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 comparing pre- and post-retrofit performance to estimate energy savings. They are often time-consuming and unreliable, especially when non-routine events (step changes or more gradual changes in how a building operates) 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 the conventional M&V method, the results can be confounded. In this study, we demonstrated that switchable interventions, such as most control retrofits, can benefit from a new M&V method that randomly samples whether to implement the baseline or the intervention strategy at a fixed interval (e.g., daily). We tested this novel randomized M&V method on a large public dataset (639 buildings) covering various climate zones and commercial building types, using a virtual chilled water supply temperature reset based on outdoor weather as the intervention. The results show that compared to the conventional method, the randomized method provides more accurate savings estimations with a meidum of 74% accuracy improvement and is even faster. Additionally, we found that when non-routine events are present, the randomized method estimates savings that are much closer to the ground-truth savings than the conventional method, demonstrating much improved reliability. We also assessed the impact of normalizing for different weather, starting the M&V at different time of the year, continuing randomization with a different sampling ratio after satisfying all criteria, dropping samples affected by carryover when switching between strategies, and identified the optimal sampling interval for that scenario using this large dataset.

# 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. Specifically, an M&V analyst should account for two types of baseline changes: (1) static changes, which are known building operational changes: such as renovation, equipment addition or removal; and (2) gradual changes, which often occur subtly and unknown, such as the incremental adoption of LED lighting or more efficient plug loads. Identifying and quantifying these changes are particularly labor-intensive and challenging. Consequently, those limitations reduce the feasibility of M&V and complicate 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 building baseline changes, we proposed a M&V method that adopts the randomized crossover design (Raftery et al. 2024), a gold standard from medical, agricultural and online controlled experimental (i.e. A/B testing) studies (Blackston et al. 2019; Duan, Kravitz, and Schmid 2013; Gupta et al. 2019; Mie et al. 2022; Munro, Wager, and Xu 2023; Raseduzzaman and Jensen 2017). An early application of this randomized design to the field of M&V in buildings was first used in (Raftery et al. 2018), which compared the energy performance of two different supply air temperature reset strategies in a building. Subsequent improvements to the M&V method include adding a sequential evaluation framework and defining stopping criteria to end the M&V period early after reaching a target level of uncertainty, among other improvements, such as using blocked randomization by weekday. The full framework is detailed in a previously published study (Raftery et al. 2024) along with a case study application to a real building, and other example use cases. 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 one intervention, the balanced randomized schedule would equally sample five Mondays with the baseline strategy in operation and five Mondays with the intervention strategy in operation. The major 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, which is a small subset of all the interventions performed in buildings. However, for all applicable use cases, it allows analysts to detect energy savings sequentially shortly after the test begins. In addition, once the desired estimate uncertainty target is achieved, analysts can terminate the M&V study 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 (a common non-routine event) and other more subtle long term changes in building energy use are likely to be evenly distributed among measurements, leading to a more accurate and unbiased assessment of the intervention effect.

## 1.3 BDG2 dataset

To demonstrate and differences between the two M&V methods, the Building Genome Dataset 2 (BGD2) is used, which 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 used the metered electricity and outdoor weather measurements for evaluating the M&V methods.

## 1.4 Literature review

### 1.4.1 Whole building approach

ASHRAE Guideline 14 and the IPMVP provide options for whole-building M&V use cases. For code compliance using the prescriptive option, 12 months of baseline and 12 months of post-retrofit measurements are required. In addition, the expected savings should be larger than 10% and the baseline energy model fitting accuracy should have CV(RMSE) lower than 25% for 12-month data collection. Thus, most research related to M&V for the 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 the baseline model might deteriorate (i.e., become 'stale') over the extended period of pre- and post-analysis.

### 1.4.2 Changes in building baseline measurements

Changes in a building that influence its energy usage and are unrelated to the intervention being studied by the M&V method are commonly referred to as non-routine events. Current guidelines define non-routine events as 'static factors' that need to be adjusted after projecting the baseline in the post-retrofit period (ASHRAE 2023; Efficiency Valuation Organisation 2012). 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 provide a general approach for consideration. For example, IPMVP requires the facility owner and the M&V analyst to periodically perform inspections of all equipment and operations during the reporting period, which is labor-intensive and error-prone. ASHRAE Guideline 14 recommends performing engineering calculations or computer software simulations to adjust the post-retrofit baseline. Additionally, it is relatively rare to have access to all measurements needed for adjustments, such as occupancy, and thus the analyst normally assumes those factors remain unchanged throughout the study. For example, some studies use linear interpolation to estimate counter-factual baseline for the demand response program (Beil, Hiskens, and Backhaus 2015; Keskar et al. 2020). Consequently, one study points out that it is inaccurate to assume no change in the operating conditions during the response period (Huang, Katipamula, and Lutes 2023). Other studies realized such limitation in field study, and emphasized the importance of requiring matched groups to control for exogenous factors beyond weather differences when comparing between baseline and intervention (Demand Side Analytics 2022; Huang, Katipamula, and Lutes 2023).

Overall, those non-routine events mostly refer to clearly observable changes, but there are also more gradual and subtle changes in buildings, which are hardly noticed. We define those gradual changes as 'operational drift' in this study. One likely encountered operational drift situation is the filter clogging in air handling units due to particle accumulation. This causes supply fans to gradually consume more energy to maintain the 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 overestimated as the decreased energy use is incorrectly attributed to the intervention rather than the filter replacement. An operational drift has not been formally defined in the literature or standards, but in this study, we consider it as a special type of non-routine event, as it also represents a change in facility operation that occurs gradually over a longer period of time.

## 1.5 Objectives

As mentioned, the goal of an M&V project is to determine the effect, typically energy savings, of some intervention in the building. In this study, we focus on switchable interventions and an example of such an intervention is a control strategy 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 made it even simpler by adjusting the setpoint based on outdoor weather conditions (see description in Section 2.2). By conducting such an analysis, we aim to:

1. Determine, using a large sample of buildings, how much more accurately the randomized M&V method would estimate the savings of the proposed intervention compared to the conventional M&V, and how much more quickly it would reach a result.
2. Determine how much more robust the randomized method is compared to the conventional method when static non-routine events and operational drift are present in the dataset. The BDG2 dataset contains real-world building energy usage data collected over two years, which inherently contains varying degrees of usage changes. By virtually introducing intervention effects into the existing dataset, we can calculate the ground-truth energy savings, which serve as a reference point for method comparison.
3. Open-source implementation of the proposed randomized M&V method. 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 code 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.

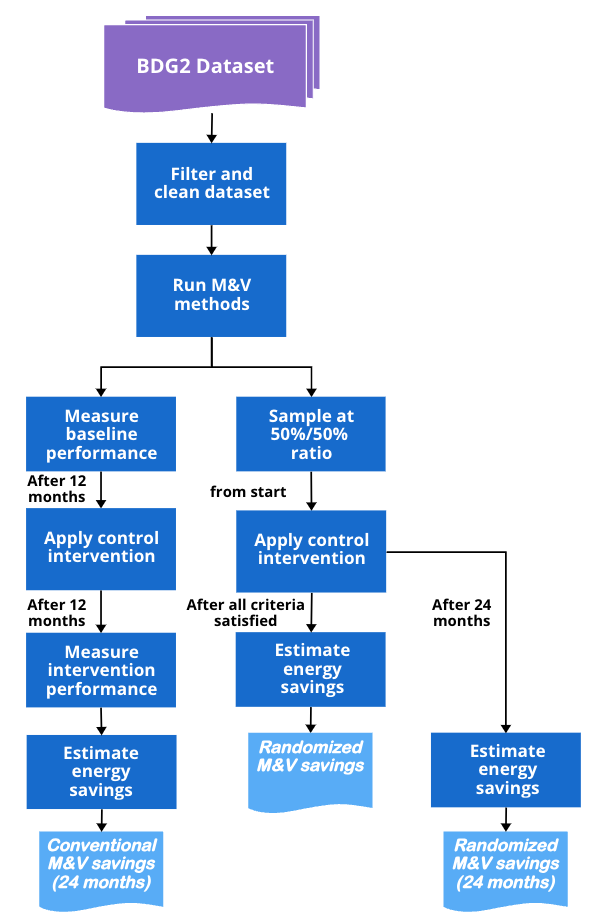


Figure 2.1: Flow chart showing the methodology for comparing the estimated savings of randomized M&V with the conventional M&V

## 2.1 Filter and clean dataset

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 in 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% of electrical energy usage intensity according to the statistics provided by the Building Performance Database (Lawrence Berkeley National Laboratory 2015; 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.

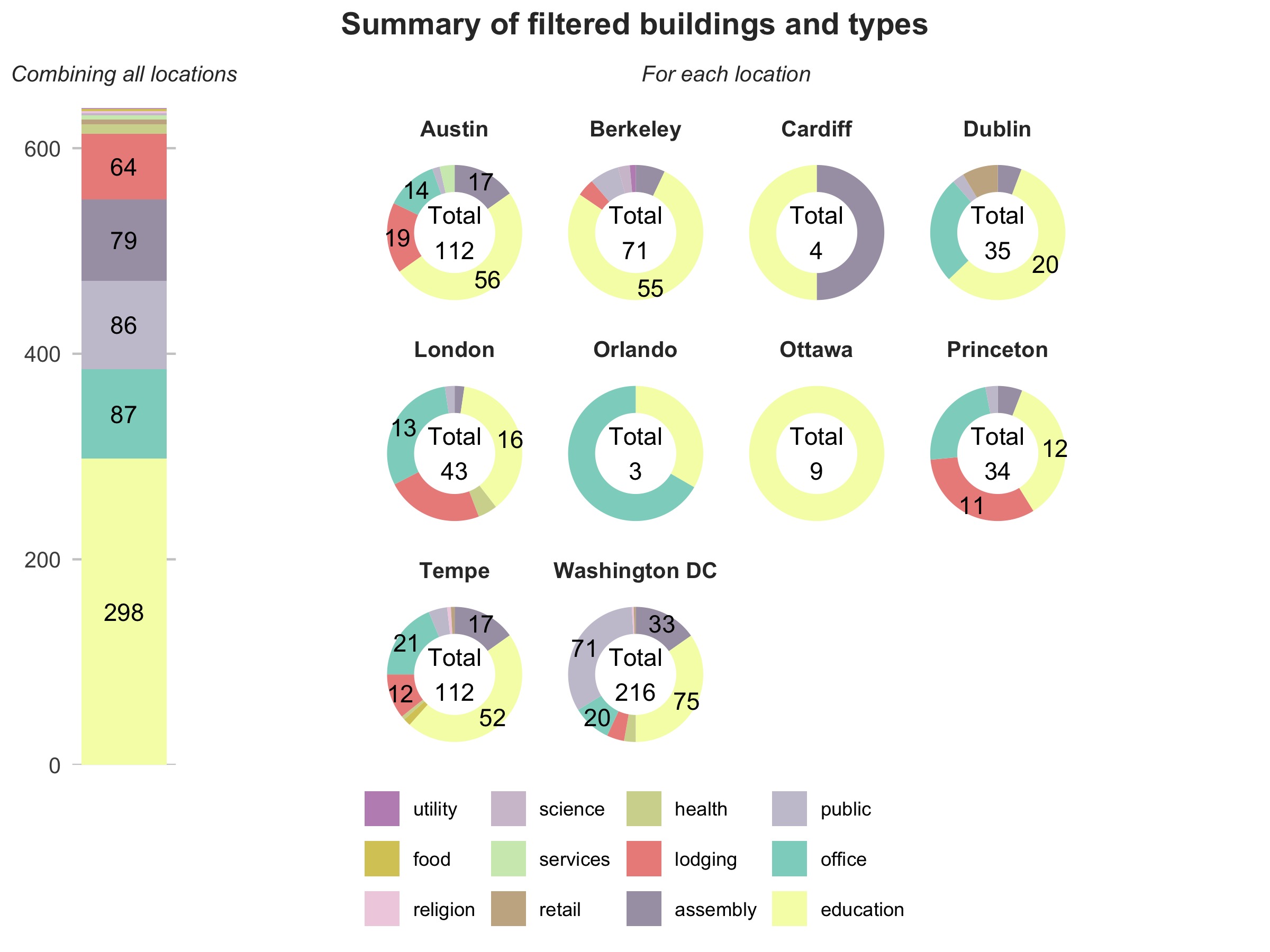


Figure 2.2: Site summary of the filtered buildings from BDG2 dataset (counts < 10 are omitted for visualization; left: aggregated counts of buildings for each type; right: breakdown building counts for each building type at each location)

For all qualified buildings, we also create a feature to identify buildings with very stable electricity usage overall by identifying those with no statistically significant difference (P-value > 0.05) between the two-year electricity usage. Applying this filter to real buildings, only 66 buildings are labeled in this 'stable' subset out of a total of more than 600 buildings. While we did not adjust for weather differences between the two years when generating this high-level feature, this result highlights that the assumption that buildings are typically stable over the timeframe involved in conventional M&V methods is rarely true. As mentioned in 3.1, such variability is largely associated with non-routine events, especially operational drifts since the assessment, even under ideal situations, lasts for two years. And according to the statistics shown here, it is a typical case observed in real buildings applications.

## 2.2 Apply virtue intervention

Figure 2.3 shows the algorithm for the proposed control intervention that resets 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 operates 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.

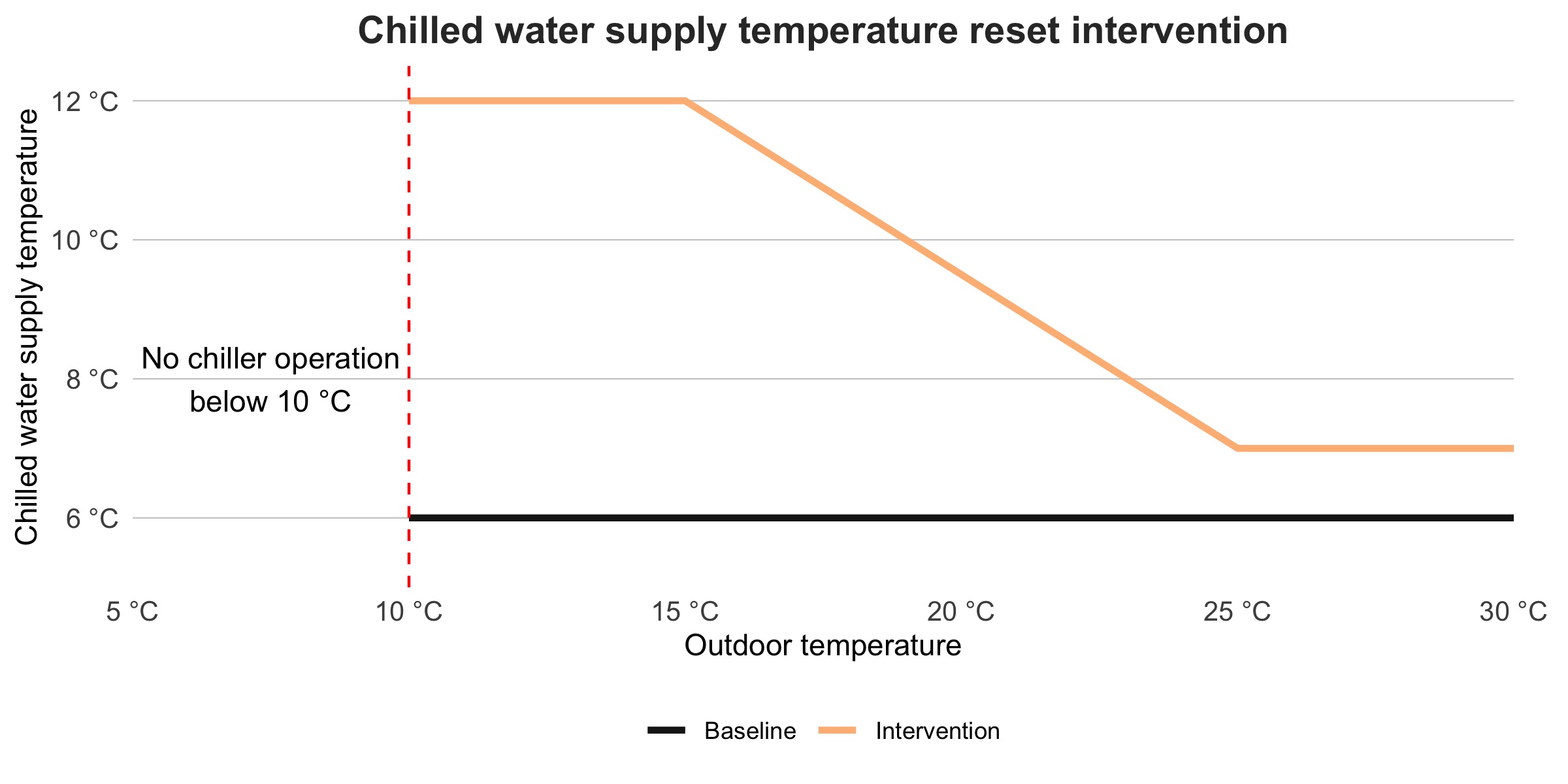


Figure 2.3: Proposed intervention strategy: chilled water supply temperature reset based on outdoor temperature.

We used a simple relationship to map the chilled water supply temperature reset to the electrical energy savings:

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 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 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 the stopping criteria enable an early stop at the end of the satisfied blocking period.
* At least 80% of the dry-bulb temperature range in the annual TMY data is sampled by both strategies.
* 90% confidence that energy savings exceed or do not exceed 0% using the sequential probability ratio test (SPRT). Medium effect size (d = 0.5) quantified by Cohen's d and calculated SPRT statistics either fall below the lower threshold or exceed the upper threshold.
* Sample strategies at 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 representing 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 for reference.

Figure 3.1 shows the overall results of M&V methods comparison and for clarity, we plotted conventional M&V results in grey, randomized M&V results after satisfying all stopping criteria in green, and the randomized M&V results sampled for the entire 24 months in blue. Subplots (a) and (b) calculate the savings estimation error as the absolute deviation from the true savings (i.e. ) and show the results distribution using the 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 conditions, and subplot b) shows the savings normalized on the typical meteorological weather of the building location after using a time-of-week-temperature (TOWT) model (Mathieu et al. 2011). As a result, the conventional M&V method exhibits a median deviation of 5% in savings estimation, whereas the randomized method demonstrates significantly smaller deviations. If the analyst stops immediately after satisfying all stopping criteria, the deviation is reduced to approximately 1.5%. Extending the M&V period to match the same time period of the conventional method (24 months) further improves accuracy to 0.7% deviation. Furthermore, by comparing subplots (a) and (b), we observed a minor improvement in estimation accuracy when a model is fitted to the data to account for weather differences. The model fitting is most beneficial for the shorter dataset (randomized and stopping when all criteria are met, typically 24-36 weeks of data), where it reduces deviation from 1.5% to 1.2% (around 20% improvement). This is because the weather conditions between the first and second year for each of the locations are relatively similar on average over this time period, and thus the effect of adjusting for temperature as an independent variable in the model is relatively small. We found 83% of all filtered buildings show less error when using the randomized M&V, and by comparing the error distribution in subplot (b), we noticed the medium deviation drops from 4.6% to 1.2% (around 74% improvement) at the time of satisfying all stopping criteria. Subplot (c) compares the overall timeline of the two M&V methods. The results show that almost all buildings achieve accurate M&V results and meet the stopping criteria within 24 weeks (for this intervention), with the remainder meeting all stopping criteria within 36 weeks. The additional 12 weeks are usually driven by a need to span sufficient weather conditions representative of a full year of data. Meanwhile, the red arrow highlights the excess time required by the conventional M&V method (24 months). Thus, the randomized M&V method finds more accurate results in approximately a quarter of the time required for conventional M&V. Even if it was the case that an existing 12-month baseline data was available at the start of the M&V process, randomized M&V would still obtain a more accurate result, in less time, than obtaining the 12-month retrofit dataset required by conventional M&V.

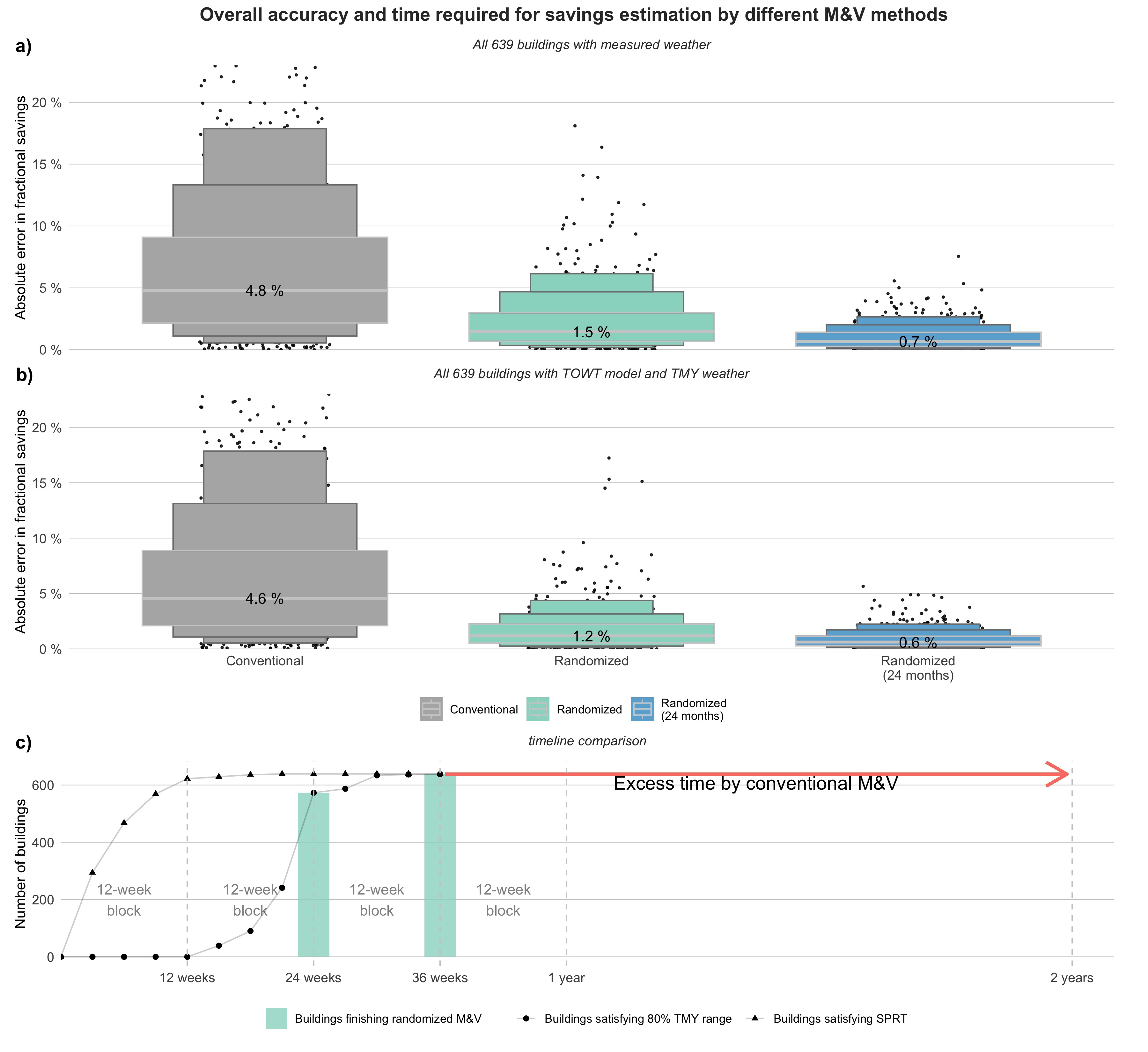


Figure 3.1: Overall comparison results between the conventional M&V method and the proposed randomized M&V method (both upon reaching the stopping criteria for each case, and over a fixed two-year period).

Figure 3.2 shows the biased (instead of absolute) estimated deviation, highlighting that the uncertainty range associated with the conventional method is significantly larger compared to even the randomized method that only takes 36 weeks, according to subplot (c) in Figure 3.1.

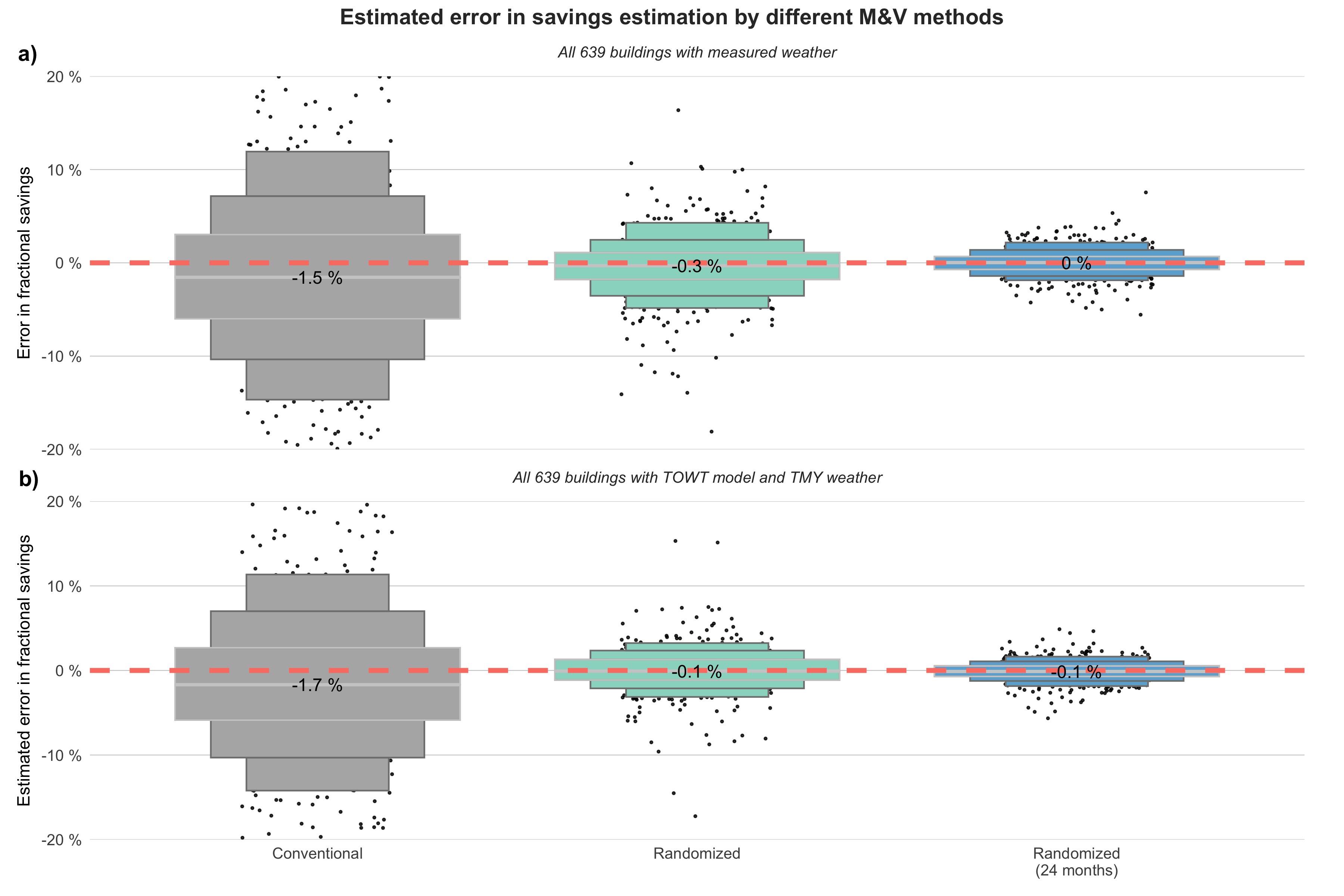


Figure 3.2: Distribution of estimated deviation from the ground-truth savings calculated by the two M&V methods (red dashed line highlighting 0% error)

Additionally in the literature, we found another M&V study using over 500+ commercial buildings (on a different open-source dataset), but their study scope was to compare the predictive performance of a variety of M&V baseline models from simple weekly 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 for 600+ buildings with the TOWT model and TMY weather with their TOWT model prediction assessment in the following table, highlighting that the model and overall building datasets are performing similarly.

Table 2. Savings accuracy comparison with a similar literature

|  | Reference study results (Granderson et al. 2016) | Our study results |
| --- | --- | --- |
| 25th percentile | -5.85 | -5.87 |
| 50th percentile | -1.25 | -1.68 |
| 75th percentile | 3.86 | 2.68 |

## 3.1 Non-routine events impact

As mentioned in Section 2.1, the majority of the selected buildings are affected by operational drift, which is the reason why we observe baseline electricity usage changes throughout the two-year period. To evaluate the robustness of the two M&V methods to the effect of operational drift and non-routine events, we repeated the M&V run on the original dataset prior to the application of the reset control. In other words, there *is no intervention applied here*, so the most reliable M&V method will be the one that detects the closest to 0 kW savings using the measurements. Therefore, the here are calculated as:

Figure 3.3 shows that the randomized M&V method (both at the time of satisfying all stopping criteria and at the end of a two-year period) produced consistent results. We included the absolute deviation () figure in the supplementary material. Similarly to Figure 3.1, stopping when meeting all stopping criteria generally yields slightly larger uncertainty than continuing for 24 months. However, it still yields a more accurate result than the conventional method. Interestingly, similar to (Granderson et al. 2016) where they obtained a median normalized deviation around -1.25%, our analysis on the conventional method shows that the typical deviation is -1.5%, meaning that the typical building is decreasing energy use slightly in this dataset in the absence of any known intervention. Thus, if a building analyst were to apply an intervention to this building and use conventional M&V to measure the effect, they would typically 'detect' 1.5% energy savings even when the intervention had no actual impact on energy performance.

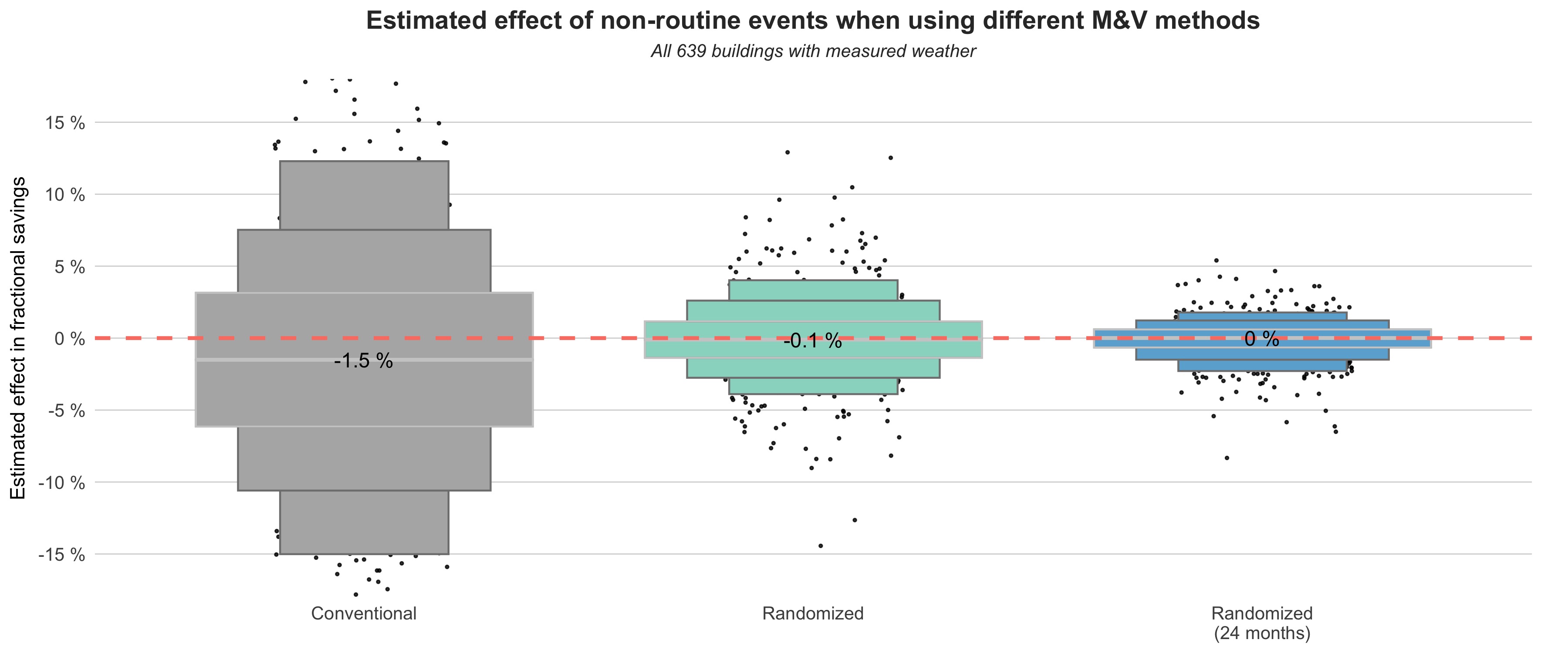


Figure 3.3: Comparison between the two M&V methods in detecting no intervention effect when buildings are subject to static non-rountine events and gradual operational drift.

# 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. CV(RMSE) is a normalized measure, thus it enables direct comparison across different modeling results. Since we have the ground-true baseline measurements in the post-retrofit period, we can assess the error caused by model adjustment (i.e., how accurate is the 'counterfactual' baseline provided by the two M&V methods). Figure 4.1 shows the difference between the true post-retrofit baseline and adjusted baseline through TOWT modeling and since the quantity of the training set is the same, this highlights the impact of the data sampling technique. For example, if a non-routine event happens (such as a tenant occupying an entire floor of a building moving out), the TOWT model can yield significant deviations. This is particularly the case for the conventional M&V method since it samples continuously throughout the pre-retrofit period. However, randomized sampling is less impacted by these events since it samples only 50% of the pre-retrofit baseline throughout a blocking period. Consequently, the figure shows, there is a noticeable improvement in modeling accuracy when using the randomized method.

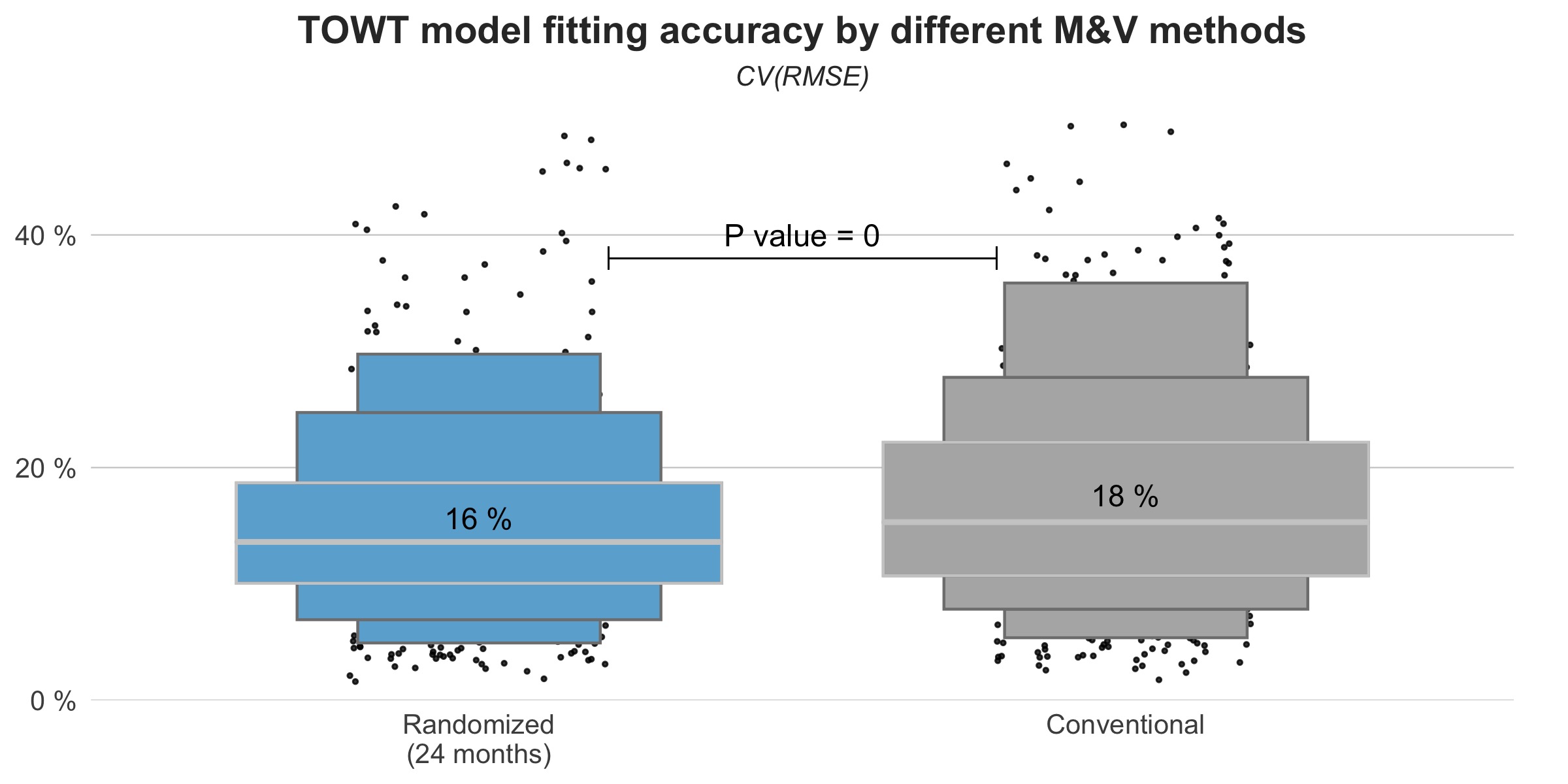


Figure 4.1: TOWT modeling accuracy distribution for all buildings included (with each data point representing one building)

## 4.2 Sampling interval impact

In some cases, particularly those involving large thermal mass and active thermal energy storage systems or operating continuously, there can be a substantial time-lagged effect associated with interventions. This means that the effect of one control strategy may persist, or 'carryover', after switching to another. To mitigate such carryover effects, building analysts may choose to drop non-consecutive days in the switchback experiment. For example, consider a heat pump water heater where the tank is charged at very different times of the day by the intervention strategy compared to the baseline strategy. When switching strategy, the tank may be much warmer or cooler than typical for that strategy, thus 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 discarding measurements from those non-consecutive days can eliminate carryover. However, this approach reduces the total data collected: for instance, sampling every 3 days and dropping days with non-consecutive control strategies would drop 1/6 of measurements on average. Sampling every 2 days means dropping a quarter of the measurements. Therefore, by increasing the sampling interval, there is an accuracy penalty due to a reduction in randomization (how often samples occur), and when decreasing the sampling interval, there is an accuracy penalty from dropping more non-consecutive days. To quantify such trade-off, Figure 4.2 shows the accuracy of different sampling intervals and the consequence of dropping the non-consecutive days in the dataset at the time when randomized M&V produces a result (after satisfying all stopping criteria).

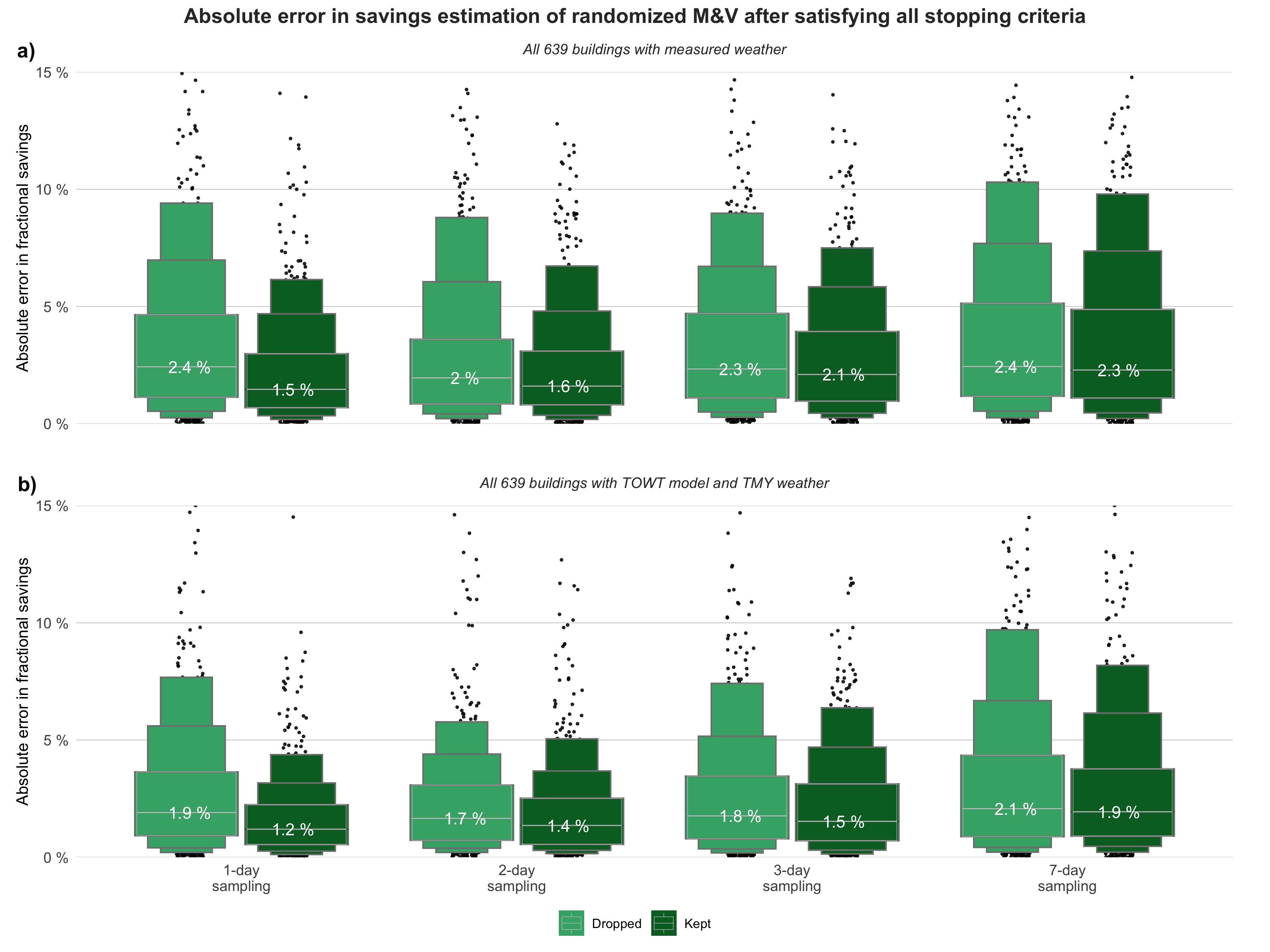


Figure 4.2: Comparison of different sampling interval impact on M&V estimation accuracy (dropped: all non-consecutive days were dropped; kept: all measurements were kept)

To quantify the impact of reduced randomness, all measurements are kept when calculating the distribution shown in the darker-colored set. The results suggest that without dropping non-consecutive days in subplot (a), a gradual increase in error from 1.5% to 2.3% (1.2% to 1.9% with weather normalization in subplot (b) when the sampling interval increases from daily to weekly intervals. This is due to the fact that increasing the sample interval means reducing randomization. To quantify the combined impact of data loss due to preventing the carryover effect, we generated the lighter-colored set, which drops all non-consecutive days. As a result, we noticed an increase in estimation error when sampling daily to 2.4% (to 1.9% when TOWT is fitted), which led to a similar accuracy compared to weekly intervals. Therefore, considering this trade-off, we recommend using a two-day sampling interval (or three-day sampling interval if the building owner prefers less frequent switchover) if the carryover effect is likely but expected to last less than one day, and then normalizing the estimation via energy modeling. The supplementary material contains similar figures showing biased deviation distribution, as well as the same results when continuing the randomized M&V for 24 months. Those results indicate the same conclusion; that sampling every two days and dropping non-consecutive samples yields the optimal result when carryover effects are present.

Sampling at different intervals and excluding non-consecutive days also impact other stopping criteria, such as the time needed to satisfy statistical uncertainty and weather conditions thresholds. In Figure 4.3, we compared the percentage of buildings satisfying all stopping criteria under different sampling intervals, with or without dropping non-consecutive days. The results indicate that randomized M&V concludes more quickly when all data are retained, particularly at shorter intervals, as expected. For instance, when sampling at a one-day interval, 90% of buildings meet all criteria within 24 weeks if no data are dropped, but this number decreases significantly to 52% when non-consecutive days (account for 1/3 of the data) are excluded, and 3% of the buildings require another 12 weeks to satisfy the weather criterion. As the sampling interval increases, this difference gradually diminishes: when sampling at a weekly interval, removing non-consecutive days has no impact on the timeline of meeting stopping criteria.

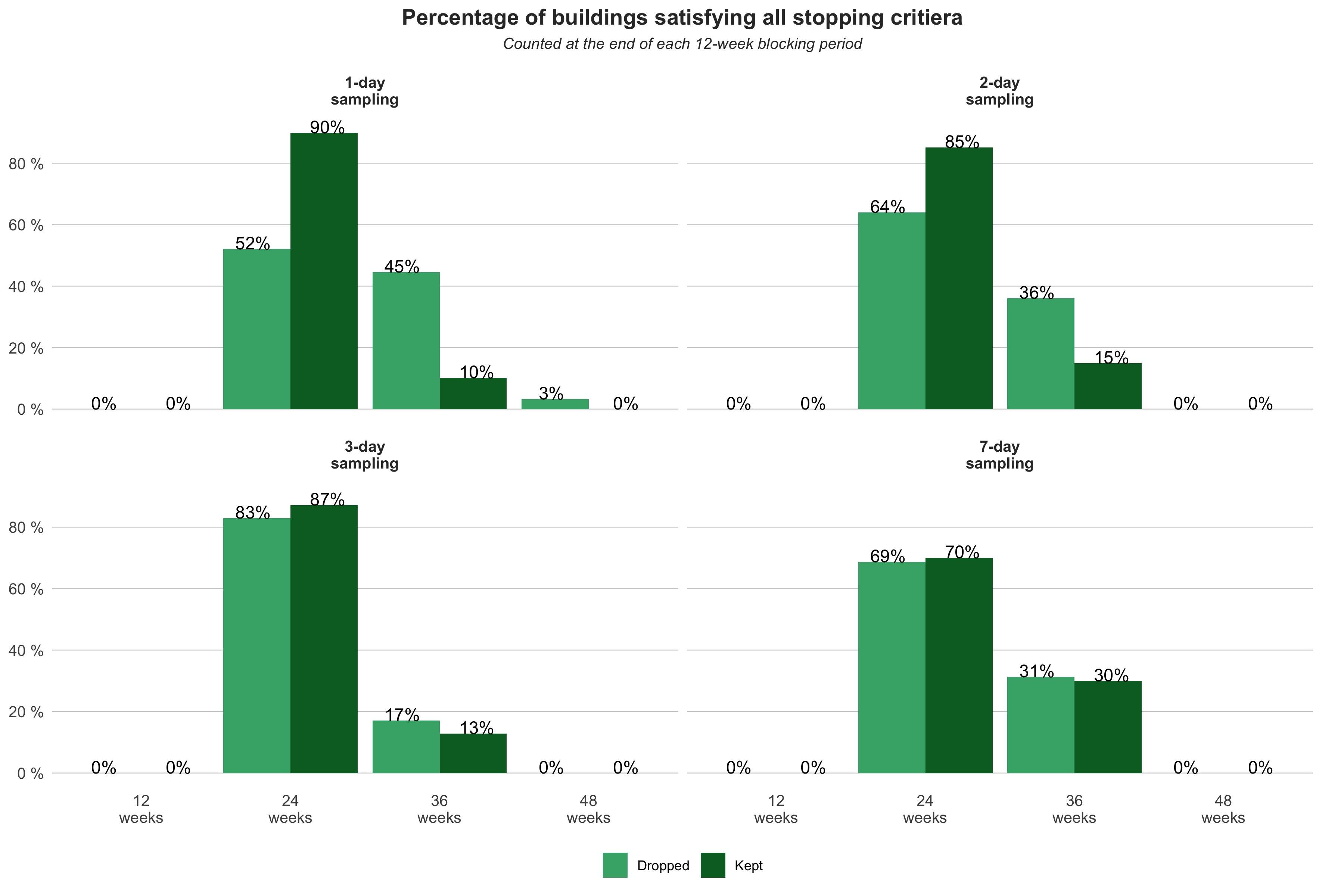


Figure 4.3: Comparison of time required to satisfy all stopping criteria when considering dropping non-consecutive days (or not)

## 4.3 Sampling ratio impact

In our previous study (Raftery et al. 2024), we recommended that building owners adjust the sampling ratio after meeting all stopping criteria, as continuing baseline sampling helps maintain M&V accuracy with limited impact on the energy savings actually achieved by the building. To further examine its impact on M&V accuracy, Figure 4.4 compares two sampling ratios (20%/80% and 10%/90%) and two associated scenarios with the previously shown results when sampling at 50%/50%. The "Continued" group (shown in a lighter color) follows the standard randomized M&V framework with an initial 50%/50% sampling ratio, then switches to the indicated sampling ratio after satisfying the stopping criteria, continuing this until the end of the two-year period. In contrast, the "Randomized" group uses the indicated ratio from the outset and continues throughout the two year. Subplot (a) shows that absolute error is significantly lower without the initial 50%/50% sampling, due to baseline drift and non-routine events that most buildings experience over two years (as noted in Section 2.1). Subplot (b) confirms this finding even after weather normalization.

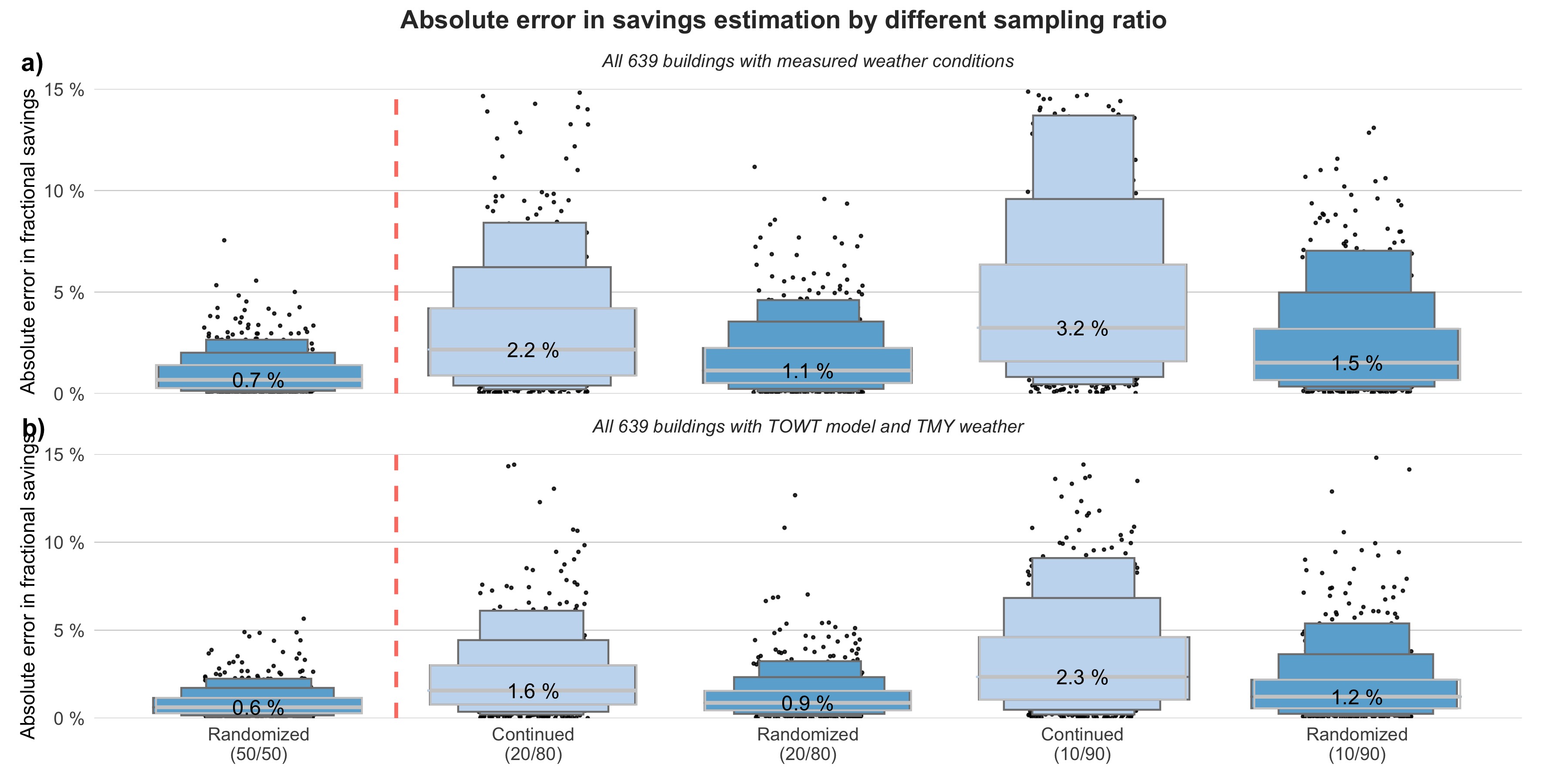


Figure 4.4: Comparison of different sampling ratio impact on M&V estimation accuracy over the entire 24 months (previous results of sampling at 50%/50% shown on the left side of the red dashed line)

We further illustrated this point in the supplementary material where we plotted results using only the ‘stable’ building subset (66 buildings), as defined in Section 2.1. Since they are less influenced by non-routine events, the differences between sampling strategies are much reduced in this subset and can be largely corrected through weather normalization modeling.

As shown in the previous Figure 3.1, the highest accuracy is achieved when sampling at a 50%/50% ratio. Therefore, we continue to recommend that building owners follow the proposed framework and maintain a 50%/50% sampling ratio until all stopping criteria are met. The additional results from Figure 4.4 further add to the guidance: if building baseline changes, retrofit savings estimation should simply discard the existing baseline measurements and revert back to 50%/50% sampling ratio as described in our framework. On the other hand, if the building owner has decided to test for two years and would like to capitalize more savings at the same time, sampling between the baseline and intervention at 10%/90% from start is recommended.

## 4.4 Blocking period impact

Selecting different blocking periods affects how frequently the M&V analyst evaluates stopping criteria and may influence the decision when a savings is detected at a sufficient level of certainty. Here, we summarize the advantages and disadvantages of shorter versus longer blocking periods. As discussed in our previous study, the second step of this randomized M&V method involves choosing an appropriate blocking period to minimize potential biases caused by incomplete randomization. For most practical scenarios, we recommend a blocking period of approximately 12 to 16 weeks, as it typically covers the duration of a season, ensuring a representative sampling of operational conditions. In general, longer blocking periods are advantageous when evaluating multiple intervention strategies or using sampling intervals greater than one day. In such cases, extended blocking ensures a balanced distribution of both day-of-week and outdoor weather conditions.

Opting for a shorter blocking period, such as 6 weeks, has the potential benefit of allowing the stopping criteria to be met earlier. We plotted the percentage of buildings finishing all stopping criteria when the blocking period is set as 6 weeks in Figure 4.5. As a result, when compared to Figure 4.3, the only difference we found is when sampling at a daily interval and not dropping non-consecutive days, 12% of buildings can get an estimate earlier. This means those buildings are likely to meet all stopping criteria between 12 weeks and 18 weeks but when sampled over a 12-week block, the savings can only be determined at week 24. Although sampled baseline and intervention days can be different when the blocking period is changed, the results indicate most buildings still require 9 months to meet outdoor weather conditions when starting M&V from January.

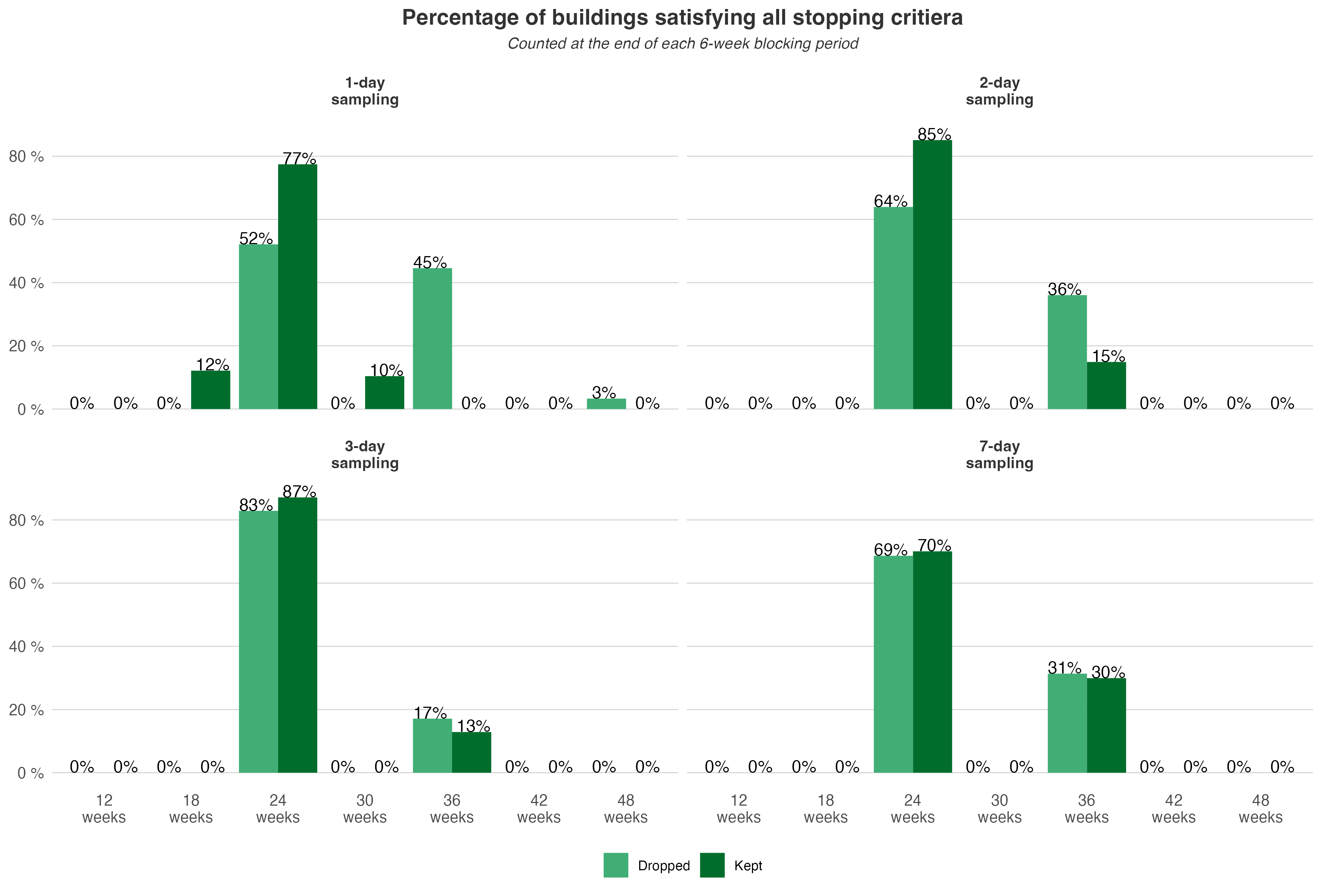


Figure 4.5: Time required to satisfy all stopping criteria when using a 6-week blocking period

Sampling at a shorter blocking period also means that the analyst must perform the hypothesis tests more frequently, which can necessitate adjusting the significance level using methods such as the Bonferroni correction to control for increased type-I error risk. On the plus side, a shorter blocking period can reduce the adverse impact of missing or problematic data. For example, if an intervention requires fine-tuning after initial implementation, removing data from the first blocking period would be less detrimental compared to longer blocking periods, preserving the overall integrity of the experiment.

## 4.5 M&V starting time impact

In this study, we set the baseline period to begin on January 1st, which offers the advantage of capturing colder outdoor weather conditions early in the M&V process. However, when the baseline measurements miss the initial winter period, meeting the required weather condition criteria can potentially take a longer time. To assess this impact, we repeated the randomized M&V analysis for all buildings using a new start date in March. As shown in Figure 4.6, the new M&V process results in a noticeable delay in satisfying the stopping criteria. Compared to Figure 4.3, where a January start allowed most buildings to finish within 36 weeks, a March start extended the required duration to 48 weeks for the majority of buildings. Furthermore, when a longer sampling interval is used or non-consecutive days are excluded (to avoid carryover effects), some buildings may require up to 60 weeks to complete the randomized M&V. As expected, to collect sufficient data representing winter conditions, the sampling may need to continue until the next January (around week 48). However, as the overall data collection period extends, the influence of dropping non-consecutive days diminishes and becomes less of a concern.

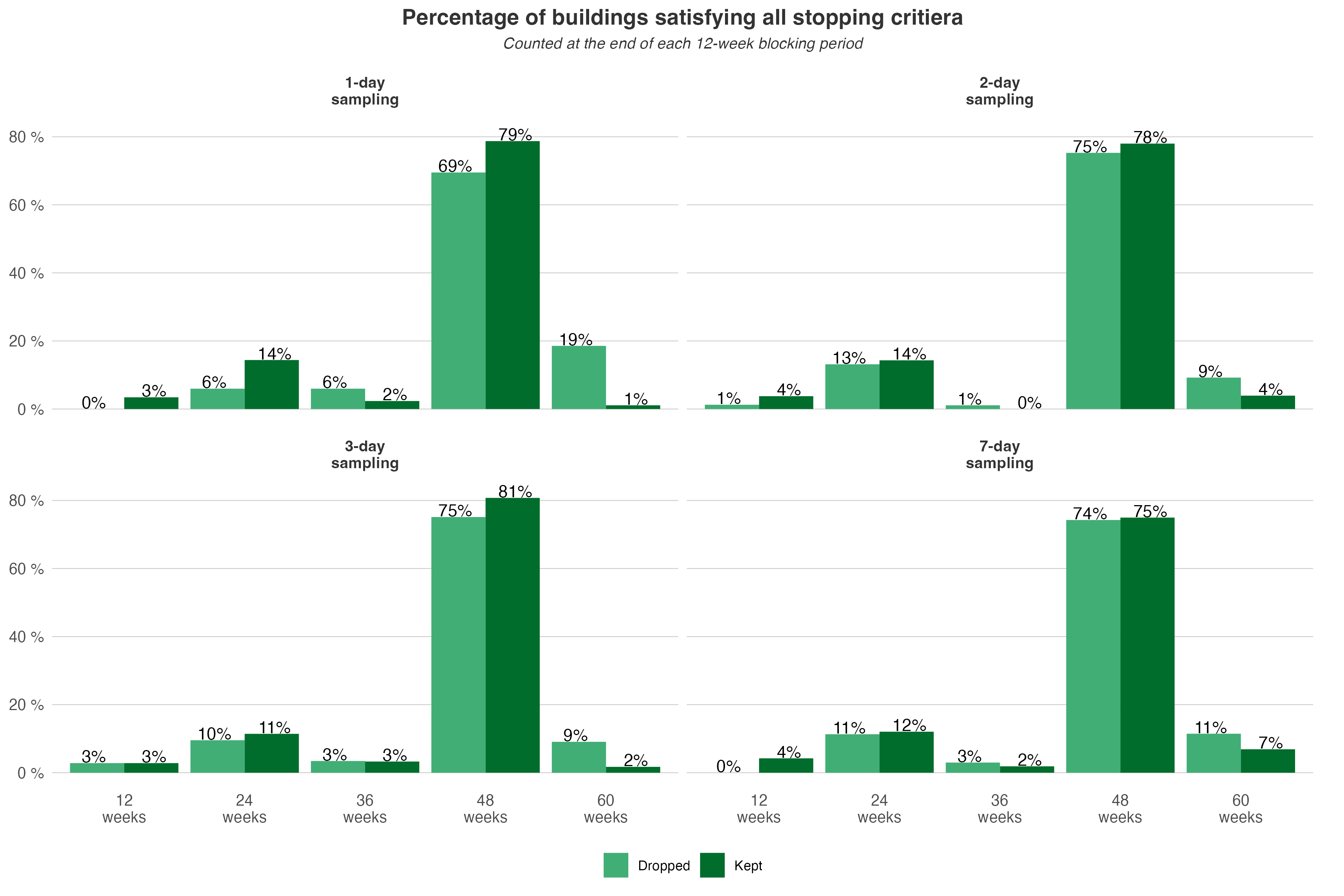


Figure 4.6: Time required to satisfy all stopping criteria starting from March

Similarly, we evaluated the time required to satisfy all stopping criteria when using a shorter blocking period, and included the plot in the supplementary material. We found that compared to Figure 4.6, a significant proportion of buildings completed the randomized M&V by week 42 (mid-December), which is one blocking period earlier than having a 12-week blocking period. This result further highlights that, in general, starting M&V in the middle of the winter or summer periods will yield results spanning the full range of weather conditions faster than starting at other times of year, such as during a swing season.

## 4.6 Limitations

Aside from the general limitation that this method only applies to the subset of interventions in buildings that are easily switchable, we identify two key limitations in this study:

1. The simulated intervention (chilled water temperature reset) 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.
2. For the main analysis, we assumed no carryover effect and that a daily sampling interval would suffice for most commercial buildings, but certainly exceptions exist. 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 randomized whole-building measurement and verification (M&V) method, comparing its performance to the conventional approach described in the IPMVP and in ASHRAE Guideline 14, using a large, open-source commercial building dataset. The randomized M&V method leverages the randomized experimental design concept from other scientific fields, along with statistical sequential inference techniques, to determine when savings are detected at sufficient certainty. 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 after 36 weeks (with the majority finished by 24 weeks) once all stopping criteria are 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, we verified that with a reduced M&V timeline, the randomized method can estimate savings more accurately by showing that the absolute error is only 1 - 2% for a typical building. Whereas for the much longer two-year conventional method, the estimation error was approximately 5% in a typical building. We also evaluated the impact of non-routine events, especially more gradual and subtle operational drift 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. In contrast, we found that those events have a very negligible impact on the savings estimated using the randomized method, demonstrating robustness to this issue. Furthermore, we consider scenarios where there is a known carryover effect from switching between strategies in the building and assess the impact of dropping samples where the strategy is non-consecutive and varying the sampling interval. Through that process, we showed that for the typical building, if carryover is present, using a 2-day sampling interval and dropping days with non-consecutive control strategies operating yields the optimal design for the median building in this dataset. Lastly, we investigated the time required to complete the randomized M&V method under different scenarios. We found that measuring a sufficient range of outdoor weather conditions remains the most stringent criterion. But even for the most non-ideal case, the randomized method can still produce an estimate much faster than 2 years, and produces an estimate with far lower error.

# 6 CRediT authorship contribution statement

**Aoyu Zou**: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Software, Writing - original draft, Writing - review & editing. **Paul Raftery**: Conceptualization, Funding acquisition, 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

A 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, 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, 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 Laboratory. 2015. “The Buildings Performance Database Overview.” <https://buildings.lbl.gov/sites/default/files/bpd-overview.pdf>.

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. 2023. “Treatment Effects in Market Equilibrium,” January. <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, Shuyang Li, Baihong Jin, Min Ting, Gwelen Paliaga, and Hwakong Cheng. 2018. “Evaluation of a Cost-Responsive Supply Air Temperature Reset Strategy in an Office Building.” *Energy and Buildings* 158 (January): 356–70. <https://doi.org/10.1016/j.enbuild.2017.10.017>.

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)