Demonstrating the Reliability of Randomized Measurement and Verification for Switchable Control Retrofits Using a Large Open-Source Dataset

Aoyu Zou [[1]](#footnote-20), Paul Raftery 1, Stefano Schiavon 1, Carlos Duarte 1, Gail Brager 1

Conventional measurement and verification (M&V) methods rely on pre- and post- retrofit comparison to estimate energy savings. They are often time-consuming and unreliable, especially when non-routine events such as reduced occupancy, or equipment failures occur during the M&V process. Those events are unrelated to the intervention but significantly affect building energy consumption and thus when the analyst applies conventional M&V suggested by industry guidelines, the results can be largely confounded. In this study, we demonstrated that switchable interventions, such as most of the control retrofits, can benefit from a new M&V method which randomly samples whether to implement the baseline or the intervention strategy each day. We tested this novel randomized M&V method on a large public dataset covering various climate zones and commercial building types, using a virtual chilled water supply temperature reset based on outdoor weather as the intervention. Our results show that compared to the conventional method, the randomized method provides faster and more accurate savings estimation. Additionally we found that when non-routine events are present, the randomized M&V approach estimates savings that are much closer to the ground-truth savings than the conventional M&V method, demonstrating much improved reliability.

# 1 Introduction

## 1.1 Conventional M&V

Measurement and Verification (M&V) is the process of quantifying energy savings from energy efficiency projects by comparing actual energy consumption against a baseline, adjusting for factors like weather and occupancy. This process ensures that improvements in energy performance are accurately evaluated. In the United States, practitioners often refer to ASHRAE Guideline 14, the International Performance Measurement and Verification Protocol (IPMVP), and the Federal Energy Management Program (FEMP) for standard guidelines (DOE 2008; Efficiency Valuation Organisation 2012; ASHRAE 2023). These guidelines outline standardized methods for quantifying energy savings, whether through calibrated simulations or monitored measurements for specific equipment or systems (isolation methods) or for entire buildings (whole-building methods). In this study, we will focus on the energy savings quantified at the whole-building level where the measurements are obtained from utility bills or whole-building meters, however the overall conclusions would also apply to isolation methods. Typically, the process begins with baseline measurements taken during the year before implementing the  energy-efficiency retrofit, followed by the same measurement procedure during the year after the intervention. After collecting two years of data, an M&V analyst fits an energy prediction model, using variables such as outdoor temperature and time (Mathieu et al. 2011) to project baseline energy consumption in the post-retrofit period. The difference between the counter-factual baseline and the measured intervention period energy consumption represents the energy savings. One of the key drawbacks of this method is its reliance on a two-year timeline to quantify savings, during which baseline measurements can become outdated due to changes in building performance caused by non-routine events unrelated to the intervention. This limitation reduces the feasibility of rapid M&V and complicates the quantification of estimation uncertainty, thus impacting the accuracy and timeliness of savings assessments.

## 1.2 Randomized M&V

To address the limitations of conventional M&V methods and the challenges posed by non-routine events, we propose a novel M&V method that adopts the randomized crossover design, a gold standard from medical and agricultural studies (Blackston et al. 2019; Duan, Kravitz, and Schmid 2013; Gupta et al. 2019; Mie et al. 2022; Munro, Wager, and Xu, n.d.; Raseduzzaman and Jensen 2017). Another improvement is that we proposed a sequential evaluation framework and defined stopping criteria to end the M&V if the target effect is detected. This is to avoid unnecessary measurement collection over the full 2-year M&V cycle. The full framework is detailed in a previous published study with all stopping criteria and example usecases outlined (Raftery et al. 2024). In summary, this method provides M&V analysts with a randomized schedule that alternates between baseline and intervention implementation while ensuring balanced sampling across days of the week and seasons. For example, given a 10-week M&V period for 1 intervention, the balanced randomized schedule would equally sample 5 Mondays with the baseline strategy in operation and 5 Mondays with the intervention strategy in operation. The limitation of this method is that it is only applicable to a subset of retrofit projects where interventions can be easily switched on and off. However, for all applicable use cases, it allows analysts to detect energy savings sequentially right after test begins. In addition, once the desired savings target is achieved, analysts can terminate the M&V and switch to 100% intervention. The key advantage of randomization is that if control strategies are sampled with equal probability, the influence of other confounding factors such as occupancy change (or commonly known as non-routine events) is likely to be evenly distributed among measurements, leading to a more accurate and unbiased assessment of the intervention’s impact.

## 1.3 BDG2 dataset

The Building Genome Dataset 2 (BGD2) is an extensive open-access dataset designed to advance research and development in building energy efficiency and control strategies acting as a test-bed for modeling, simulation, and algorithm development (Miller et al. 2020). BGD2 contains over 1000 buildings’ metadata and realistic operational measurements between 2016 and 2017 from across North America and Europe, making it one of the most comprehensive collections of building-related data available for scientific use. The dataset includes various commercial building types such as offices, education facilities, public, and retail buildings, and provides detailed information on their physical characteristics (e.g. energy ratings, heating types and floor area) and hourly measurements of chilled and hot water, electricity, gas usage as well as site outdoor weather conditions. In this study, we mostly use the metered electricity and outdoor weather measurements for running the M&V methods.

## 1.4 Literature review

### 1.4.1 Whole building approach

ASHRAE Guidelinge 14 and IPMVP provides options for whole building M&V use cases. For code compliance using prescriptive option, 12 months of baseline and 12 months of postretrofit measurements are required. In addition, the expected savings should be larger than 10% and the baseline model fitting accuracy should have CV(RMSE) lower than 20%. Thus, most research related to M&V for whole-building approach focuses on the accuracy of baseline modeling, exploring model performance from simple regression models to more complex machine learning techniques. One study reviewed various models suitable for M&V applications as well as selected input features (Alrobaie and Krarti 2022) and another study provided a definitive methodology to apply machine learning models for M&V use cases (Gallagher et al. 2018). In addition, a few studies investigated the critical performance metrics to evaluate the developed baseline models (Granderson et al. 2015; Granderson and Price 2014) and compared a variety of models using those metrics (Granderson et al. 2016). These studies made significant contributions by emphasizing the uncertainty associated with the model-fitting process, a key factor in accurately determining energy savings. Furthermore, other researchers addressed this issue by leveraging statistical formulation and inference to improve baseline energy models (Burkhart, Heo, and Zavala 2014; Heo, Choudhary, and Augenbroe 2012; Walter, Price, and Sohn 2014). However, a gap still remains in the literature regarding the rigorous quantification of uncertainties directly associated with calculated savings, for instance, accounting for the potential bias that baseline model might deteriorate (i.e. becomes ‘stale’) over an extended period of pre- and post-analysis.

### 1.4.2 Non-routine events impact

Non-routine events during M&V commonly refers to unexpected changes in a building that influences its energy usage. These changes can greatly affect measured energy consumption in buildings and are typically unrelated to the intervention strategy. Therefore, those changes are considered ‘static factors’ respect to measured independent variables. A common non-routine event in energy-saving M&V projects is a change in occupancy or a significant shift in occupant behavior, equipment run time, and operating conditions (e.g. set points, lighting and ventilation levels). However, current standards or guidelines only provides general approach for consideration. For example, IPMVP requires facility owner and the M&V analyst periodically perform inspections of all equipment and operations during reporting period, which is labor intensive and error-prone. ASHRAE guideline 14 recommends performing engineering calculations or computer software simulations to adjust post-retrofit baseline. Additional, it is very rare to have access to all measurements such as occupancy, and thus the analyst normally assume those factors remain unchanged throughout the study. Specifically, studies investigating the effects of demand response on building energy efficiency commonly uses linear interpolation to estimate counter-factual baseline (Beil, Hiskens, and Backhaus 2015; Keskar et al. 2020), and one study points out that it is inaccurate to assume no change in the operating conditions during the test period (Huang, Katipamula, and Lutes 2023). One likely encountered non-routine event due to operation condition change is filter clogging in air handling units due to particle accumulation. This can cause supply fans to gradually consume more energy to maintain required duct static pressure (Feng and Cao 2019; Zhai and Nathaniel Johnson 2017). If M&V analysts are unaware of such changes and lack an appropriate adjustment method (e.g., replacing filters before the intervention begins), the savings could be underestimated as increased energy use is incorrectly attributed to the intervention rather than the mechanical issue. Some studies realized the limitation of current M&V methods, which only consider adjusting for outdoor weather, is insufficient and emphasized the importance of requiring matched comparison groups to control for exogenous factors beyond weather differences when comparing between baseline and intervention (Demand Side Analytics 2022; Huang, Katipamula, and Lutes 2023). However, those methods are still unable to accurately quantify uncertainties due to changing baseline.

## 1.5 Objectives

As mentioned, the goal of an M&V project is to determine the effect—typically energy savings—of an switchable energy-efficient intervention. The goal is to determine, for a large sample of buildings, how much more accurately the novel M&V method would estimate the savings of such an intervention compared to conventional M&V, and how much more quickly it would reach a result. An example of such an intervention is a control retrofit developed by a software-as-a-service company that adjusts the chilled water plant’s supply water temperature, which can be commonly found in the literature (Duarte et al. 2023; K.-P. Lee and Cheng 2012; Qiu et al. 2022; Jin, Du, and Xiao 2007; Taylor 2012). In our case, we make it even simpler by adjusting setpoint based on outdoor weather conditions with more description shown in Section 2.2. Therefore, we defined the M&V scenario as follows:

“A company aims to sell its supply temperature reset control software package to a customer, in this case, the building owner, with a promotion that it will reduce the building’s electricity usage. If the building owner decides to purchase the service, the company agrees to charge a service fee based on a percentage of the measured energy 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 BG2 dataset. By conducting such analysis, we aim to:

1. Compare the energy saving estimation accuracy between the conventional and the randomized method. This study extends the comparison to a large sample of buildings, covering a variety of types and climate zones. The comparison metrics include both estimation accuracy and M&V completion timeline.
2. Verify the enhanced robustness of the randomized method. By using realistic measurements from real-world buildings, which include various sources of noise, we aim to reflect the challenges faced by building analysts in real projects. As we will perform numercial simulation on the existing dataset, the ground-truth savings can be calculated as the reference for comparison.
3. Open-source implementation of the proposed randomized M&V method using a public dataset. We ensured the reproducibility of the method by making the analysis code open source including randomized schedule generation, sequential statistical analysis, energy modeling and normalized saving calculation. In addition, we also included codes for extended use cases of the randomized method, such as changing sampling ratio and sampling intervals. Using the available open resources, building analysts should be able to seamlessly integrate and apply them in their own M&V projects.

# 2 Method

We outlined the methodology of the study in Figure 2.1 and extended several key components in this section.

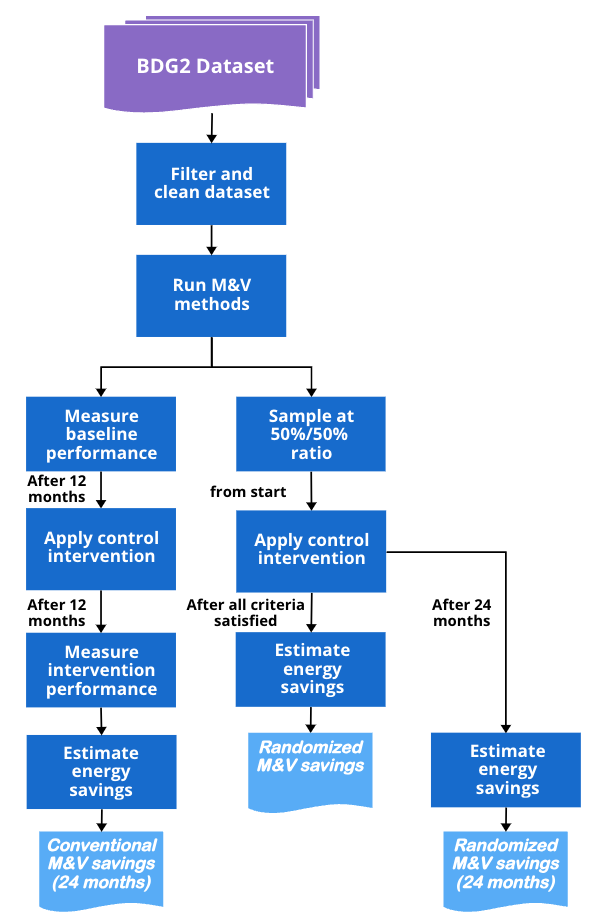


Figure 2.1: Flow chat showing the methodology for comparing the estimated savings of randomized M&V with the conventional M&V

## 2.1 Filter and clean dataset

In this study, we extracted the electricity measurements from the BDG2 dataset and filter out all qualified buildings based on the following criteria:

1. Missing values < 1000: given the hourly resolution of all measurements, this is equivalent to 1.5 months of missing days.
2. Mean electricity usage > 0 kWh: target buildings should have active electricity usage.
3. : any increase or decrease of building electricity usage in the second year should be less than 25% of that in the first year.
4. Electric EUI < 750 kWh/: excludes buildings at the top 5% electrical energy usage intensity according to the statistics provided by the Building Performance Database (Lawrence Berkeley National Lab, n.d.; Mathew et al. 2015).
5. Warehouse and parking types are excluded: target buildings have less demand flexibility to implement a chilled water set point reset control.

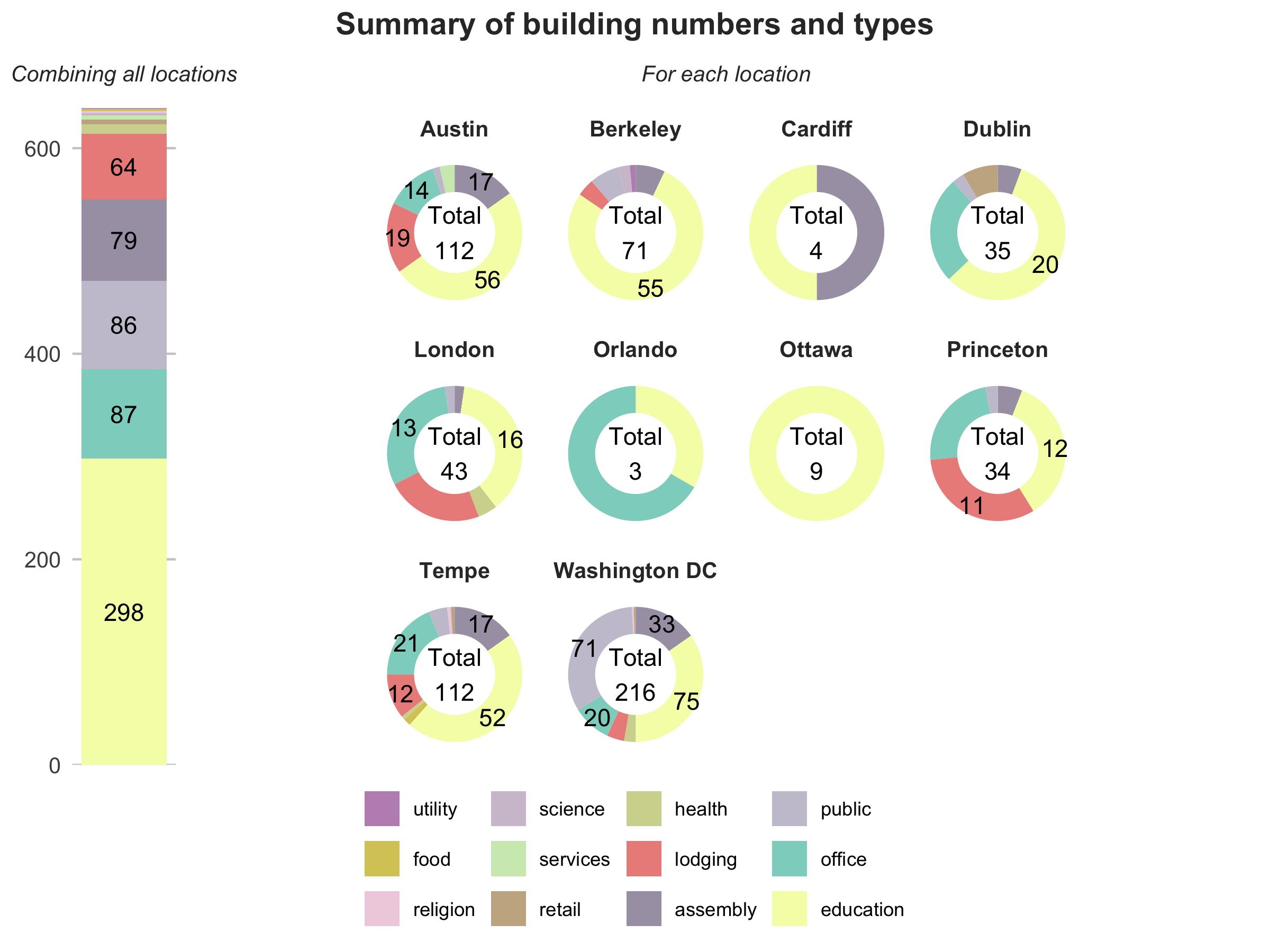


Figure 2.2: Site summary of the stable building subset (counts < 10 are omitted for visualization; left: aggragated counts of buildings for each type; right: breakdown building counts for each building type at each location)

For all qualified buildings, we further added one more stringent criterion to explore all buildings with very stable electricity usage between the two years:

1. No statistical significant difference (P-value > 0.05) between the two-year electricity usage, which are sometimes assumed to be the case for whole-building measurement and verification.

As a results, there are only 66 buildings were labeled in this ‘stable’ subset (out of total of 600 buildings) implying that the assumption made by conventional M&V is overly simplified. As mentioned in Section 1.4.2, such variability is largely associated with non-routine events and according to the statistics shown here, is commonly observed in real applications.

## 2.2 Apply virtue intervention

Figure 2.3 shows the algorithm for the proposed control intervention that reset the chiller supply temperature based on the outdoor weather conditions, which can be commonly found in the literature (Y. J. Lee et al. 2022; Čongradac and Kulić 2012). For both strategies, we assume that the chiller is activated when the outdoor temperature exceeds 10°C. The baseline strategy, representing the existing measurements from the dataset, operates with a constant water supply temperature. The intervention strategy, as illustrated in the figure, adjusts the water supply temperature dynamically, resetting it from 7°C to 12°C.

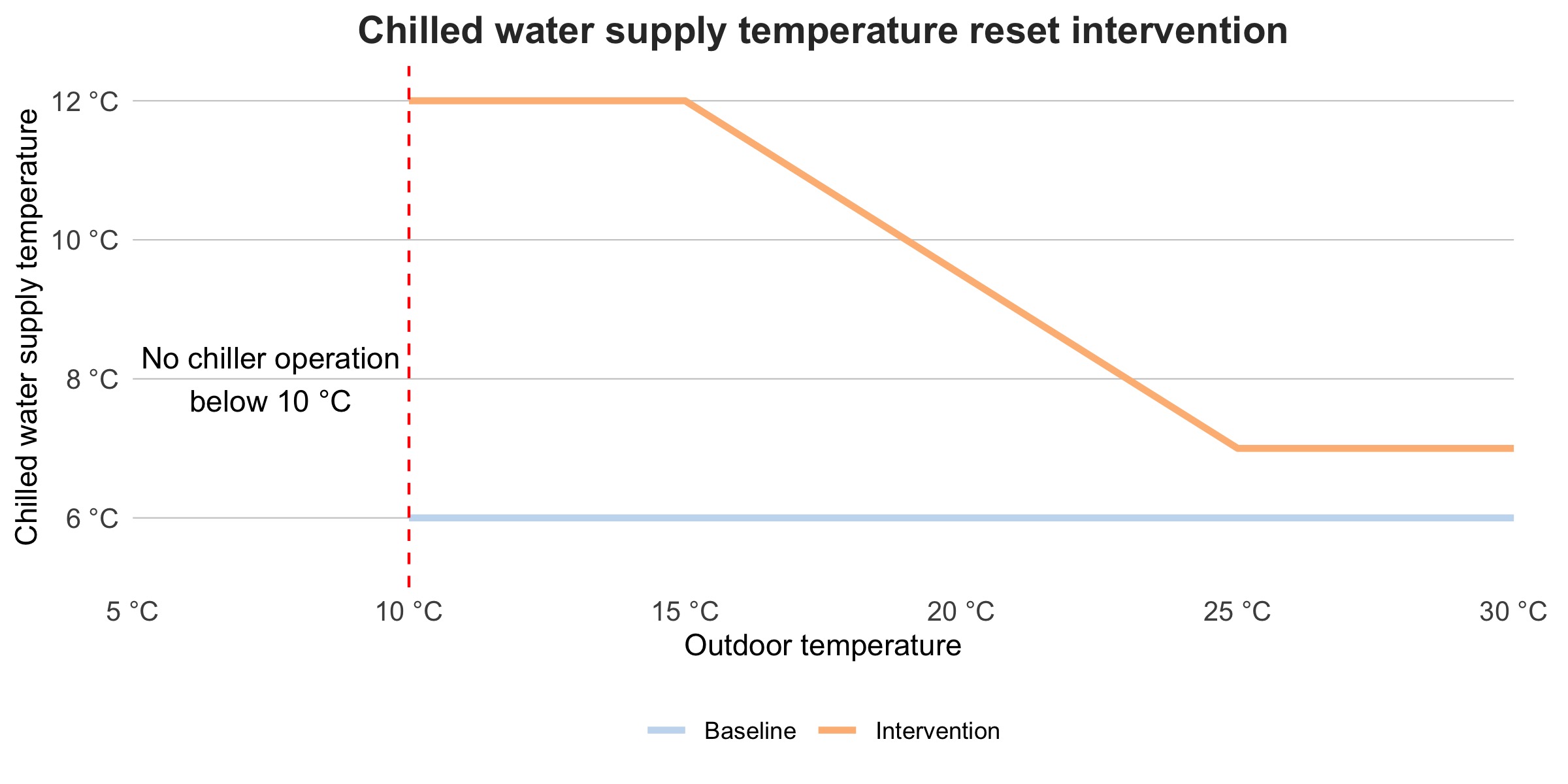


Figure 2.3: Proposed intervention strategy: chilled water supply temperature reset based on outdoor temperature

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

We assume on average, HVAC systems account for approximately 50% of a building’s total electricity consumption, and the chilled water plant further consumes around 50% of the HVAC electricity. While this assumption largely simplifies the diverse energy usage across various building types, for the scope of this paper, we assume that 25% of the total building electricity is used by the chilled water plant, (Administration 2012). Typically, the savings from an intervention are not proportional to the building’s hourly electricity usage, which is generally the challenge for M&V. To address this, we mapped the resulting electricity savings as a percentage of the plant’s normal operation, calculated as its mean electricity usage over the two-year period. This percentage is influenced by factors such as 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. For simplicity, those parameters were not rigorously calibrated for each building and were applied uniformly in the filtered dataset. Interested readers can also change the parameters in our open-source code to simulate different scenarios.

Table 1. 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 Run M&V methods

We described more in detail the workflow of both conventional and randomized M&V methods in the previous study (Raftery et al. 2024). The conventional method is a simple pre- and post- comparison on a 12-month baseline and 12-month intervention timeline. The randomized method defines sampling requirements as:

* 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.

and stopping criteria as:

* 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 dry-bulb 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.
* test with an equal sampling ratio (50% baseline, 50% intervention).

# 3 Results

The accuracy () is calculated as:

where indicates true electricity usage and indicates the estimated electricity usage either through the conventional method or the new randomized method. We present the distribution of savings estimation accuracy using boxen plots, also known as letter plots, which is an advanced variation of the box plot designed to extend beyond the interquartile range (IQR) by progressively dividing the data into smaller percentiles, revealing more detail in the tails of the distribution. We set the division parameter to k = 4, meaning the entire box area represents the data distribution from 6.25% to 93.75%. As the steps move closer to the center line (50% median), the distribution range progressively narrows with the next step represents 12.5% to 87.5%, followed by 25% to 75%. In addition, for each boxen plot, we also showed the mean value of the distribution on top.

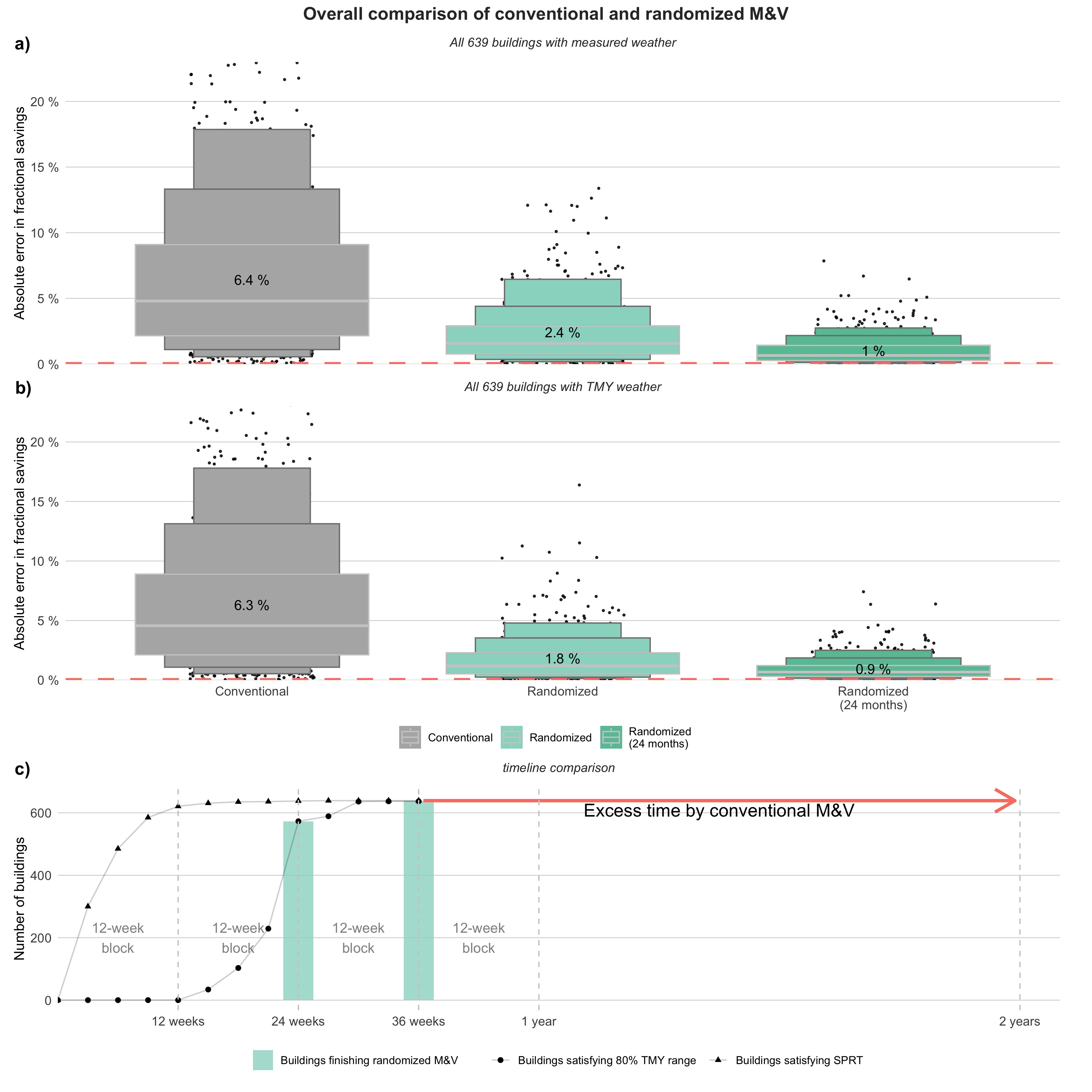


Figure 3.1: Overall comparison results between the conventional M&V method and the proposed randomized M&V method (both at the stopping criteria and over a two-year period).

Figure 3.1 shows the overall results of M&V methods comparison. Subplots a) and b) calculate the savings estimation error as the absolute deviation from the true savings (i.e. ) and shows the results distribution using conventional M&V method in the first column, the randomized M&V method that stops after satisfying all stopping criteria in the second column, and the randomized M&V method continues 50%/50% sampling throughout 2 years in the third column. In addition, subplot a) shows the savings estimated from measured weather condition and subplot b) shows the savings normalized on typical meteorological weather file after using TOWT model. As shown clearly in the plots, the conventional M&V method exhibits an average deviation of 6% in savings estimation, whereas the proposed randomized method demonstrates significantly smaller deviations. If the analyst stops immediately after satisfying all stopping criteria, the deviation is reduced to approximately 2%. Extending the M&V period to match the conventional method further improves accuracy to 1%. Furthermore, by comparing subplot (a) and (b), we observe a minor improvement (less than 1%) in estimation accuracy due to model fitting. This is because the weather condition over the two years for all selected locations are similar. Thus the effect of adjusting temperature as an independent variable in the model is not significant. Overall the results suggest that projecting the baseline into the post-retrofit period that only considers weather variation is inefficient. Subplot (c) presents the overall timeline comparison of the two M&V methods. The results indicate that a significant portion achieve accurate M&V results within 24 weeks and all buildings satisfy the stopping criteria within 36 weeks, mostly influenced by weather variability. Meanwhile, the red arrow highlights the excess time required by the conventional M&V method, which only leads to less reliable results.

We also plotted the mean error in Figure 3.2 and we noticed that the uncertainty range associated with the conventional method is significantly larger compared to the randomized method that only takes 36 weeks according to subplot c) in the Figure 3.1.

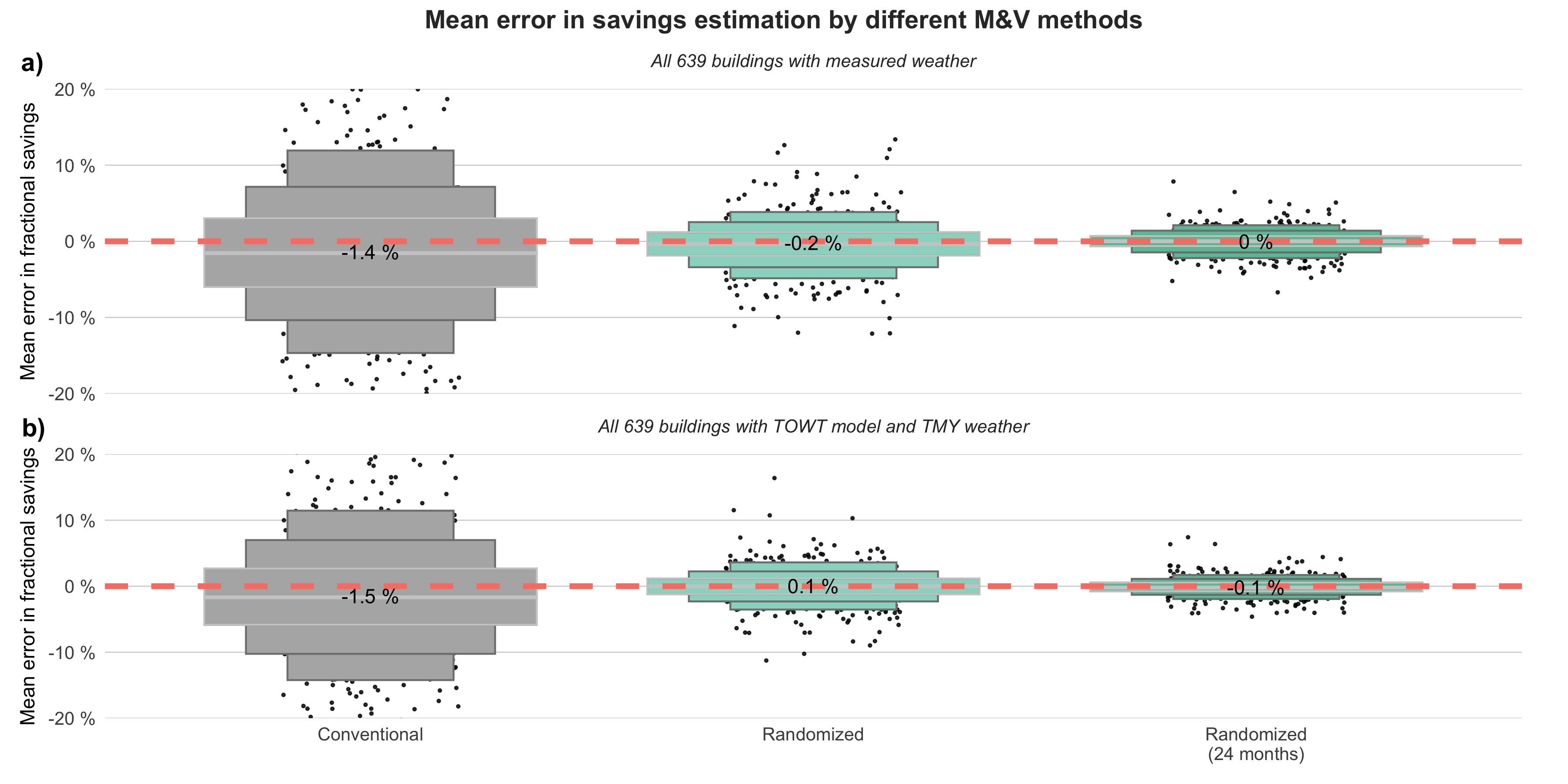


Figure 3.2: Distribution of mean deviation from the ground-truth savings calculated by the two M&V methods

Additionally in the literature we found another M&V method study using over 500+ commercial buildings, but their study scope was to compare a variety of M&V baseline models from simple week mean to more complex machine learning methods using the normalized mean bias error (NMBE)(Granderson et al. 2016). Since we also calculated the normalized savings estimation error, we compared our normalized accuracy of 600+ buildings with TOWT model and TMY weather with their TOWT model prediction assessment in the following table. As each percentile is closely matched, our results are verified.

Table 1. Savings accuracy comparison with a similar literature

|  | Reference study results (Granderson et al. 2016) | Our study results |
| --- | --- | --- |
| 25th percentile | -5.85 | -5.84 |
| 50th percentile | -1.25 | -1.68 |
| 75th percentile | 3.86 | 2.67 |

## 3.1 Non-routine events impact

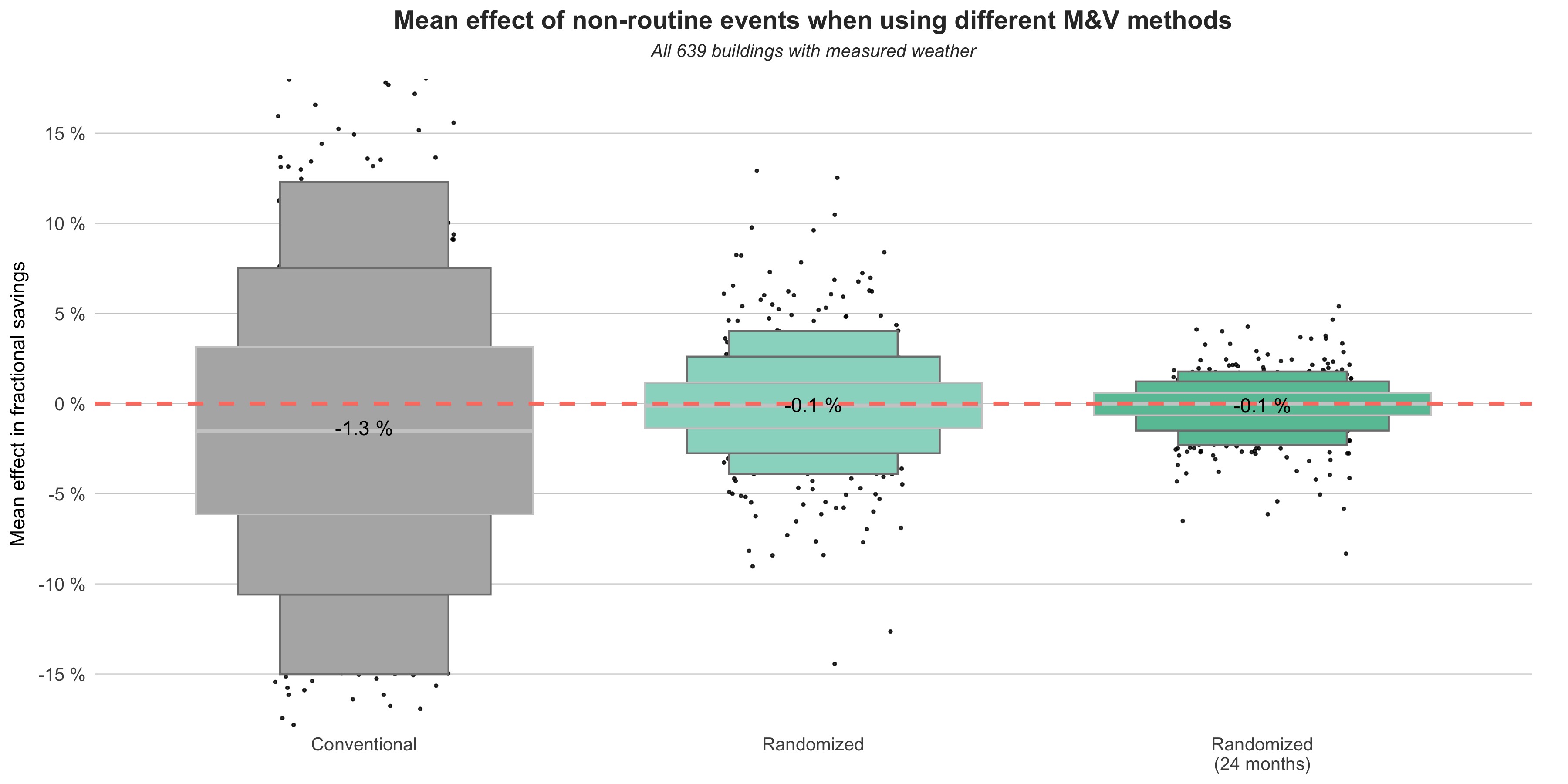


Figure 3.3: Comparison between the two M&V methods in detecting no intervention effect

As mentioned in Section 2.1, the majority of the selected buildings are affected by non-routine events, which is the reason why we observe baseline electricity usage changes throughout the two year. To evaluate the reliability of the two M&V methods in blocking the effect of non-routine events, we repeated the M&V run on the original dataset prior to the application of the reset control. In other words, as there is no intervention applied or the intervention is known to have no impact on whole-building electricity usage, a more reliable M&V method should detect closer to 0 kW savings using the measurements. Therefore, the here are calculated as:

Figure 3.3 shows the randomized M&V method (both at the time of satisfying all stopping criteria and at the end of two-year period) produced consistent results. We included the absolute deviation () in the appendix. Similarly, stopping early can lead to slightly larger uncertainty similar to Figure 3.1. However, as for conventional method, the uncertainty range is much larger than the randomized method and it shows on average the deviation is 1.3% meaning a building analyst can ‘detect’ 1.3% energy savings when there isn’t any using the conventional M&V method.

# 4 Discussion

## 4.1 TOWT modeling accuracy

To evaluate modeling accuracy, we used the Coefficient of Variation of Root-Mean Squared Error (CV(RMSE)) as the primary error metric. Since CV(RMSE) is a normalized measure, it enables direct comparison across different model fitting results. According to ASHRAE Guideline 14, whole-building baseline models should achieve a CV(RMSE) below 25% for computing savings or 30% when less than 12 months’ of measurements are available (ASHRAE 2023). In our study, Figure 4.1 compares the model fitting accuracy of the two M&V methods and it shows around 87.5% of the models fitted in the randomized M&V method are code compliance while 80% for the conventional M&V method. As the quantity of the training set is the same, this highlights the impact of data sampling on model fitting accuracy, and for TOWT, the distribution of outdoor weather conditions sampled in the training dataset is influential. The conventional method trains the model using the entire first-year dataset, if the location is heating or cooling dominant, the data-driven model is trained using colder or warmer outdoor conditions leading to biased model coefficients. Although the random sampling in our study does not specifically block the outdoor temperature factor, the CV(RMSE) distribution shown in the figure indicates a significant improvement in model fitting accuracy.

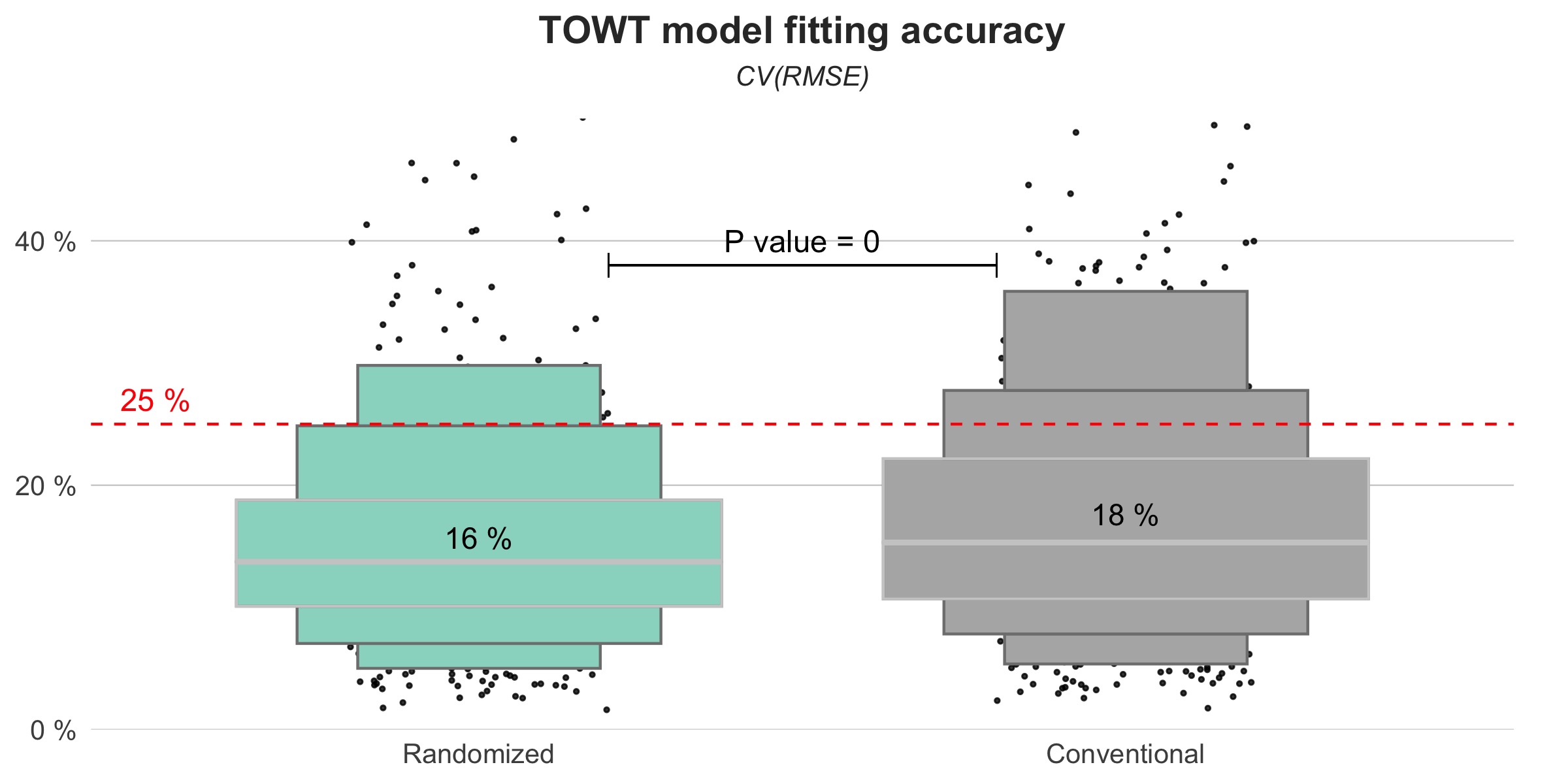


Figure 4.1: TOWT model fitting accuracy distribution for all buildings included (with each data point representing one building)

Additionally, we also noticed in our study that despite the regression models capture well on the mean energy consumption of the building but the tend to underestimate the 15-min daily peak load (Granderson et al. 2021), which is not useful for assessing savings for demand response events.

## 4.2 Sampling interval impact

In certain situations, particularly those involving large thermal mass in active thermal energy storage systems, there is a known time-lagged effect associated with the intervention. This means that the effect of one strategy may persist after switching to another. To mitigate such carryover effects, building analysts may choose to increase the sampling interval and drop non-consecutive days in the switchback experiment. For example, with a hot water tank that is charged by an intervention strategy, residual heat may still be present when the next strategy begins, influencing subsequent measurements, particularly the performance measured on the first day (i.e. a non-consecutive day). In such cases, increasing the sampling interval (e.g., sample strategies every 2 or 3 days instead of daily) and discard measurements from those non-consecutive days can eliminate potential carryover effect. However, this approach reduces the total data collected: for instance, sampling every 3 days would drop 1/6 of measurements and sampling every 2 day means would remove 1/4 of measurements. Therefore, by increasing the sampling interval, there is a reduction in statistical power due to less randomness, and when decreasing the sampling interval, there is a accuracy penalty from dropping more non-consecutive days. To quantify such trade-off, 4.2 shows the accuracy of different sampling intervals and consequence of dropping the non-consecutive days in the dataset over the entire two-year period.

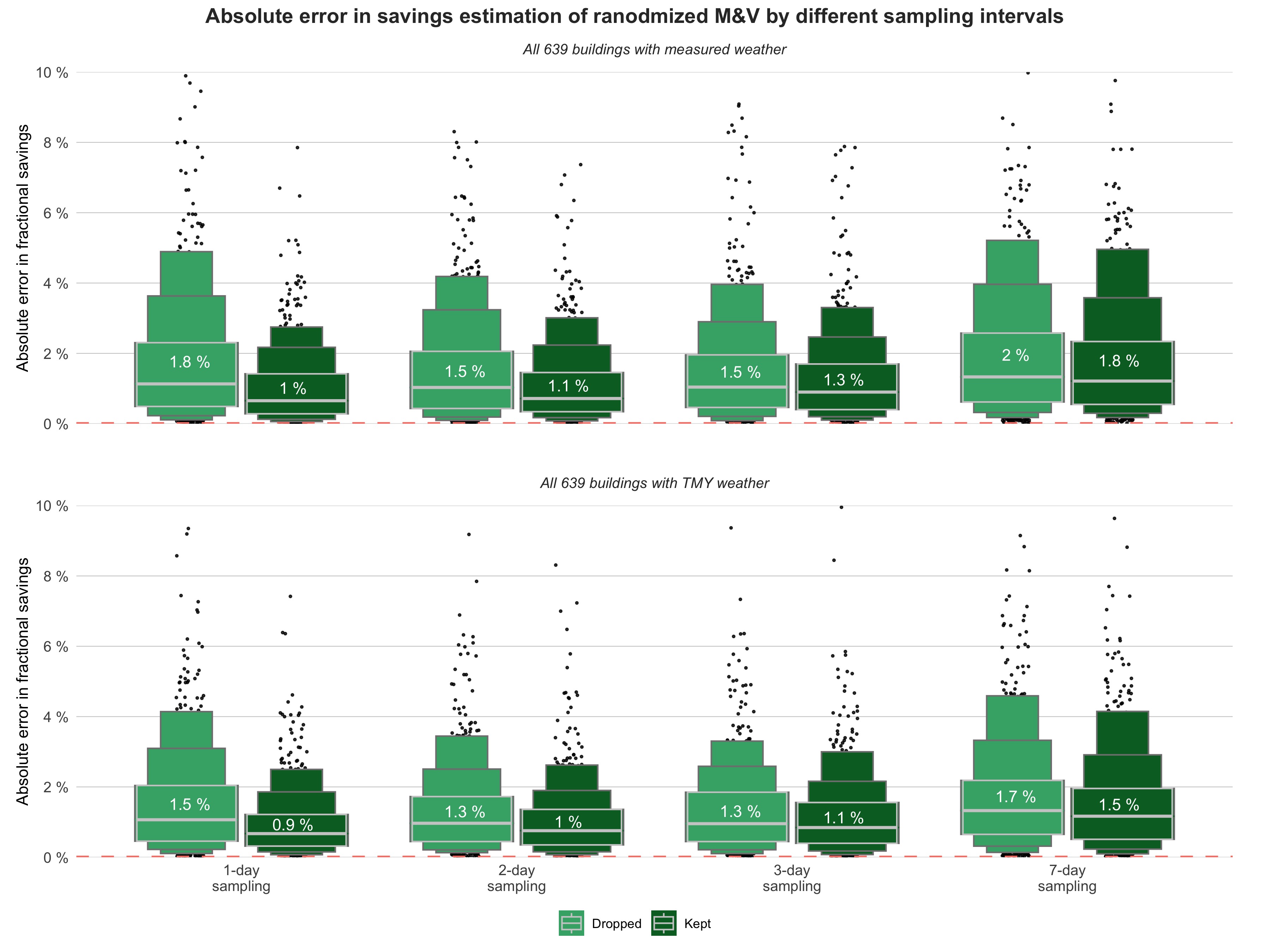


Figure 4.2: Comparison of different sampling interval impact on M&V estimation accuracy (dropped: all non-consecutive days were dropped; kept: all measurements were kept)

To quantify the impact of reduced randomness, all measurements are kept when calculating the distribution shown in the darker colored set. The results suggests a gradual increase in error from 1% to 1.8% when sampling from daily to weekly intervals. To quantify the combined impact of data loss due to preventing carryover effect, we generated the lighter colored set, which drops all non-consecutive days. As a result, we noticed a significant increase in estimaion error when sampling daily, which lead to a similar accuracy compared to weekly intervals. Therefore, considering this trade-off, we recommend using a three-day sampling interval if the carryover effect is likely but expected to last less than one day (as more sampling is required for a two-day swithback), and then normalizing the estimation via energy modeling. We also included the mean deviation distribution and also the accuracy distribution calculated after all stopping criteria satisfied in the supplementary material and the results shows the same indication.

## 4.3 Sampling ratio impact

Another advantage of using randomized M&V is the flexibility of changing sampling ratio after the target savings detected. For instance, the building owner can continue sampling at a 50%/50% ratio between the baseline and the intervention to further reduce the uncertainty associated with the savings. Alternatively, they could switch to 100% intervention to maximize cost savings, though this approach risks the baseline becoming outdated, which is another trade-off to consider. A middle-ground approach would be to sample at an 20%/80% ratio or 10%/90% between the baseline and the intervention. To compare those possible choices to a building owner, figure 4.3 shows the results of changing the original 50%/50% sampling ratio after satisfying all stopping criteria and continuing a new ratio till the end of the two-year period. We also provided the reference case when the building owner continues with no change. The results shows as the sampled baseline days and intervention days become unbalanced, the savings estimation error grows from 1% to 5%. This is reasonable since sampling at 10%/90% is similar to the conventional M&V method. We included the mean deviation version in the supplementary material.

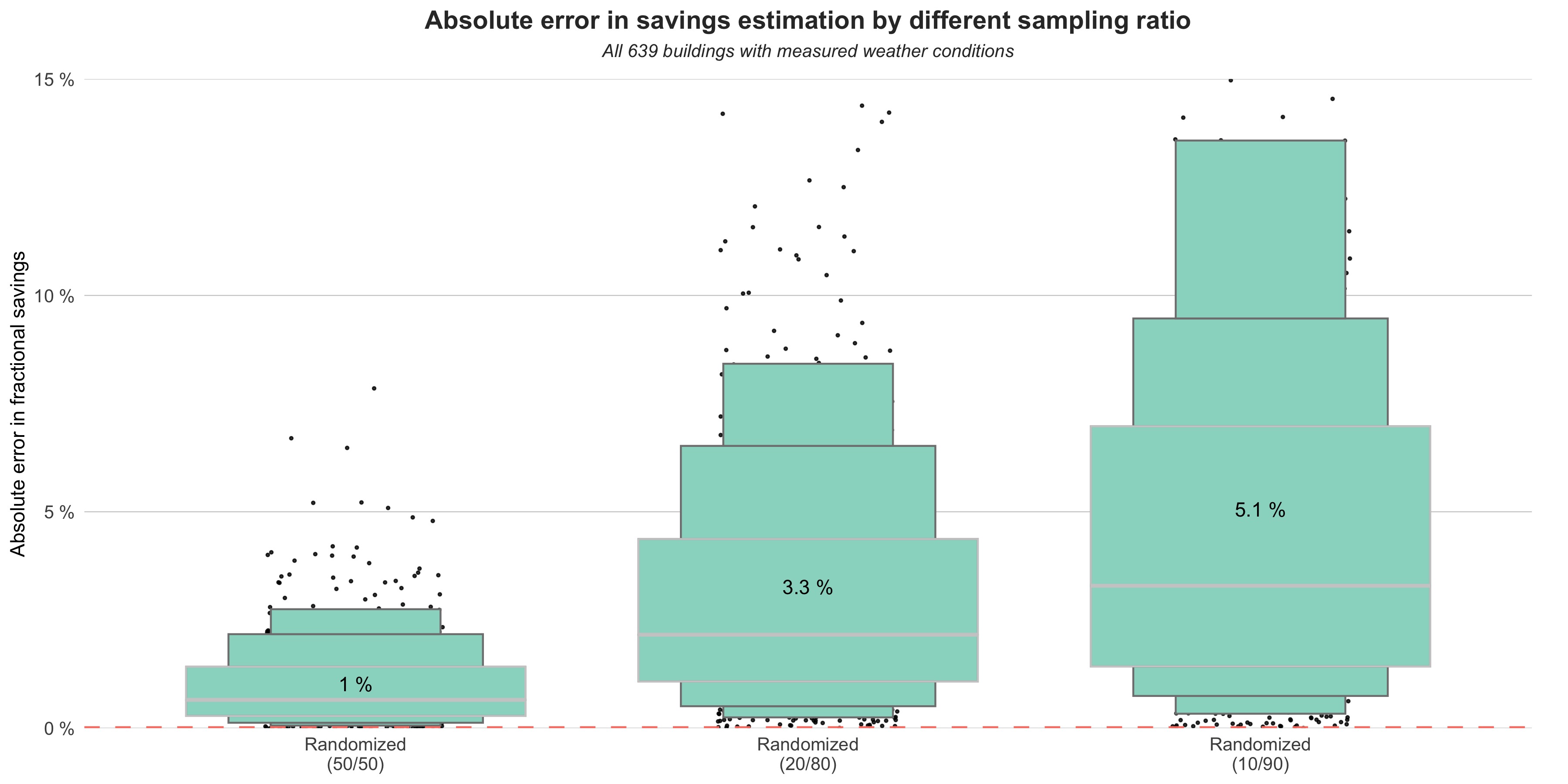


Figure 4.3: Comparison of different sampling ratio impact on M&V estimation accuracy

## 4.4 Limitations

We identify two key limitations in this study:

1. The simulated intervention remains somewhat generic and simplistic considering the diverse building types and climate zones in the BDG2 dataset. As the primary focus is to accurately detect the intervention (or any intervention) effect so for simplicity, we applied the parameters listed in Table 1 uniformly across all buildings. Yet, in some cases, raising the water temperature by 1°C might yield more or less than 8% in electricity savings due to target buildings’ demand flexibility. For example, one field validation of HVAC optimization shows additional savings might be limited if other best practice controls are already implemented (Granderson et al. 2018).
2. For the main analysis, we assumed no carryover effect and a daily sampling interval would suffice for most commercial buildings, but exceptions exist. Based on the measurement from the dataset, we have limited knowledge of how builings thermal mass respond to the proposed chilled water reset strategy. However, in the discussion, we addressed such concern by comparing different sampling strategies and the impact of dropping measurements from non-consecutive days.

# 5 Conclusion

This research demonstrated the application of a novel whole-building measurement and verification (M&V) method, comparing its performance to the conventional approach outlined in ASHRAE Guideline 14 using a large, open-source commercial building dataset. The proposed M&V method leverages the randomized experimental design concept from other scientific fields, along with statistical sequential inference techniques, to determine when target savings are detected. We used a virtual control retrofit case—resetting the chilled water setpoint based on outdoor weather conditions—and applied it to over 600 filtered commercial buildings. By comparing the savings estimations of the conventional method with the novel randomized method, we found that the randomized approach provides faster and more robust savings estimations.

Specifically, we showed that throughout 11 different locations assessed in this study, the randomized M&V can provide a saving estimation by 36 weeks (with the majority finishes by 24 weeks) once all stopping criteria satisfied. In contrast, the conventional method requires a full range of baseline and intervention measurements under normal operating conditions, typically taking 6-9 months for each phase. Most importantly, Furthermore, we verified that with reduced M&V timeline the randomized method can estimate savings more accurately by showing that the absolute error is only 1 - 2 % while for the two-year conventional method, the estimation error could be 6% in a typical building. We also evaluated the impact of non-routine events on the proposed M&V method and the results show that baseline changes in the post retrofit period can deviate savings estimated using the conventional method. As a comparison, we found those events have very negligible impact on the savings estimated using the randomized method, demonstrating ideal reliability for realistic use cases.

We also discussed the impact of model fitting accuracy when using the two methods and we found a significant improvement in our method due to random sampling. We also provide detailed assessment on using different sampling interval and sampling ratio, those results provide useful considerations to building analysts when generating randomized schedules for a variety of use cases.

# 6 CRediT authorship contribution statement

**Aoyu Zou**: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Software, Writing - original draft, Writing - review & editing. **Paul Raftery**: Conceptualization, Formal analysis, Methodology, Investigation, Project administration, Supervision, Writing - review & editing. **Stefano Schiavon**: Conceptualization, Methodology, Investigation, Supervision, Writing - review & editing. **Carlos Duarte**: Methodology, Investigation, Supervision, Writing - review & editing. **Gail Brager**: Supervision, Writing - review & editing.

# 7 Reproducibility

Reproducible example with analysis code is available (MIT license) at <https://github.com/CenterForTheBuiltEnvironment/genome_mnv>.

# 8 Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

# 9 Acknowledgements

# References

Administration, US Energy Information. 2012. “Commercial Buildings Energy Consumption Survey (CBECS).” US EIA Washington, DC. <https://www.eia.gov/consumption/commercial/data/2003/>.

Alrobaie, Abdurahman, and Moncef Krarti. 2022. “A Review of Data-Driven Approaches for Measurement and Verification Analysis of Building Energy Retrofits.” *Energies* 15 (21): 7824. <https://doi.org/10.3390/en15217824>.

ASHRAE. 2023. “ASHRAE Guideline 14 - Measurement of Energy, Demand, and Water Savings.”

Beil, I., I. A. Hiskens, and S. Backhaus. 2015. “Round-Trip Efficiency of Fast Demand Response in a Large Commercial Air Conditioner.” *Energy and Buildings* 97 (June): 47–55. <https://doi.org/10.1016/j.enbuild.2015.03.028>.

Blackston, J. Walker, Andrew G. Chapple, James M. McGree, Suzanne McDonald, and Jane Nikles. 2019. “Comparison of Aggregated n-of-1 Trials with Parallel and Crossover Randomized Controlled Trials Using Simulation Studies.” In, 7:137. Mdpi.

Burkhart, Michael C., Yeonsook Heo, and Victor M. Zavala. 2014. “Measurement and Verification of Building Systems Under Uncertain Data: A Gaussian Process Modeling Approach.” *Energy and Buildings* 75 (June): 189–98. <https://doi.org/10.1016/j.enbuild.2014.01.048>.

Čongradac, Velimir, and Filip Kulić. 2012. “Recognition of the Importance of Using Artificial Neural Networks and Genetic Algorithms to Optimize Chiller Operation.” *Energy and Buildings* 47 (April): 651–58. <https://doi.org/10.1016/j.enbuild.2012.01.007>.

Demand Side Analytics. 2022. “Population NMEC Control Group Accuracy Assessment.” CALMAC ID: PGE0476.01. <https://www.calmac.org/publications/PGE0476.01.pdf>.

DOE, US. 2008. “M&v Guidelines: Measurement and Verification for Federal Energy Projects Version 3.0.” US Department Of Energy.

Duan, Naihua, Richard L. Kravitz, and Christopher H. Schmid. 2013. “Single-Patient (n-of-1) Trials: A Pragmatic Clinical Decision Methodology for Patient-Centered Comparative Effectiveness Research.” *Journal of Clinical Epidemiology* 66 (8): S21–28. <https://doi.org/10.1016/j.jclinepi.2013.04.006>.

Duarte, Carlos, Paul Raftery, Anand Prakash, and Therese Peffer. 2023. “Field Demonstration of the Brick Ontology to Scale up the Deployment of ASHRAE Guideline 36 Control Sequences,” 9. <ps://escholarship.org/uc/item/5zt2d66r>.

Efficiency Valuation Organisation. 2012. “International Performance Measurement and Verification Protocol.” Efficiency Valuation Organisation.

Feng, Zhuangbo, and Shi-Jie Cao. 2019. “A Newly Developed Electrostatic Enhanced Pleated Air Filters Towards the Improvement of Energy and Filtration Efficiency.” *Sustainable Cities and Society* 49 (August): 101569. <https://doi.org/10.1016/j.scs.2019.101569>.

Gallagher, Colm V., Kevin Leahy, Peter O’Donovan, Ken Bruton, and Dominic T. J. O’Sullivan. 2018. “Development and Application of a Machine Learning Supported Methodology for Measurement and Verification (m&v) 2.0.” *Energy and Buildings* 167 (May): 8–22. <https://doi.org/10.1016/j.enbuild.2018.02.023>.

Granderson, Jessica, Guanjing Lin, Rupam Singla, Samuel Fernandes, and Samir Touzani. 2018. “Field Evaluation of Performance of HVAC Optimization System in Commercial Buildings.” *Energy and Buildings* 173 (August): 577586. <https://doi.org/10.1016/j.enbuild.2018.05.048>.

Granderson, Jessica, and Phillip N. Price. 2014. “Development and Application of a Statistical Methodology to Evaluate the Predictive Accuracy of Building Energy Baseline Models.” *Energy* 66 (March): 981–90. <https://doi.org/10.1016/j.energy.2014.01.074>.

Granderson, Jessica, Phillip N. Price, David Jump, Nathan Addy, and Michael D. Sohn. 2015. “Automated Measurement and Verification: Performance of Public Domain Whole-Building Electric Baseline Models.” *Applied Energy* 144 (April): 106–13. <https://doi.org/10.1016/j.apenergy.2015.01.026>.

Granderson, Jessica, Mrinalini Sharma, Eliot Crowe, David Jump, Samuel Fernandes, Samir Touzani, and Devan Johnson. 2021. “Assessment of Model-Based Peak Electric Consumption Prediction for Commercial Buildings.” *Energy and Buildings* 245 (August): 111031. <https://doi.org/10.1016/j.enbuild.2021.111031>.

Granderson, Jessica, Samir Touzani, Claudine Custodio, Michael D. Sohn, David Jump, and Samuel Fernandes. 2016. “Accuracy of Automated Measurement and Verification (m&v) Techniques for Energy Savings in Commercial Buildings.” *Applied Energy* 173 (July): 296–308. <https://doi.org/10.1016/j.apenergy.2016.04.049>.

Gupta, Somit, Ronny Kohavi, Diane Tang, Ya Xu, Reid Andersen, Eytan Bakshy, Niall Cardin, et al. 2019. “Top Challenges from the First Practical Online Controlled Experiments Summit.” *ACM SIGKDD Explorations Newsletter* 21 (1): 2035. <https://doi.org/10.1145/3331651.3331655>.

Heo, Y., R. Choudhary, and G. A. Augenbroe. 2012. “Calibration of Building Energy Models for Retrofit Analysis Under Uncertainty.” *Energy and Buildings* 47 (April): 550–60. <https://doi.org/10.1016/j.enbuild.2011.12.029>.

Huang, Sen, Srinivas Katipamula, and Robert Lutes. 2023. “An Experimental Study on Round-Trip Efficiency of a Preheating Control with a Medium Office Building.” *Energy and Buildings* 278 (January): 112622. <https://doi.org/10.1016/j.enbuild.2022.112622>.

Jin, Xinqiao, Zhimin Du, and Xiaokun Xiao. 2007. “Energy Evaluation of Optimal Control Strategies for Central VWV Chiller Systems.” *Applied Thermal Engineering* 27 (5): 934–41. <https://doi.org/10.1016/j.applthermaleng.2006.08.015>.

Keskar, Aditya, David Anderson, Jeremiah X. Johnson, Ian A. Hiskens, and Johanna L. Mathieu. 2020. “Do Commercial Buildings Become Less Efficient When They Provide Grid Ancillary Services?” *Energy Efficiency* 13 (3): 487–501. <https://doi.org/10.1007/s12053-019-09787-x>.

Lawrence Berkeley National Lab. n.d. “Building Performance Database (BPD).” <https://www.energy.gov/eere/buildings/building-performance-database-bpd>.

Lee, Kuei-Peng, and Te-Ang Cheng. 2012. “A Simulation–Optimization Approach for Energy Efficiency of Chilled Water System.” *Energy and Buildings* 54 (November): 290–96. <https://doi.org/10.1016/j.enbuild.2012.06.028>.

Lee, Young Jun, Sung Hyup Hong, Jong Man Lee, Yeo Beom Yoon, Jong Min Choi, and Kwang Ho Lee. 2022. “Chilled Water Temperature Set-Point Reset Based on Outdoor Air Temperature and Its Cooling Energy Performance in an Office Building.” *Journal of Mechanical Science and Technology* 36 (3): 1557–68. <https://doi.org/10.1007/s12206-022-0241-4>.

Mathew, Paul A., Laurel N. Dunn, Michael D. Sohn, Andrea Mercado, Claudine Custudio, and Travis Walter. 2015. “Big-Data for Building Energy Performance: Lessons from Assembling a Very Large National Database of Building Energy Use.” *Applied Energy* 140 (February): 85–93. <https://doi.org/10.1016/j.apenergy.2014.11.042>.

Mathieu, Johanna L., Phillip N. Price, Sila Kiliccote, and Mary Ann Piette. 2011. “Quantifying Changes in Building Electricity Use, with Application to Demand Response.” *IEEE Transactions on Smart Grid* 2 (3): 507–18. <https://doi.org/10.1109/TSG.2011.2145010>.

Mie, Axel, Vlastimil Novak, Mikael Andersson Franko, Susanne Gjedsted Bügel, and Kristian Holst Laursen. 2022. “Fertilizer Type Affects Stable Isotope Ratios of Nitrogen in Human Blood Plasma─results from Two-Year Controlled Agricultural Field Trials and a Randomized Crossover Dietary Intervention Study.” *Journal of Agricultural and Food Chemistry* 70 (11): 3391–99. <https://doi.org/10.1021/acs.jafc.1c04418>.

Miller, Clayton, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W. Hobson, Zixiao Shi, and Forrest Meggers. 2020. “The Building Data Genome Project 2, Energy Meter Data from the ASHRAE Great Energy Predictor III Competition.” *Scientific Data* 7 (1): 368. <https://doi.org/10.1038/s41597-020-00712-x>.

Munro, Evan, Stefan Wager, and Kuang Xu. n.d. “Treatment Effects in Market Equilibrium.” <https://doi.org/10.48550/arXiv.2109.11647>.

Qiu, Shunian, Zhenhai Li, Dalian Fan, Ruikai He, Xinghui Dai, and Zhengwei Li. 2022. “Chilled Water Temperature Resetting Using Model-Free Reinforcement Learning: Engineering Application.” *Energy and Buildings* 255 (January): 111694. <https://doi.org/10.1016/j.enbuild.2021.111694>.

Raftery, Paul, Aoyu Zou, Thomas Parkinson, and Geoff Hancock. 2024. “Reliably Estimating the Impact of an Active Control Strategy in a Building.” *Journal of Building Engineering* 98 (December): 111134. <https://doi.org/10.1016/j.jobe.2024.111134>.

Raseduzzaman, Md., and Erik Steen Jensen. 2017. “Does Intercropping Enhance Yield Stability in Arable Crop Production? A Meta-Analysis.” *European Journal of Agronomy* 91 (November): 25–33. <https://doi.org/10.1016/j.eja.2017.09.009>.

Taylor, Steven T. 2012. “Optimizing Design & Control of Chilled Water Plants: Part 5: Optimized Control Sequences.” *ASHRAE Journal* 54 (6): 56–60, 63–64, 66, 68, 70–72, 74. <https://www.proquest.com/docview/1034970953/abstract/4C29DF2E85B24972PQ/1>.

Walter, Travis, Phillip N. Price, and Michael D. Sohn. 2014. “Uncertainty Estimation Improves Energy Measurement and Verification Procedures.” *Applied Energy* 130 (October): 230–36. <https://doi.org/10.1016/j.apenergy.2014.05.030>.

Zhai, Zhiqiang (John), and Stephen Nathaniel Johnson. 2017. “Full-Scale Laboratory Test on Energy Dependence on Pressure Drops in HVAC Systems.” *Procedia Engineering*, 10th international symposium on heating, ventilation and air conditioning, ISHVAC2017, 19-22 october 2017, jinan, china, 205 (January): 2133–40. <https://doi.org/10.1016/j.proeng.2017.10.138>.

1. Center for the Built Environment, University of California Berkeley, USA [↑](#footnote-ref-20)