messy’ subset

The ‘tidy’ subset might be presentative to the measurements collected among the existing building stock since in reality, whole-building electricity measurements collected over two years normally exhibit larger changes. Those measured changes can be caused by sensor itself such as lack of calibration, or inherent changes of the building, which illustrated before, can bias the result of a M&V. Therefore, to compare the robustness of the two M&V method more realistically, we included an additional ‘messy’ subset which first exclude the ‘tidy’ subset and then re-apply the filtering rule with one amendment:

1. Absolute mean difference of the two-year electricity usage < 25%: any increase or decrease of building electricity usage in the second year should be less than 25% of that in th first year.

Figure 2.3 shows the summary of the ‘messy’ dataset. As this is less aggressive, it contains 573 buildings in 12 types from 11 different climate zones.

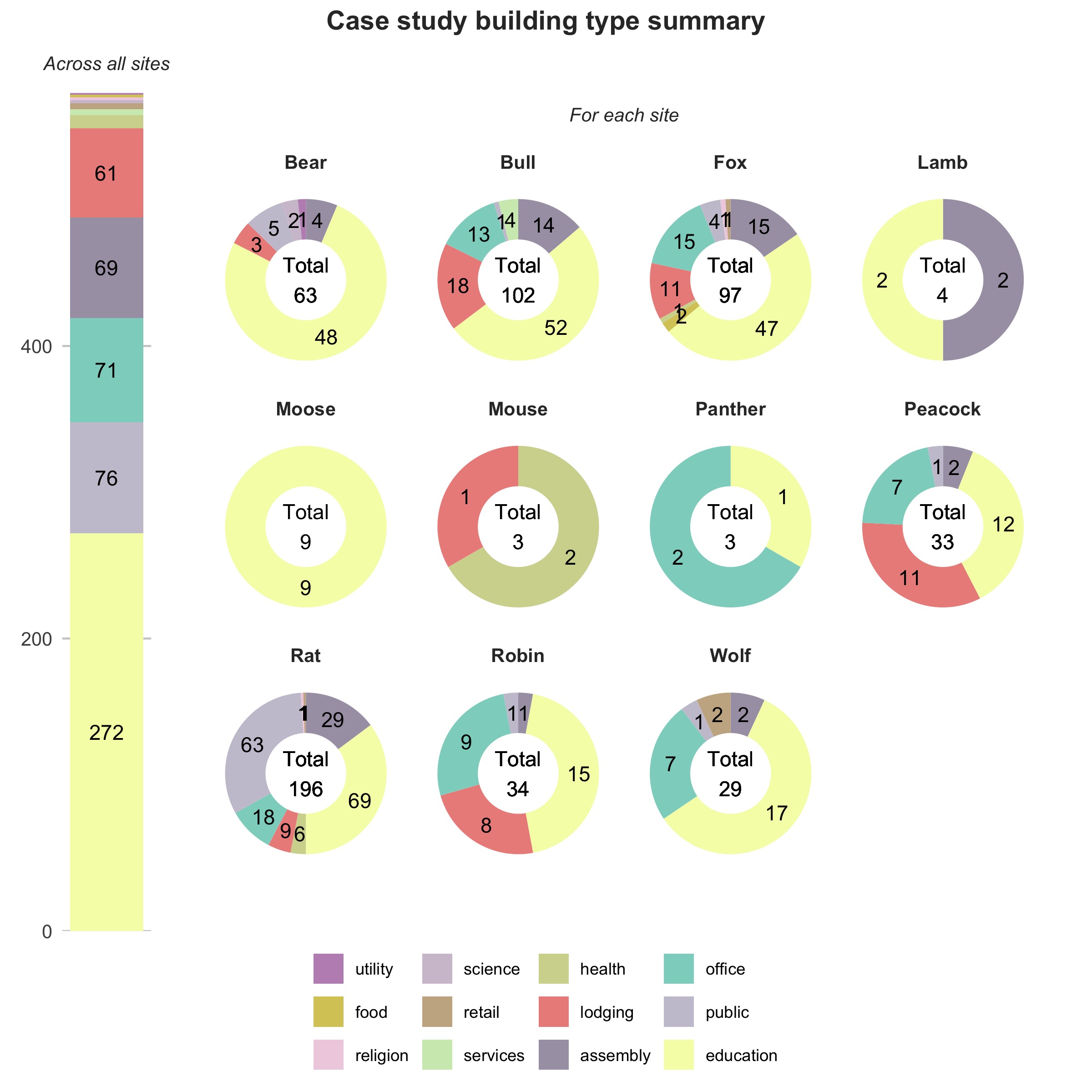


Figure 2.3: Site summary of the messy building subset

## 2.2 Apply control intervention

Figure 2.4 shows the algorithm for the proposed control intervention that reset the chiller supply temperature based on the outdoor weather conditions. Under both strategies, we assume the chiller is enabled when outdoor temperature is higher than 10°C. The baseline, which assumed to be the existing measurements from the dataset, runs at a constant water supply temperature. The intervention as the figure shows reset from 7 °C to 12 °C.

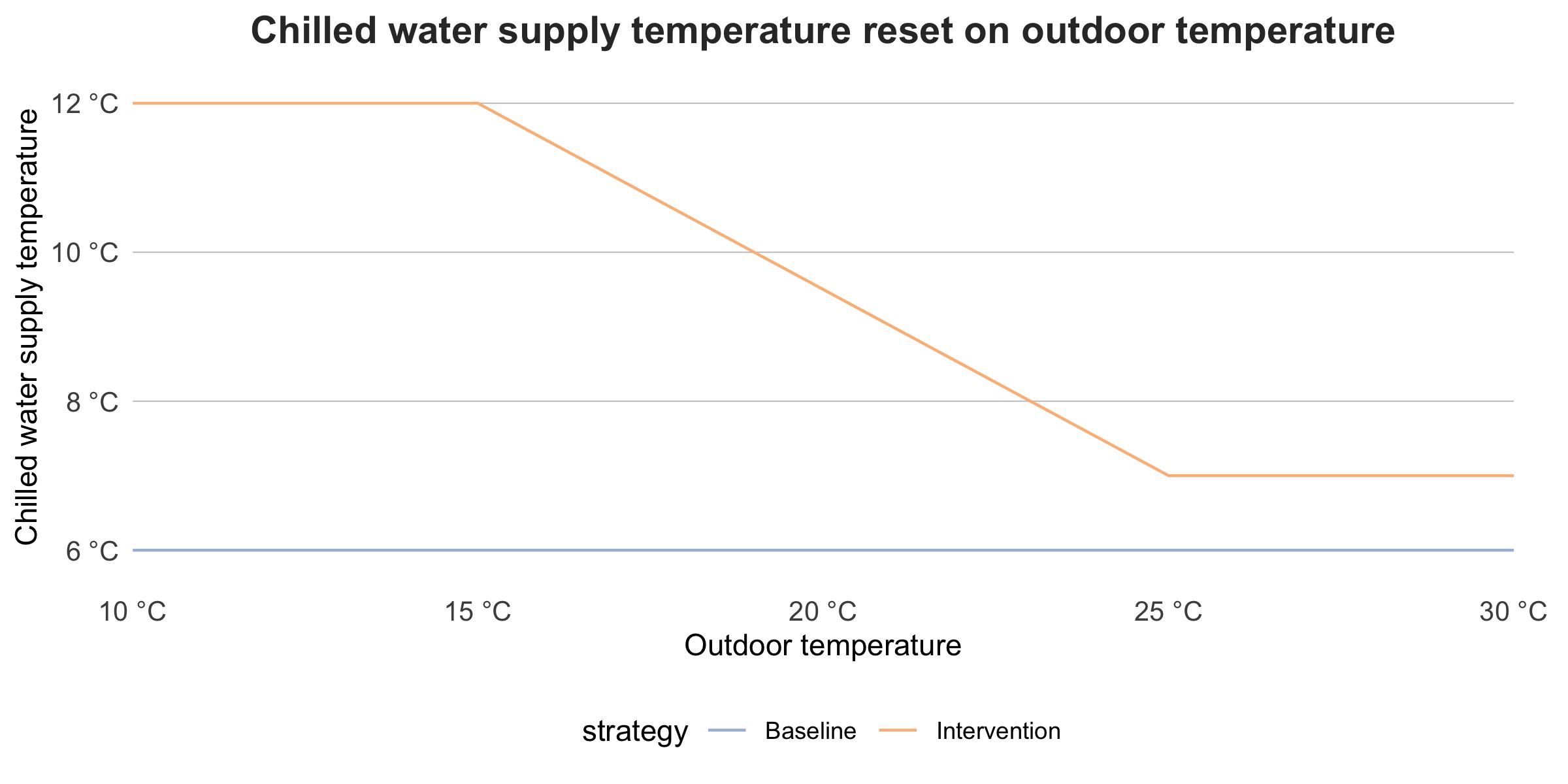


Figure 2.4: Proposed chilled water supply temperature reset based on outdoor temperature

We map the chilled water supply temperature reset to the electrical energy savings as:

We assume in general, HVAC uses around 50% of whole-building electricity and the chilled water plant further uses 50% electricity of the HVAC system. This assumption largely simplifies the unique characteristics of electricity usage in all types of commercial buildings, but for the scope of this paper, we therefore assume 25% of whole-building electricity is consumed by the chilled water plant (). In most cases, savings from an intervention is unproportional to the building hourly electricity usage, which renders M&V challenging. Therefore, we map the resulting electricity savings as a percentage of the mean electricity consumption of the plant determined by outdoor temperature (), intervention supply water temperature (), baseline supply water temperature () and hour of the day (, binary indicator whether during peak hours from 9 AM to 4 PM).

Parameters and their pre-defined values are summarized in the table below. Those parameters were not rigorously calibrated for each building, and were used uniformly across the dataset. Although this calibration process should be included, it is considered tangential to the study scope outlined in this paper.

Parameters for calculating the intervention savings

| Parameter | Description | Value |
| --- | --- | --- |
|  | % savings from setting 1 °C higher than | 0.08 |
|  | % savings adjustment during occupied hours | 1.2 |
|  | % savings adjustment during unoccupied hours | 0.8 |

## 2.3 Apply example non-routine event

To demonstrate the influence of non-routine events that are not properly adjusted, we developed three scenarios to quantify the energy consumption change associated with a hypothetical occupancy change. In summary, we assume an increase in occupancy in 2016 (i.e. refered as the ‘baseline year’ if using conventional method) resulted in 20% increase in whole-building measured electricity usage. The three scenarios specifies the increase happened during January to April (S1); May to August (S2); and September to December (S3). We applied such change only to the buildings in the ‘tidy’ subset as originally they are more likely to have no non-routine events.

## 2.4 Run M&V methods

### 2.4.1 Conventional M&V

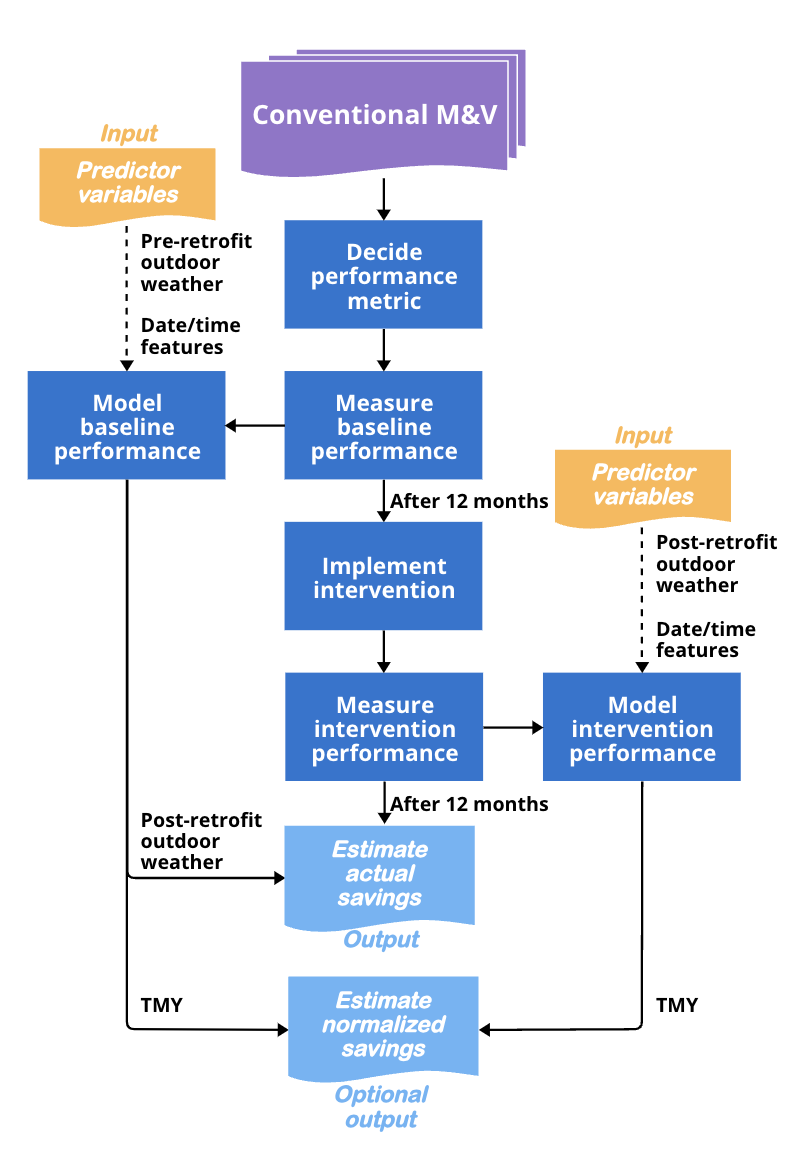


Figure 2.5: Flow chart showing the conventional M&V process for an energy-saving intervention

As described in Section 1 and Figure 2.5, conventional method for M&V is time-consuming and unable to separate non-routine event impact from measured savings. If using conventional method to estimate intervention energy savings, the result (after 24 months) is calculated by the difference between measured intervention and projected baseline in the post-retrofit period. In this study, we leveraged a piece-wise linear regression considering time-of-week and outdoor temperature (TOWT) as independent variables for projection and normalization on typical meteorological year .

### 2.4.2 Randomized M&V

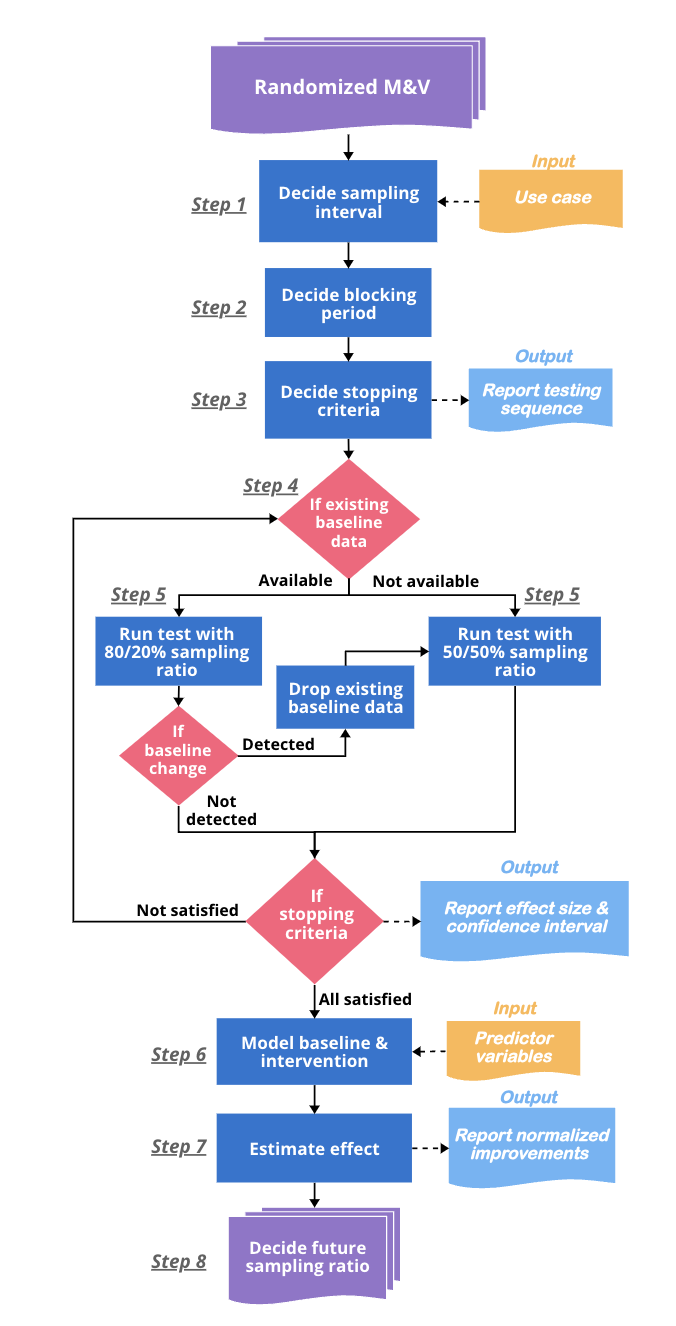


Figure 2.6: Flow chart showing the novel randomized M&V process for an energy-saving intervention

Compared to the conventional method, the randomized M&V can provide an estimation more rapid and reliable. Section 1 describes the application of randomization techniques and Figure 2.6 specifically outlined the process of conducting a randomized switchback experiment for control retrofit projects. The method sequentially evaluate the intervention effect and returns a saving estimation with 95% confidence interval when all stopping criteria are satisfied. describes all stopping criteria and recommened threshold to consider in more details and we breifly summarized the thresholds used in this paper here:

* The HVAC system operates from 06:00 to 22:00 each day, so we use a daily sampling interval with the sampling time at midnight each day.
* Block by day of the week with a block period of 12 weeks.

Stopping criteria are:

* A minimum and maximum of 12 and 108 weeks respectively. The randomized schedule covers the entire two-year period but stopping criteria enables an early stop at the end of satisfied blocking period.
* At least 80% of the drybulb temperature range in the annual TMY data sampled by both strategies.
* Test for no carryover effect using a t-test with a p-value not exceeding a defined significance threshold of 0.05.
* 90% confidence that energy savings exceed or do not exceed 0% using the SPRT test. Medium effect size (d = 0.5) quantified by cohen’s d and calculated SPRT statistics either falls below the lower threshold or exceeds the upper threshold.
* As no baseline data is available, test with an equal sampling ratio (50% baseline, 50% intervention).

To remain consistency, savings normalization on typical meteorological year is also modeled through TOWT.

# 3 Results

## 3.1 M&V methods comparison

In this section, we compare the performance of two M&V methods. The key aspects of the assessment include: 1) time required to reach a saving estimation: in most cases, a shorter M&V timeline reduces associated cost and interruption for the building owner; and 2) saving estimation accuracy: this is particularly important for any software-as-a-service company to set a reasonable price with their customers.

### 3.1.1 Savings estimation time

The timeline for the conventional M&V method is outlined in previous sections. Although ASHRAE Guideline 14 offers minimum requirements for each path, the required baseline period is either a full range of all independent variables (typically outdoor weather conditions) under normal facility operation or a 12-month worth of continuous measurements. The same requirements also applies to intervention installed in the post-retrofit period. Thus, the conventional M&V method is likely to run over 24 months or even longer due to missing data. For example if an analyst is asked to determine the energy savings of a chilled water plant retrofit but missed most of the cooling season due to delay in retrofit deployment, he/she needs to wait until the next cooling season to measure the savings.

If the randomized M&V method is applicable, even if there is a delay in the retrofit deployment, the building owner faces less risk and impact since random sampling allows measuring the full range of all independent variables simultaneously among all control strategies or modes. In addition, sampling at an equal probability (i.e. 50%/50% between baseline/intervention) effectively balances the level of independent variables measured. In other words, unless one strategy is sampled with a large amount of consecutive days (e.g. 7 days or more), it is rare that the independent variable, such as outdoor weather condition, is measured significantly different among all sampled control strategies. Figure 3.1 shows the average estimated timeline for each building across all sites and climate conditions if using randomized M&V for the applied chilled water supply temperature reset intervention. The timeline figures for each individual building are attached in the supplementary material. The figure shows for all climate zones, covering a sufficient range outdoor weather condition is the most stringent criterion. However, most buildings can satisfy the 80% range by 6 months and all buildings can quantitatively determine the savings with associated uncertainty by 9 months, which is even shorter than the length of baseline measurement required by the conventional method.

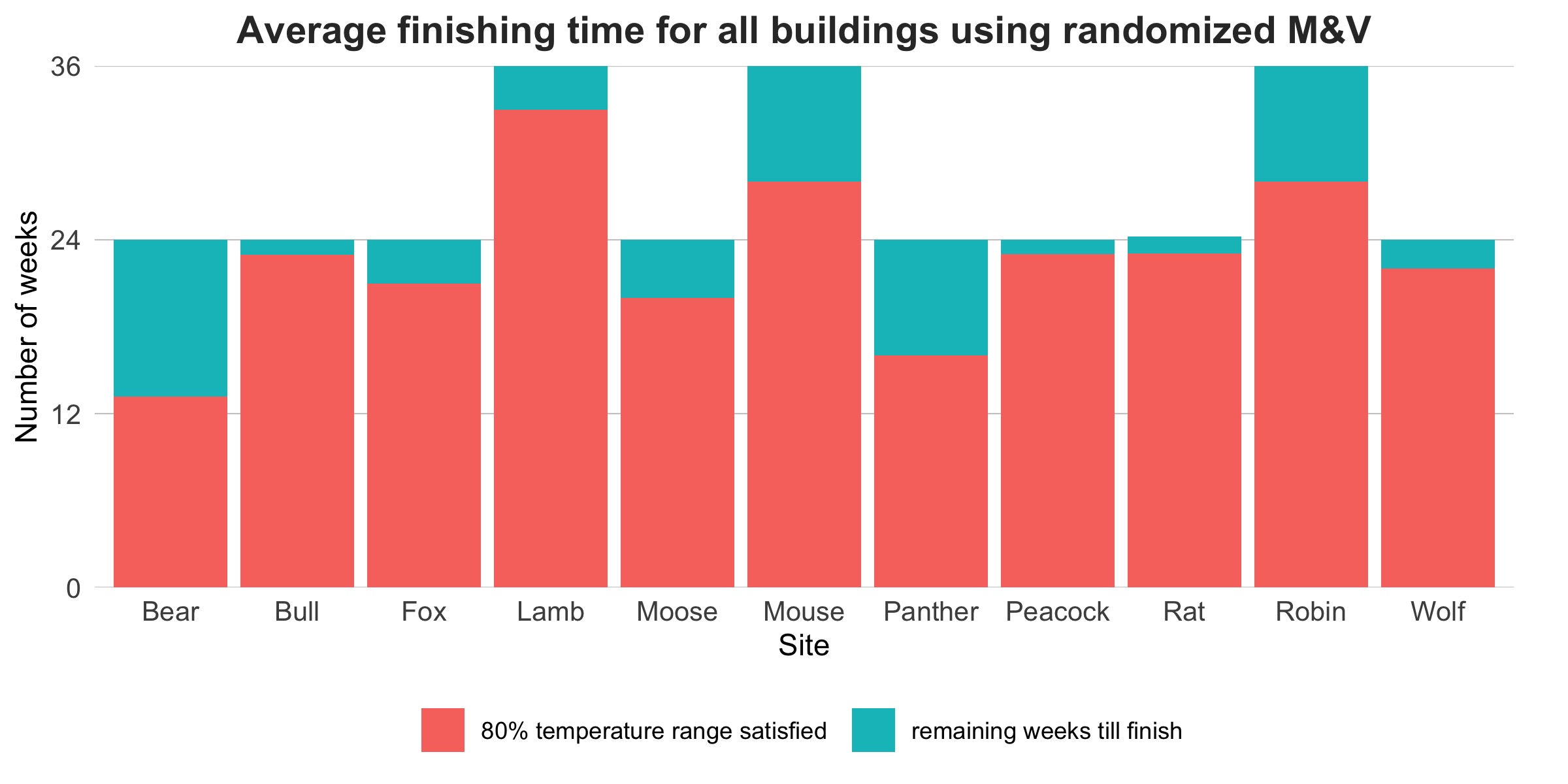


Figure 3.1: Average randomized M&V timeline summary for all buildings at each site

### 3.1.2 Savings estimation accuracy

In this study, we define the overall target savings (i.e. the ground truth) that an M&V analyst wishes to detect, either in normalized savings, fractional savings or simply the measured difference between the baseline and the intervention measurements, is the mean electricity energy reduction over the two-year period from implementing the chilled water supply temperature reset control. We estimate the M&V savings by the two methods following the process given in Figure 2.5 and Figure 2.6. In orther words, we compare the conventional M&V savings at the end of the two-year period with the randomized M&V savings after all criteria are satisfied.

Figure 3.2 shows the overall comparison of the savings estimation from all buildings normalized on the typical meteorological year weather conditions of each site. In subplot a), the narrower range of the true savings indicates the dependence of the intervention effect on the outdoor weather condition, which is intended as shown in Figure 2.4. Sites with mild climate all year round such as locations in California shows higher savings potential above 10%, while locations with more extreme climate such as Washington DC shows only 6% savings annually. In addition, although the box plot indicates some savings estimation are ‘outliers’, they should not be interpreted as such since the distribution can be biased by the unbalanced sample sizes across different site shown in Figure 2.2 and Figure 2.3. In subplot b), the results show that the conventional M&V method tends to estimate savings with a greater uncertainty and we will demonstrate in the following sections that non-routine events can have a significant impact. For the randomized method, we displayed two scenarios, the estimation at the end of two-year period with 50%/50% sampling throughout and the estimation provided after all stopping criteria satisfied (i.e. stops early at 24 weeks or 36 weeks). Since we found no significant difference between the two distributions as shown in the figure, we demonstrated that an early stop has no impact on the estimation accuracy but can significantly reduces the cost associated with M&V.

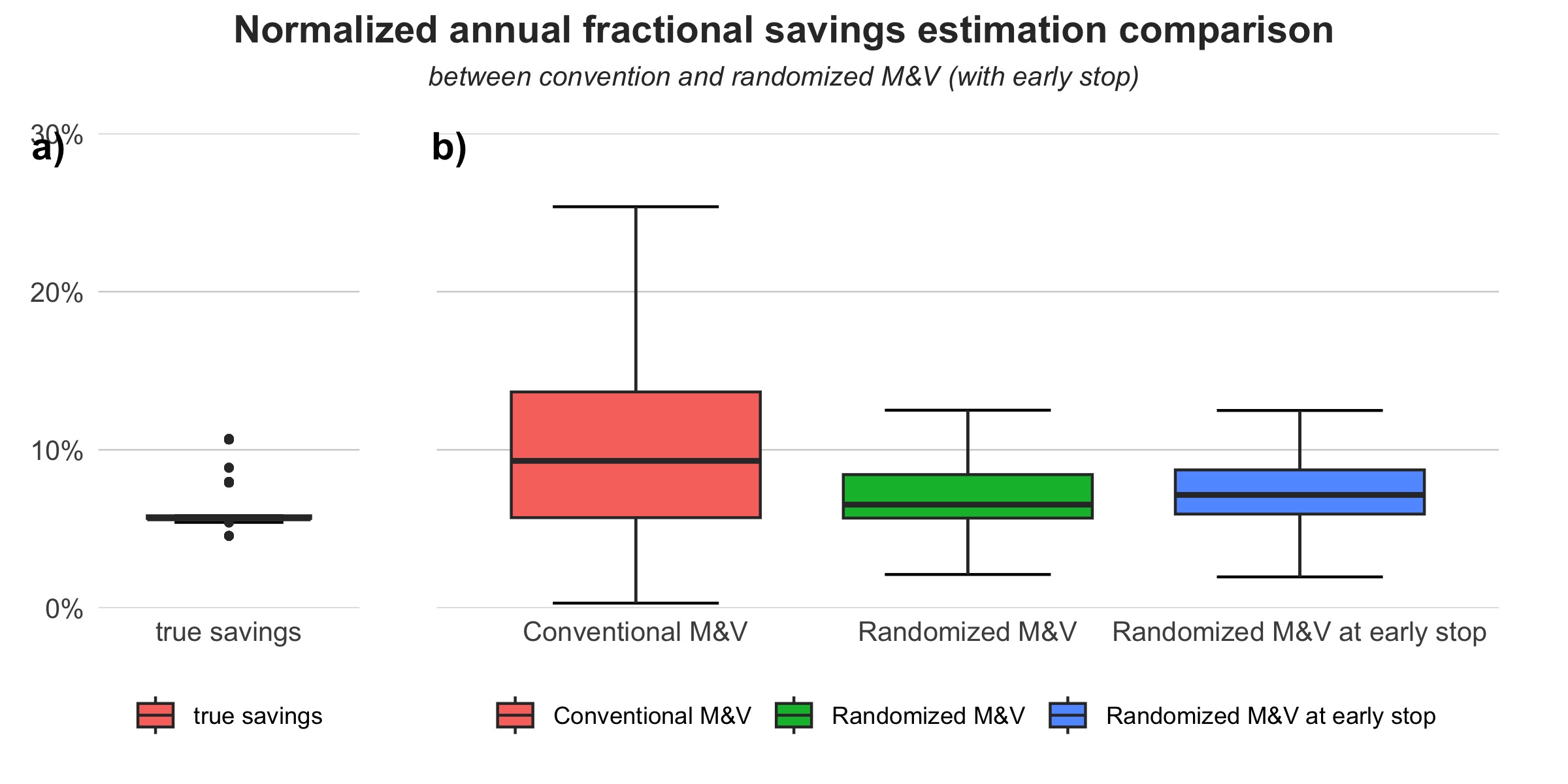


Figure 3.2: Comparison of normalized annual fractional savings on site TMY between conventional and randomized method

To assess the accuracy on individual building, we calculate the mean absolute difference between the true savings and estimated savings. Figure 3.3 shows the comparison between the two methods using the tidy subset. Because of the large variation in electricity energy usage, we express the savings as a normalized value which is a fraction of energy reduction compared to baseline. Furthermore, we focus on the median and 95% threshold of the mean absolute difference distribution in the comparitive assessment to filter out outliers. As a result, the conventional method performs better than the randomized method. One hypothesis is that the tidy subset only includes buildings exposed to similar usage throughout the two years, meaning the measurement used for regression model fitting resembles the model prediction results. Although the results indicate the conventional M&V method is preferable if a building can achieve long-term consistency in electricity usage, such condition is rarely garrenteed in reality.

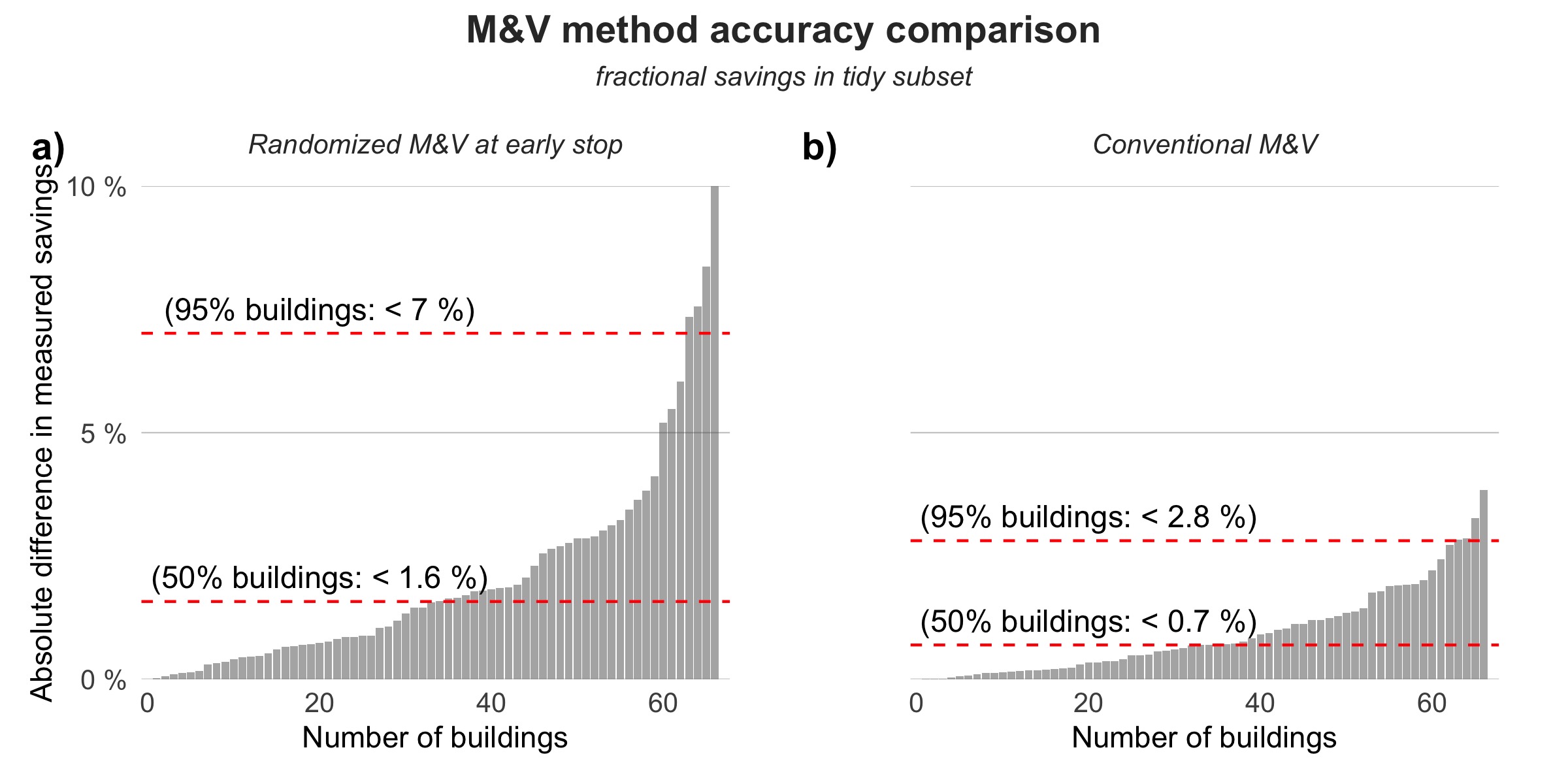


Figure 3.3: Absolute deviation of M&V estimated fractional savings from true target savings using tidy subset

A more realistic condition is shown in the messy subset in Figure 3.4. The randomized method demonstrates consistent accuracy in savings estimation, even in the presence of significant measurement noise, indicating strong robustness. On the contrary, the conventional method shows approximately in 50% of all cases, the deviations of M&V estimation are more than 5%. We showed the non-normalized mean absolute difference comparison plots in the supplementary material. The poor performance shows if the there exist a change in electricity usage, baseline projection based on regression models are unreliable. Those changes are normally referred as non-routine events, which summarizes all the influential factors that are not considered in the regression model either because they are impractical to measure, such as hourly occupancy rate, or too random with no expected patterns to quantify, such as a sudden change of use from an office to a warehouse. We will demonstrate this in more details in the following sections.

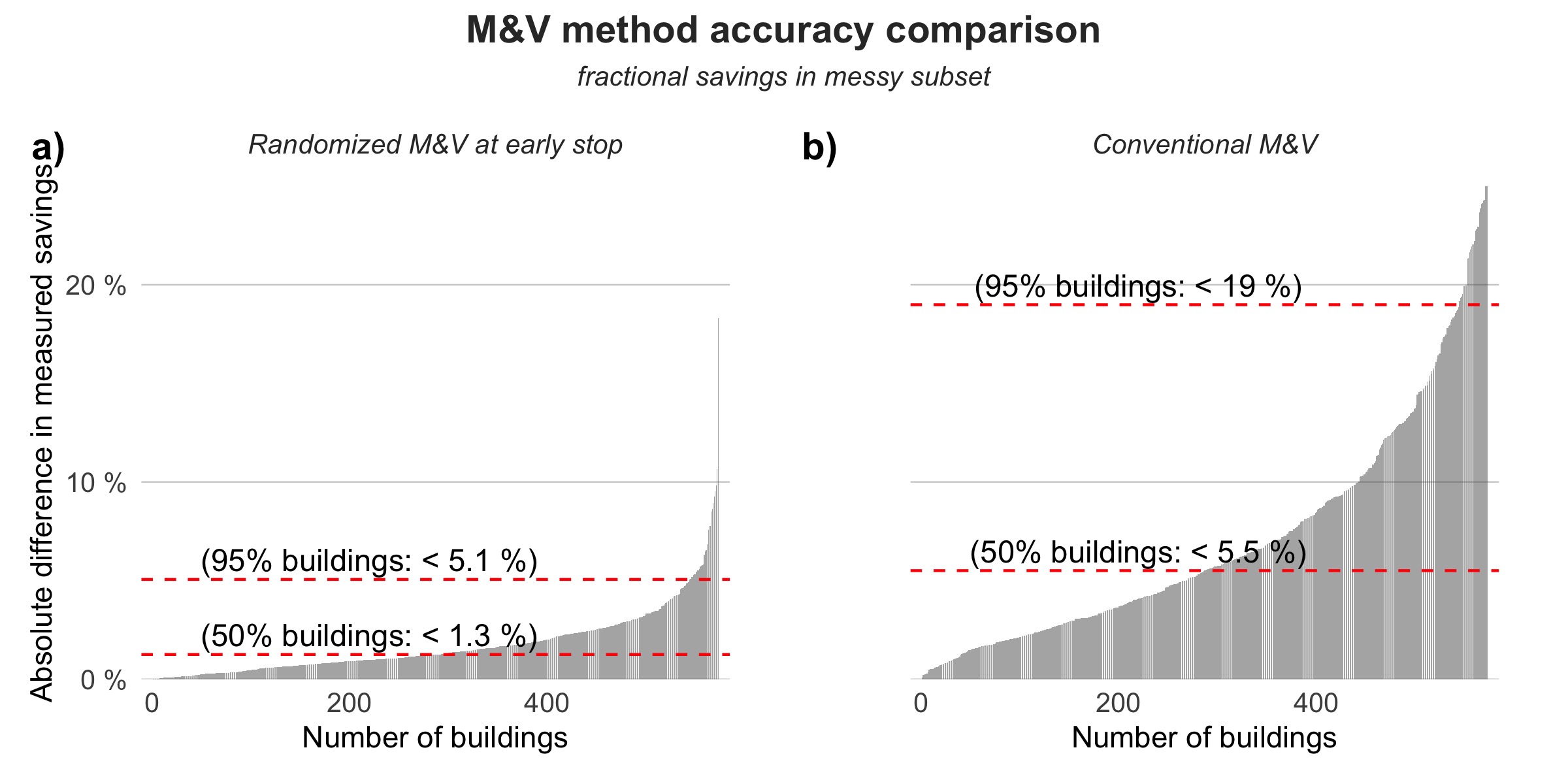


Figure 3.4: Absolute deviation of M&V estimated fractional savings from true target savings using messy subset

## 3.2 Non-routine events impact on savings estimation

In this section, we demonstrate the influence of non-routine events through two scenarios. In the first scenario, we quantified how much occupancy would influence the electricity usage and added to the tidy measurement set as they originally indicates no such change. In the second scenario, we simply run the two M&V methods on the messy measurement sets before adding the chilled water supply temperature reset. Since no intervention is added, the more reliable method should detect a ‘saving’ closer to 0.

### 3.2.1 Occupancy change

Although occupancy can be approximated in several ways in commercial buildings, such as through counting check-in, or WIFI connections, or monitoring indoor CO2 concentration as a proxy. However, it is not a cost-efficient measures to add to the routine operation for most of the buildings and there might be privacy concerns associated. The scenario proposed hypothesize that during the baseline measurement period of the target building, a floor of tenants moved out leaving the space unoccupied for four months before new tenants moved in and this led to a 20% electricity decrease as a result. The M&V protocol only requires the M&V analyst to fit a TOWT model so he/she only measures outdoor temperature. We showed one example in Figure 3.5 where from May 1st 2016 to August 31st 2016, the target building electricity decreased 20% due to reduced occupancy. Thus subplot (a) shows measured baseline with such change, which are then used for TOWT model fitting. Starting from 2017, the occupancy returns normal and only intervention strategy was measured and the projected baseline is the prediction results from the fitted model. To demonstrate, we also plotted the original measurements before adding the intervention effect as ‘adjusted baseline’. This assumes that we can accurately adjust the baseline in the post-retrofit period on back-to-normal occupancy. Subplot (b) shows that the fitted regression underestimate the true baseline condition leading to underestimated savings.

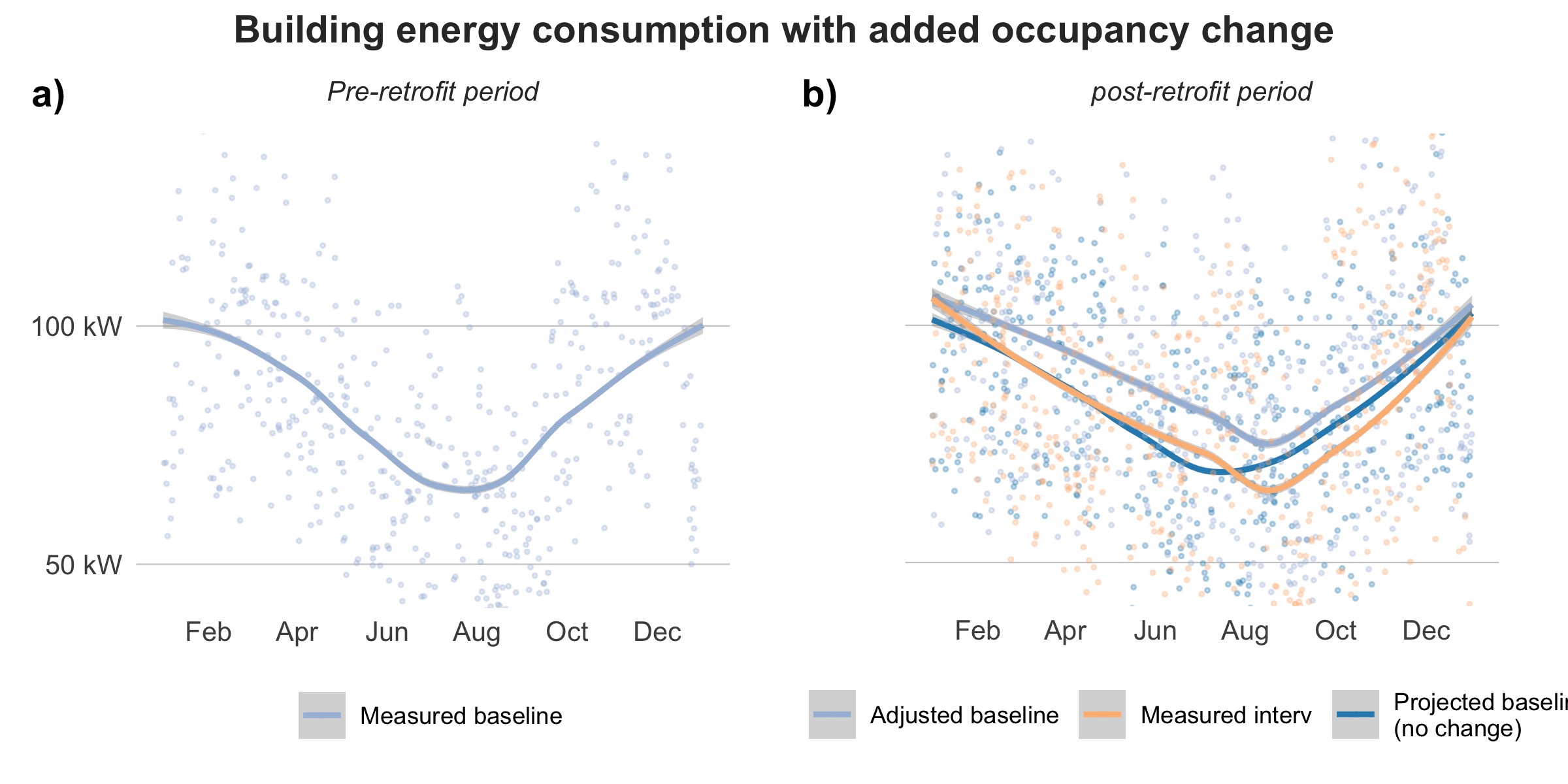


Figure 3.5: Occupancy change impact on TOWT model fitting and baseline projection

Figure 3.6 shows the normalized fractional savings estimation accuracy comparison between the two M&V methods. To compare which method is more robust to the applied occupancy change, a difference-in-difference value is plotted in the figure. Each bar in the plot represent the difference in the deviation of savings estimated by the convention method and the randomized method:

In other words, a positive value indicate the randomized method provides an estimation more aligned to the true target saving. Thus the plot shows the randomized method shows uniformly superior robustness for all sites. The red-dotted line shows the absolute deviation in savings estimation by the randomized method as a reference for scaling. The deviation is consistent with the distribution shown in Figure 2.2 and thus further reinforce its robustness to non-routine events.

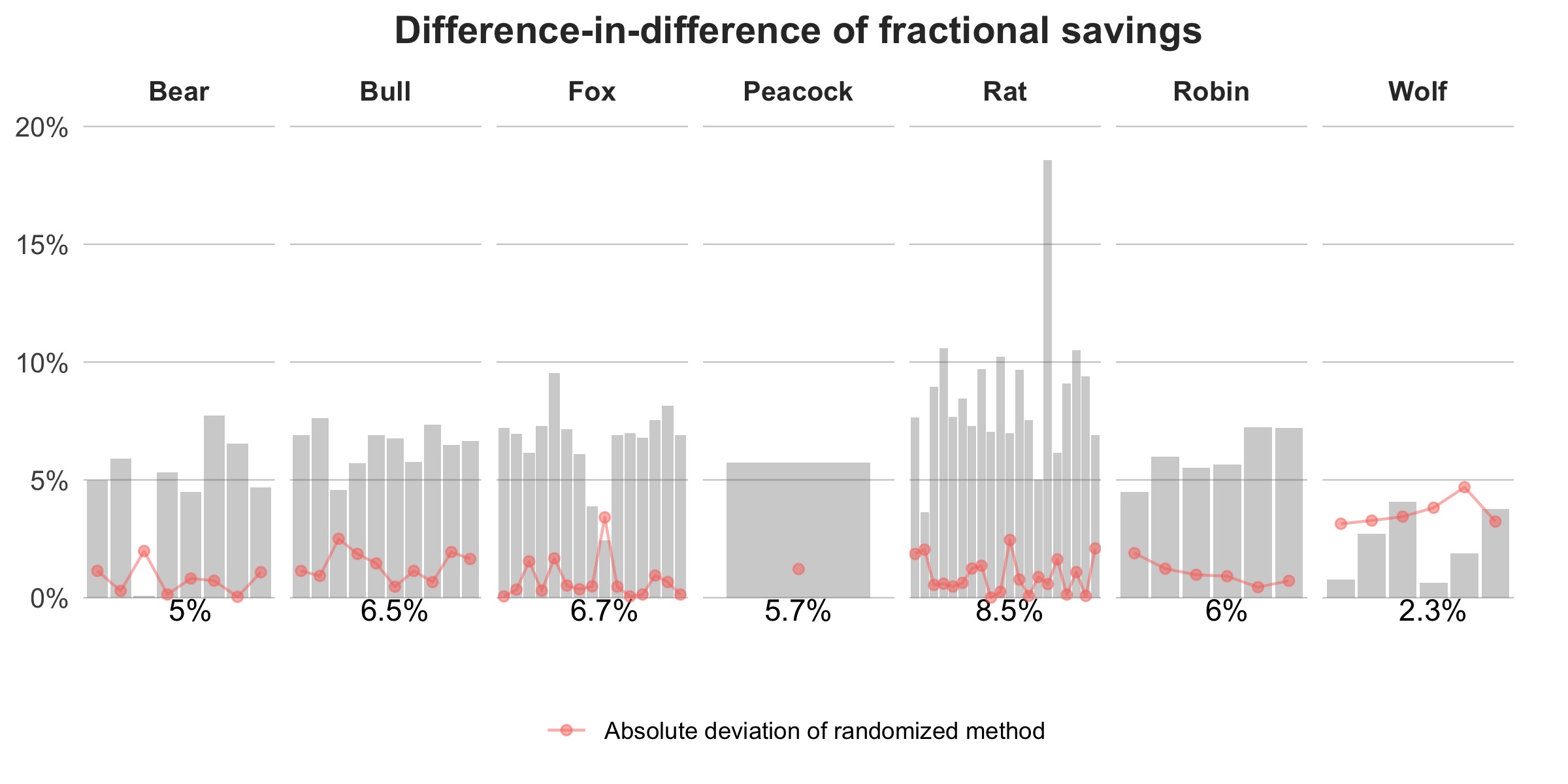


Figure 3.6: Savings estimation accuracy comparison between two M&V methods with added occupancy change (with the site average difference-in-difference displayed at the bottom of each site).

### 3.2.2 No-saving detection

The outbreak of pandemic in 2020 followed by working-from-home policy is a practical example of building energy consumption influenced by non-routine events. During unoccupied months, building energy consumption dropped compared to the same time period in the previous years. If an energy-efficient intervention was deployed onsite starting at 2020, given by the whole-building electricity measurements, the large difference in measured electricity usage in 2020 and projected baseline usage in 2019 can hardly indicate the retrofit intervention effect.

In other words, if there is no intervention applied or the intervention effect is known to be null (e.g. constantly being overriden by the baseline), an ideal M&V method should detect no savings. To test this, we kept the randomized schedule as sampled earlier but removed added chilled water plant intervention effect. Therefore, all sampled intervention days are considered additional sampled baseline days. Earlier we mentioned the ‘messy’ dataset contains some change in the two-year electricity measurement, which is purely due to various non-routine events such as occupancy change or even other measures implemented by the building manager but irrelevant to the chilled water set point reset intervention. Thus, when the proposed chilled water reset intervention was removed, a more reliable M&V method should overcome the influence of those confounding factors and inform no savings.

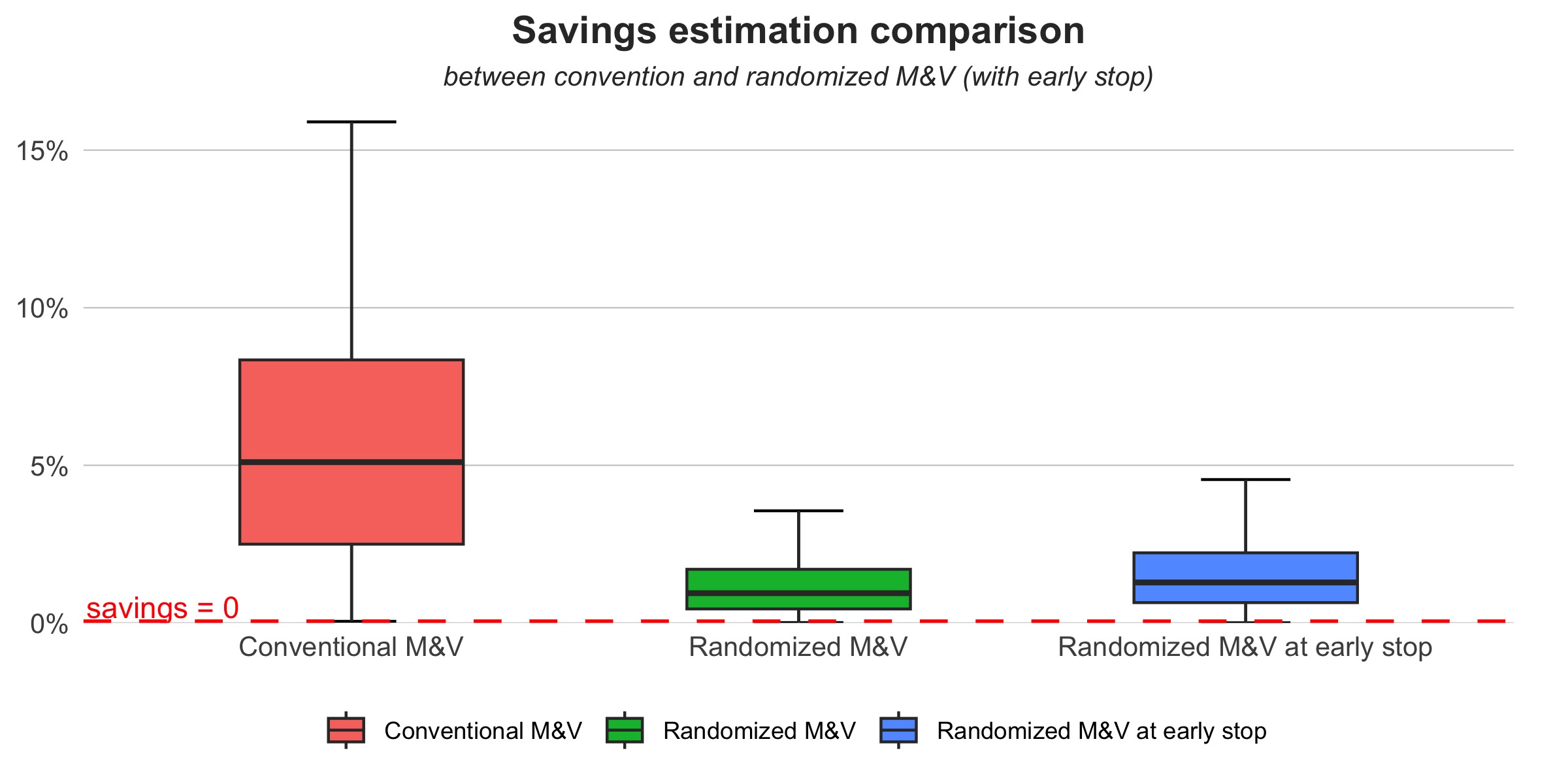


Figure 3.7: Distribution of savings estimation results by the convention M&V method and randomized method.

Figure 3.7 shows that the conventional M&V method in average detected 5% savings when there should be none. The randomized method even when stops right after satisfying all criteria, the savings estimation is much closer to 0%. This is mostly because when there is a change or disturbance applied to the target building such as lighting retrofit or thermal leakage in the envelope, randomly sample the two strategies at 50%/50% ensures the resulting effect either as a decrease in lighting electricity or an increase in reheat electricity is balanced between the two implemented strategies.

# 4 Discussion

## 4.1 TOWT modeling accuracy

Although this paper argues that regression model prediction results can be largely biased by various non-routine events, we further clarify that this is not related to regression model fitting accuracy. Time-of-week temperature model is used here for its simplicity and convenience since the model only requires outdoor temperature measurements and time of week as independent variables. The model assumes a linear composition of building energy consumption as temperature-dependent and time-dependent load. The time-dependent component accounts for day-to-day variations and the temperature dependent component considers a piecewise linear relationship across a range of temperature intervals.

To assess modeling accuracy, we used the Coefficient of Variation of Root-Mean Squared Error, or CV(RMSE) as the error metric. Since this metric is calculated as a normalized value, it is useful to compare different model fitting results. Guideline 14 requires that whole-building baseline model fitting accuracy should maintain a CV(RMSE) lower than 30%. In addition One study focusing on the baseline energy data-driven model fitting indicates that TOWT performs as accurate as other more advanced machine learning models and the calculated CV(RMSE) distribution for a large sample of commercial buildings indicates a median of 20%. Figure 4.1 plots the distribution of the model fitting accuracy calculated separately for the two measurement sets. The result shows that the TOWT model performance is even better in this study compared to the literature and no significant difference between the two subsets. This is mostly related to low-quality measurement sets filtering in the pre-processing mentioned in Section 2.

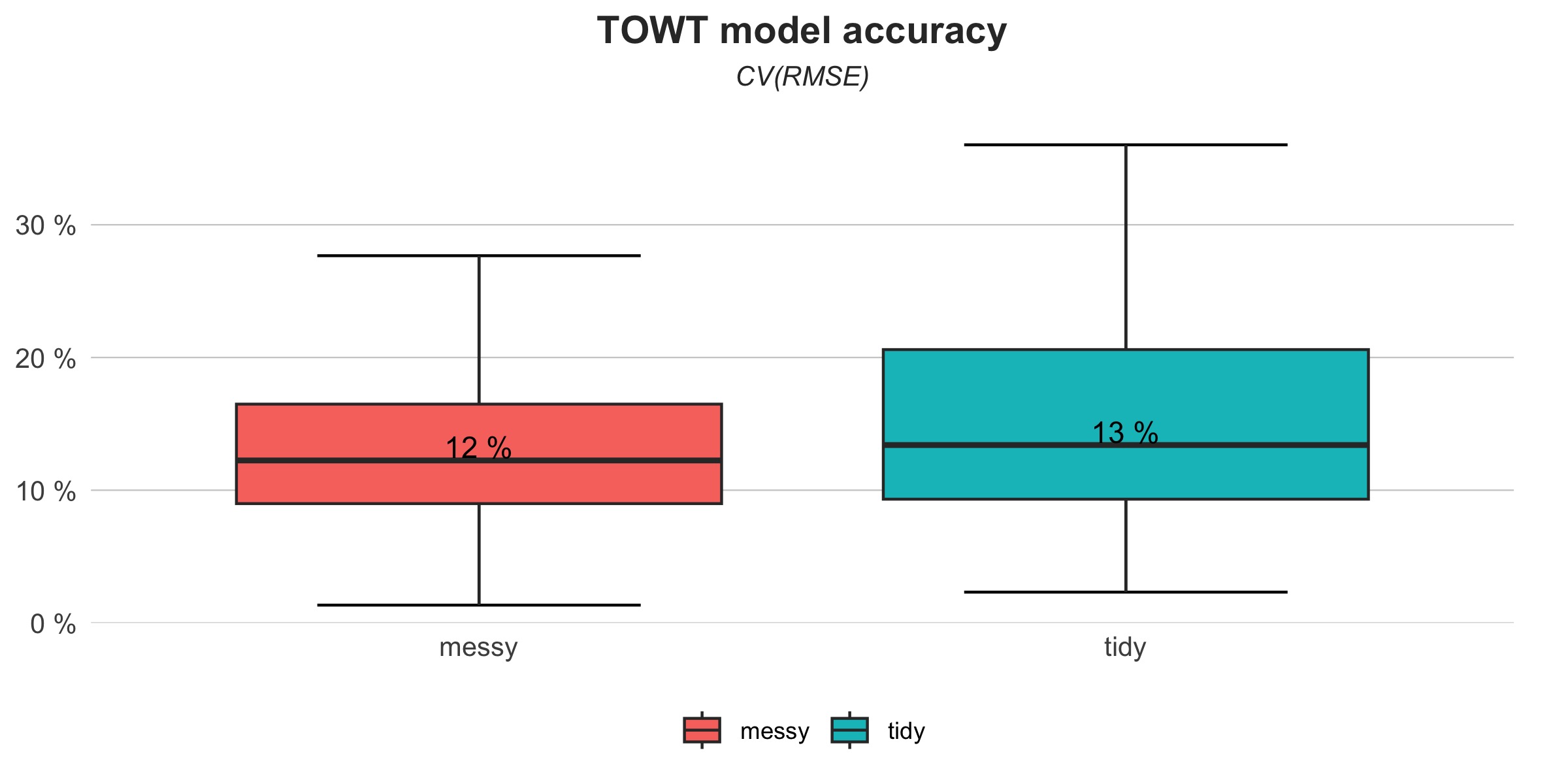


Figure 4.1: TOWT model fitting accuracy distribution on filtered measurement set

In general, the limitation is that despite the regression model tends to capture well on the mean building energy consumption but can underestimate and peak or overestimate the lower base load , which is not useful for assessing demand response. Additionally, a data-driven model can learn the training dataset very efficiently with high accuracy, but the well-trained model is challenging to generalize prediction outside the range of training set.

## 4.2 Sampling ratio for randomized M&V

Another advantage of using randomized M&V is the flexibility of changing sampling ratio after the M&V. For example, the building owner can continue sampling at 50%/50% between the baseline and the intervention to further reduce the uncertainty range associated with the savings. Or he/she can switch to 100% intervention to optimize energy savings, but whether the existing baseline becomes outdated is unknown. A compromised approach is to continue sampling at 80%/20% between the baseline and the intervention. Figure 4.2 shows after the analyst reports the randomized M&V results, a new schedule sampling at 80%/20% was implemented for another 36 weeks.

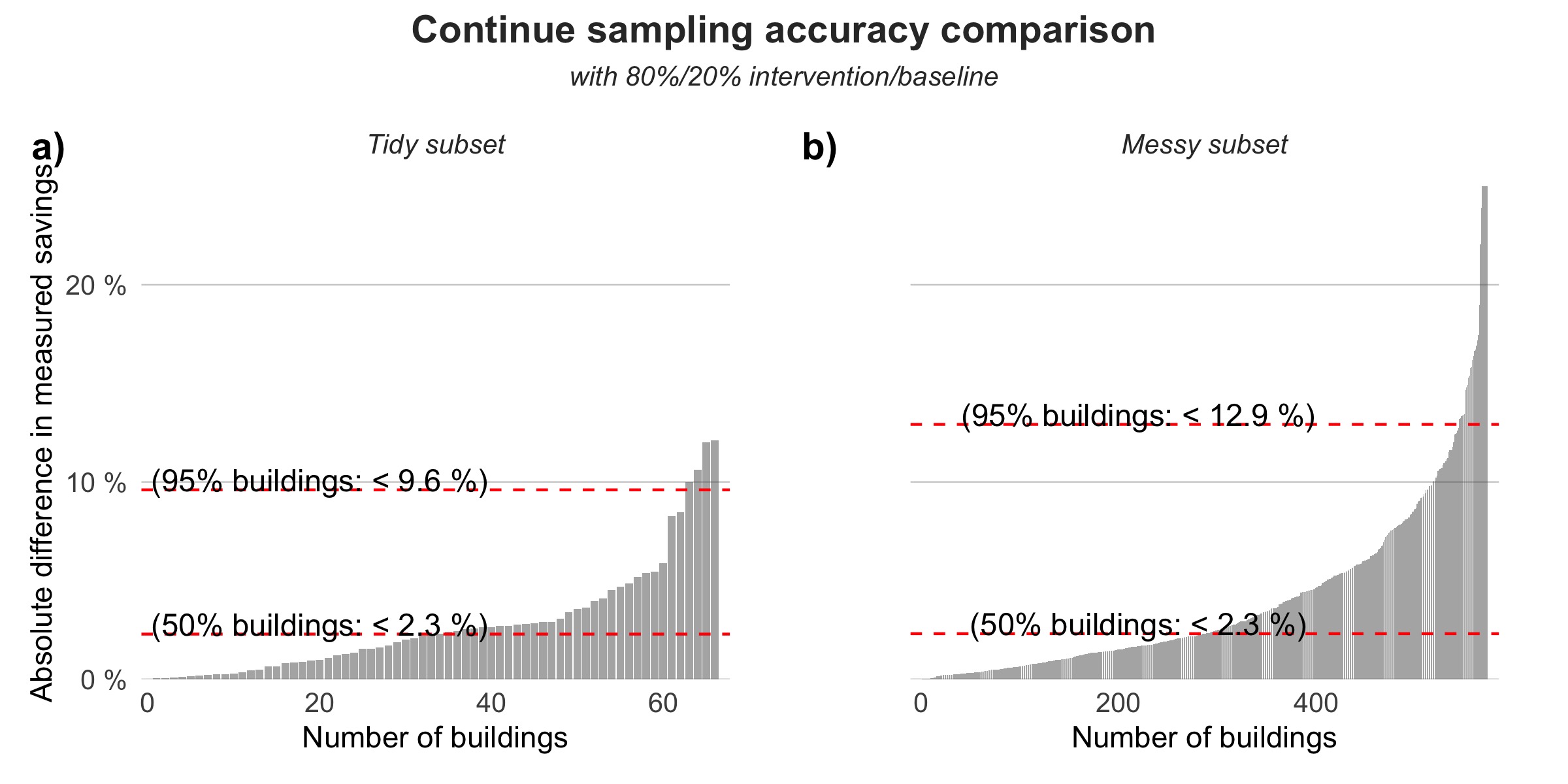


Figure 4.2: Savings estimation accuracy for continue sampling at 80%/20% for 36 weeks after all stopping criteria are satisfied

The deviation remains consistent in the tidy subset but increases slightly in the messy subset, suggesting that changes in energy usage in the target building create a trade-off between measured savings and estimation accuracy when down-sampling the baseline. In this scenario, opting for an early stop at 24 or 36 weeks with 50% intervention and resampling 80% of the data for 36 weeks still allows the building owner to capture 65% of the full-range savings.

## 4.3 Limitations and future study

Considering the diversity of the measurement sets, the assumption made in this study is relative general. For example,

# 5 Conclusion

1. Center for the Built Environment, University of California Berkeley, USA [↑](#footnote-ref-20)
2. Correspondence to [aoyuzou@berkeley.edu](mailto:aoyuzou@berkeley.edu) [↑](#footnote-ref-21)