Lawrence Berkeley National Laboratory
LBL Publications

Title
A hybrid statistical-engineering approach to enhance the performance of non-routine event detection in building energy savings estimation

Permalink
https://escholarship.org/uc/item/63z977bc

Journal
Energy and Buildings, 360

ISSN
0378-7788

Authors
Granderson, Jessica
Fernandes, Samuel
Crowe, Eliot

Publication Date
2026-06-01

DOI
10.1016/j.enbuild.2026.117318

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

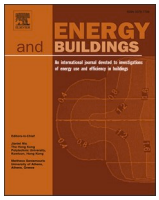
Energy & Buildings 360 (2026) 117318



Contents lists available at ScienceDirect

Energy & Buildings

journal homepage: www.elsevier.com/locate/enb





A hybrid statistical-engineering approach to enhance the performance of non-routine event detection in building energy savings estimation

Jessica Granderson, Samuel Fernandes *, Eliot Crowe

Lawrence Berkeley Laboratory, CA, United States

A B S T R A C T

In recent years, advanced measurement and verification (M&V) methods utilizing interval meter data have become increasingly important for quantifying energy savings from building efficiency projects. Driven by policy initiatives such as California's Assembly Bill 802 and Missouri's M&V 2.0 guidance, these methods leverage advanced metering infrastructure (AMI) data to assess energy savings through pre- and post-intervention analysis. However, the presence of non-routine events (NREs) poses a significant challenge to accurate energy savings estimation, as these events can obscure the effects of efficiency measures. Traditional NRE detection methods, reliant on manual inspections and expert judgment, hinder scalability and standardization in savings estimation. Recent advancements in automated detection methodologies, including machine learning and statistical techniques, show promise but face challenges related to false positives and data quality. This study builds on previous research by integrating engineering rules into a change point detection algorithm to enhance NRE identification accuracy. By incorporating domain-specific knowledge, the proposed approach aims to improve the precision and interpretability of detection results. The expanded test dataset allows for a comprehensive evaluation of the detector's performance across various NRE types. Key questions addressed include the detector's accuracy concerning different NRE characteristics and the potential for improvement through algorithmic refinements. The findings contribute to the ongoing discourse on enhancing M&V practices, ultimately supporting more reliable energy savings assessments in building efficiency projects.

1. Introduction

In recent years, the use of advanced measurement and verification (M&V) methods relying on interval meter data has gained prominence as a means for quantifying energy savings from building efficiency projects. This trend has been driven by various policy initiatives and state-level activities, such as California's enactment of Assembly Bill 802 [1], Missouri's guidance on adopting measurement and verification 2.0 (M&V2.0) [2], the New York Public Service Commission's order considering advanced M&V in utility programs [3], as well as advanced M&V pilots in Connecticut [4]. Advanced M&V utilizes advanced metering infrastructure (AMI) data (at hourly or shorter intervals) from before an efficiency intervention (the "baseline period") and after the intervention (the "reporting period") to determine project savings. This is achieved through development of energy model(s) that correlate energy consumption with independent variables. Concurrently, various tools, guidance documents and regulatory frameworks have been developed to support advanced M&V practices, including methods for detecting and addressing non-routine events (NREs)- a persistent challenge in this domain [5,6]. NREs, which refer to operational changes that can significantly impact energy consumption patterns, can confound energy savings estimations if not properly identified and accounted for [7]. They are atypical occurrences that significantly impact building energy consumption, deviating from the baseline conditions established during the initial measurement period [8]. These events can range from equipment malfunctions and operational changes to occupancy fluctuations. NREs may be broadly classified into three informal categories based on their duration: short-term, temporary, and permanent. Short-term NREs typically last for one or two days, temporary NREs span several weeks to several months, and permanent NREs indicate long-lasting changes. In the M&V context, NREs cause changes in building energy use that are not attributable to installed efficiency measures and are not accounted for in the baseline model's independent variables. NREs need to be detected first and then must be accounted for as 'adjustments' to appropriately attribute the estimated energy savings to the specific efficiency interventions that were implemented.

Traditionally, NRE detection has relied on manual methods, such as visual inspection of time-series energy consumption data and periodic surveys conducted using professional judgment and engineering expertise. As such it remains a barrier in scaling and standardizing meter-based savings estimation. In addition, there is an insufficient availability of ground truth data from actual buildings, coupled with the absence of standardized synthetic test datasets. Automated methods often struggle to adapt to dynamic consumption patterns, requiring more

* Corresponding author.
E-mail address: sgfernandes@lbl.gov (S. Fernandes).

https://doi.org/10.1016/j.enbuild.2026.117318
Received 26 September 2025; Received in revised form 24 February 2026; Accepted 11 March 2026
Available online 15 March 2026
0378-7788/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

sophisticated approaches. In recent years, researchers have explored various methodologies to address the challenge of NREs. Tonooka et al [9] proposed a middleware platform that leverages building energy management system (BEMS) big data analytics to automatically detect abnormal energy consumption. Their approach utilizes time-series energy use data collected from building subsystems, applying pattern recognition and statistical analysis techniques to identify deviations from expected operational behavior, thereby supporting early fault detection and operational diagnostics. Fenza et al. [10] presented a change point detection method for identifying anomalies in building energy consumption data, utilizing the cumulative sum algorithm to detect shifts in energy usage patterns. This drift-aware methodology allows for more accurate and timely identification of irregularities, contributing to the reliability and security of smart grid operations. However, the implementation of this approach may require significant computational resources and continuous monitoring, which could pose challenges for utilities and building operators with limited infrastructure. Buzau et al. [11] propose a supervised machine learning method for detecting non-technical losses (NTLs)- energy losses that are not due to physical defects in the infrastructure but rather to issues such as energy theft, faulty meters, or billing errors. Their method leverages smart meter data, but heavily depends on the quality and completeness of the available data and requires extensive labeled datasets for training. Oprea et al. [12] leverage hybrid supervised and unsupervised machine learning techniques to identify anomalies in electricity consumption data, but the authors use a finite dataset that limits the generalizability of findings to other regions or populations. In addition to these supervised techniques, multi-agent-based unsupervised methods have also been proposed for detecting anomalies in energy consumption at a smart campus scale [13]. While these approaches show promise, challenges remain in improving detection accuracy, reducing false positives, and ensuring scalability across different energy systems. Himeur et al. [14] proposed an outlier detection approach to identify abnormal energy consumption patterns in buildings, utilizing daily energy consumption and peak energy data. They established thresholds based on historical data to flag potential outliers. However, the author's approach heavily relies on historical data and may struggle to detect new, unforeseen energy consumption patterns. Goldberg [15] explored two methods for detecting and addressing NREs in residential and commercial AMI data. The paper explores how NREs can be systematically identified and incorporated into Normalized Metered Energy Consumption (NMEC) analyses. They highlight the challenges of accounting for these events—such as changes in operations, occupancy, or equipment—that, if ignored, can lead to inaccurate energy savings estimates. The authors propose structured approaches and best practices to make handling NREs more consistent and potentially "routine" within M&V frameworks. While these methods have shown promising results, there is no standardized test, data or metrics with which to compare different methods, and NRE detection remains a barrier to scaling and standardizing meter-based savings estimation. In addition, there is still no consensus regarding which performance metrics should be used to measure and report the performance of NRE detection systems and their underlying algorithms [16]. Furthermore, the field lacks comprehensive ground truth datasets from real building operations and standardized synthetic test datasets, which limits the ability to conduct rigorous comparative evaluations of NRE detection methods and hinders the establishment of benchmark performance standards across different building types and operational contexts.

Touzani et al. [7] leveraged a data driven methodology to automate, and therefore stream- line, the process of detecting NREs in the post-retrofit period and making associated savings adjustments. The statistical detector employs a penalized change point detection method combined with a dissimilarity metric to identify irregularities in building energy consumption. The method is designed to function independently of existing energy models, enhancing its applicability across various building types and operational conditions, and meaning it is not solely dependent on historical consumption data. The study was conducted on a relatively limited set of test conditions, which did not provide insight into the nature and type of events not detected by the algorithm. The primary drawbacks of the statistical approach in Touzani et al. stem from its model-agnostic nature, which, while flexible, lacks contextual sensitivity to operational or engineering-specific patterns. This can lead to false positives or missed detections when statistical changes do not correspond to meaningful operational shifts. A key finding of this research was that the detector was relatively accurate in identifying NREs in a simulated dataset but was oversensitive, resulting in a higher false positive rate than desired. Moreover, limited interpretability arises as purely statistical anomalies may not align with practitioner-relevant events. For detailed results of the performance of the NRE detector, refer to [7].

Hybrid statistical-engineering approaches, which integrate human expertise with rule-based model training, have demonstrated significant improvements in event detection across various scientific domains. For instance, Kovacevic et al. [17] developed a system that combines rule-based and machine learning techniques to extract temporal expressions and clinical events from medical narratives, achieving high accuracy rates, but also computational overload. Similarly, Ge et al. [18] proposed a human–machine collaboration framework for sound event detection, leveraging both machine learning models and human annotations to enhance detection performance. However, the approach may face limitations in scenarios where background knowledge is incomplete or not well-defined, potentially affecting detection accuracy. Additionally, in a study by Fanaee-T & Gama [19], the authors explored event labeling by combining ensemble detectors with background knowledge, highlighting the effectiveness of integrating diverse detection methods and expert insights. However, the reliance on human input may increase the time and labor required for annotation, potentially limiting scalability. These examples underscore the efficacy of hybrid approaches in enhancing event detection performance that leverage statistical techniques.

Building on prior successes in combining statistical methods with expert input, this study refines the change point detection algorithm of Touzani et al. [7] by integrating supplementary engineering rules into the statistical framework. By incorporating engineering insights into the change point detection process, the proposed approach in this paper aims to enhance the accuracy and reliability of NRE detection in building energy consumption data. Incorporating domain-specific knowledge—such as existing building energy consumption, equipment schedules, control sequences, or maintenance logs—into the detection process can improve both the precision and interpretability of the results. Supplementary rules are motivated by their ability to refine the output of statistical detectors by filtering or adjusting previously identified changepoints or flags. While these rules are effective at reducing the incidence of false positives and may influence true positive rates, they do not introduce new changepoints and therefore cannot increase sensitivity. Driven by this limitation and the need for advanced M&V practices and the need to handle non-routine events in portfolio-scale operations, the proposed method is intended as a triage mechanism, not an autonomous alarm suppressor. The detector identifies flagged time periods—potential NREs—alongside confidence and dissimilarity-context signals. These outputs allow practitioners to determine which events warrant investigation and savings attribution adjustment. Ultimately, the system acts as a prioritization filter, helping engineers and auditors focus on high-impact events to support scalable M&V 2.0 workflows without replacing human judgment.

We also expand the test dataset in order to assess the effectiveness over a broader set of conditions from the 2019 study. We summarize the development of the expanded test dataset and the results of applying the supplementary rules and internal parameter adjustments for the overall population of test cases, as well as based on key differences in test case typology (e.g., test cases varying by retrofit type, NRE type, NRE magnitude, etc.). This research is grounded in addressing the following

key questions:

1. What are the characteristics of the detector's accuracy across various types of NREs, such as temporary or permanent, short-duration or long-duration, large or small magnitude, variable or constant load shift? Does the detector exhibit better performance for certain types of NREs compared to others?
2. Can the detector's accuracy be improved through refinement of the internal parameters of the algorithm or with supplementary engineering rules?

This paper is organized as follows: Section 2 describes the original change point detector (baseline), supplementary engineering rules, internal parameter adjustments made, description of the simulated dataset and accuracy evaluation metrics; Section 3 describes the baseline detector accuracy; Section 4 describes the results from applying the supplementary rules to the detected NREs; and Section 5 presents discussion and conclusions.

2. Methodology

2.1. Baseline change point detector algorithm overview

The algorithm developed by Touzani et al. [7] for detecting potential NREs during the post-implementation period of an efficiency measure involves four key steps, outlined in the flowchart in Fig. 1. The fifth step in the flowchart, highlighted in grey, signifies the incorporation of supplementary engineering rules.

This approach assumes that in the absence of NREs, the behavior of pre- and post-retrofit energy consumption time series remains consistent in relation to the energy model's independent variable values. Dissimilarity, or dissimilarity score, is a measure where higher values indicate greater deviation (difference) from the baseline, and is used to detect changes. Traditional measures like Euclidean distance, Minkowski distance, and Fréchet distance focus on value proximity but neglect temporal structure, making it essential to consider both behavior and value for effective NRE detection [20]. This methodology leverages the differences in behavior to identify significant deviations, ensuring accurate detection of NREs. Fig. 2 illustrates the hourly energy consumption over the post-period (blue line) for a single test case, with the time period over which a simulated NRE is present marked in red, and the identified changepoints (represented by gray spikes on the charts) classified as either true positive (TP) or false positive (FP). In this example, an accurate NRE detector would flag the start of the NRE period when energy consumption steps up, and the end of the NRE when consumption steps back down. We see in this case that the detector is accurately flagging these two-time stamps (indicated TP), but flags an additional five timesteps that were not NREs (indicated as FP). The corresponding dissimilarity time series is shown in the lower panel. Dissimilarity measures the dynamic distance (composed of temporal and value proximities) between the baseline model's predictions and the post-period's actual energy data. This distance is calculated on a daily basis, allowing the detection of day-to-day deviations that may indicate operational changes or NREs. This dissimilarity series is an integral component of the NRE detector, which applies a changepoint algorithm to identify the start and end of potential NREs.

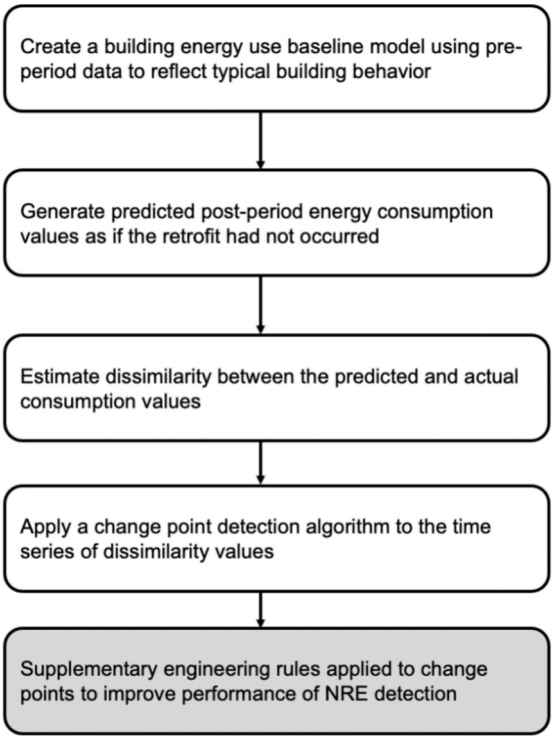


Fig. 1. Flowchart illustrating the NRE detection with the original algorithm (not highlighted) and using supplementary engineering rules (highlighted).

2.2. Supplementary engineering rules

The supplementary rules begin with changepoints detected by the detection algorithm and apply separate analytical rules to the time series consumption data, residuals, and/or the dissimilarity time series. We developed four supplementary rules to improve changepoint detection for non-routine events in energy consumption data:

2.2.1. Rule 1: dissimilarity time series rolling average

Initial attempts to evaluate dissimilarity time series using iterative calculations of mean and standard deviation, assuming normal distribution with $\pm 3\sigma$ thresholds, proved computationally expensive, and the normality assumption could not be verified. We implemented an alternative approach using rolling averages with fixed starting points to identify sustained energy use changes. This method calculates:

1. Rolling average of dissimilarity with a fixed starting point
2. Slope between identified change points

Both metrics serve as filtering criteria for the identified changepoints. Our method included the assessment of different segment lengths and slope thresholds, outlined in table 1.

In Fig. 3, the top panel presents a comparison between the baseline model predictions (depicted in red) and the actual measured energy consumption (blue). Grey vertical lines indicate changepoints identified by the base detector, which operates by detecting structural breaks in the observed data. The bottom panel illustrates the dissimilarity metric (primary Y axis), where light teal points represent the raw dissimilarity values computed between baseline and actual time series segments. Overlaid in orange is the rolling average of the dissimilarity signal (secondary Y axis), which serves to smooth short-term fluctuations and emphasize persistent deviations. Rule 1 leverages this rolling dissimilarity behavior—specifically its slope and segment duration—to distinguish true NREs from noise-induced anomalies. The hypothesis is that a steeper slope is more likely to indicate an NRE.

To determine the optimal configuration, a grid search was performed across the parameter combinations listed in Table 1. The objective function was defined to maximize the reduction in False Discovery Rate (FDR) while maintaining a minimum True Positive Rate (TPR) of 60%.

The analysis revealed that shorter segment lengths (e.g., 7 days) were overly sensitive to noise, failing to sufficiently reduce false positives, while longer segments (e.g., 21 days) often smoothed over shorter-duration NREs, significantly degrading sensitivity. The combination of a 14-day minimum segment and a 0.75 slope threshold provided the most

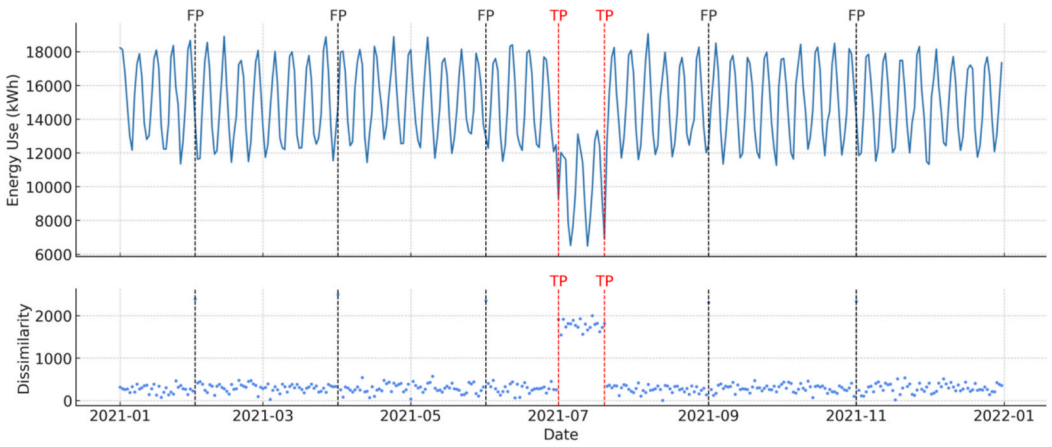


Fig. 2. Example NRE test case time series. The top figure shows the energy consumption over the post period in blue. The time period over which a simulated NRE is present is marked in red, and the changepoints are identified as true and false positives. The bottom figure shows the corresponding dissimilarity time series. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

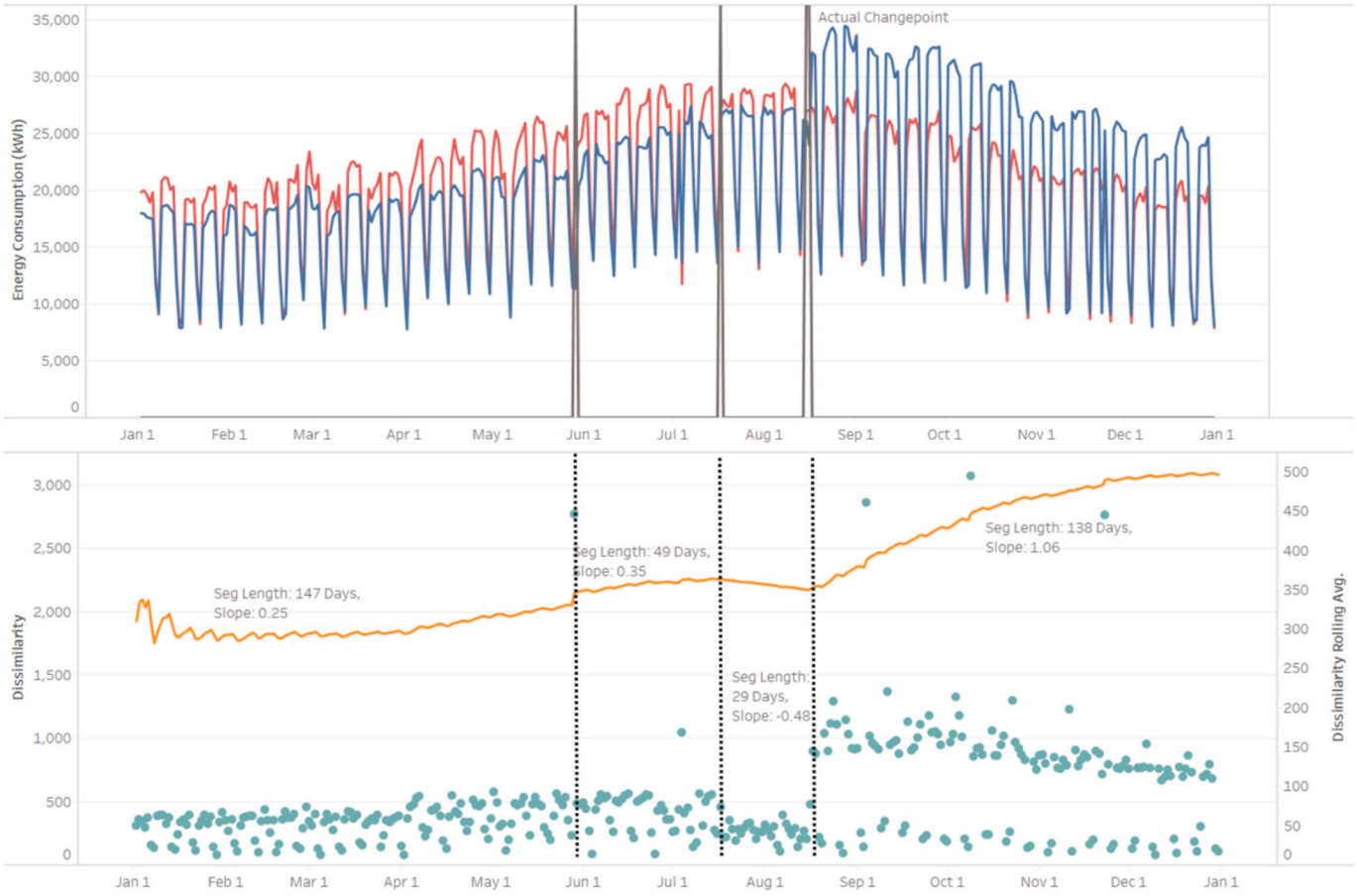


Fig. 3. Rule 1 implementation. Top chart: baseline model predictions (red), actual energy use (blue), identified changepoints by base detector (grey). Bottom chart: dissimilarity (light teal), rolling average of dissimilarity (orange). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

balanced trade-off, filtering out transient noise while retaining sensitivity to genuine operational shifts.

Regarding computational efficiency: while the *selection* of the rolling average method over iterative normality tests was driven by computational constraints, the specific choice of parameter values (e.g., 14 days vs. 21 days) had negligible impact on runtime complexity, as the algorithm operates in linear time regardless of window size.

2.2.2. Rule 2: mini-model prediction comparison

This rule constructs a predictive model using hourly consumption data from the post-period prior to the first identified changepoint, under the assumption that this changepoint is a true NRE. The mini-model is trained only on this initial segment and then used to predict energy use for the remainder of the post-period. The comparison between predicted and actual values is used to visually assess whether subsequent changepoints reflect real shifts in building behavior. Model performance

Table 1
Segment lengths and slope threshold combinations tested.

Segment length (days)	Slope threshold
7	0.50
10	0.50
14	0.50
21	0.50
14	0.40
14	0.60
14	0.65
14	0.70
14	0.75

depends heavily on the fit quality and coverage of training data. If the first changepoint occurs early in the post-period, the model may lack sufficient data to capture the building's operational dynamics, leading to poor predictive accuracy. In Fig. 4, the top panel shows actual energy consumption (blue), the mini-model fit over the pre-changepoint period (red), and the model's predictions for the remainder of the time series (green). Vertical brown lines mark changepoints flagged by the base detector. The divergence between actual and predicted energy use indicates potential NREs not well captured by the model. The bottom panel displays the dissimilarity series (purple), which measures the dynamic deviation between model predictions and actual energy use. The visualized pattern underscores how poorly fitted or insufficiently trained models may limit this rule's utility as an automated filter — but it can still provide value as a diagnostic or visual aid in identifying and interpreting non-routine events. Brown vertical lines denote changepoints identified by the base detection algorithm, which mark points where the observed energy behavior diverges significantly from the predicted baseline. These changepoints serve as indicators of potential NREs.

The method underperforms when change points occur early in the dataset due to insufficient model training. This technique functions optimally as a visual confirmation tool when models have adequate data coverage across varied operational conditions.

2.2.3. *Rule 3: Directional analysis of daily-averaged residuals*

This method analyzes directional changes in residuals between baseline predictions and post-period measurements. Implementation involves: (1) calculating residuals between baseline predictions and actual post-period data, (2) identifying directional shifts as NRE indicators. In Fig. 5, the top panel shows a comparison between the baseline model's predicted energy use (red) and the actual measured energy use (blue). Vertical dashed lines indicate changepoints detected by the base algorithm, which signal shifts in behavior that deviate from the modeled baseline. The bottom panel displays the cumulative residuals, calculated as the running sum of the differences between actual energy use and baseline predictions during the post-change period.

This approach performs adequately only in datasets with significant NREs that force directional changes in cumulative residuals. The technique serves primarily as a visual validation tool for qualified datasets rather than as a standalone detection mechanism.

2.2.4. *Rule 4: fractional change in segment energy consumption*

The final supplementary rule evaluates fractional changes in average energy consumption between adjacent segments defined by change points. Segments exhibiting fractional changes exceeding predetermined thresholds may be indicators of true NREs. Fig. 6 illustrates a time series of daily building energy consumption (in kWh), with changepoints identified by the base detection algorithm denoted by

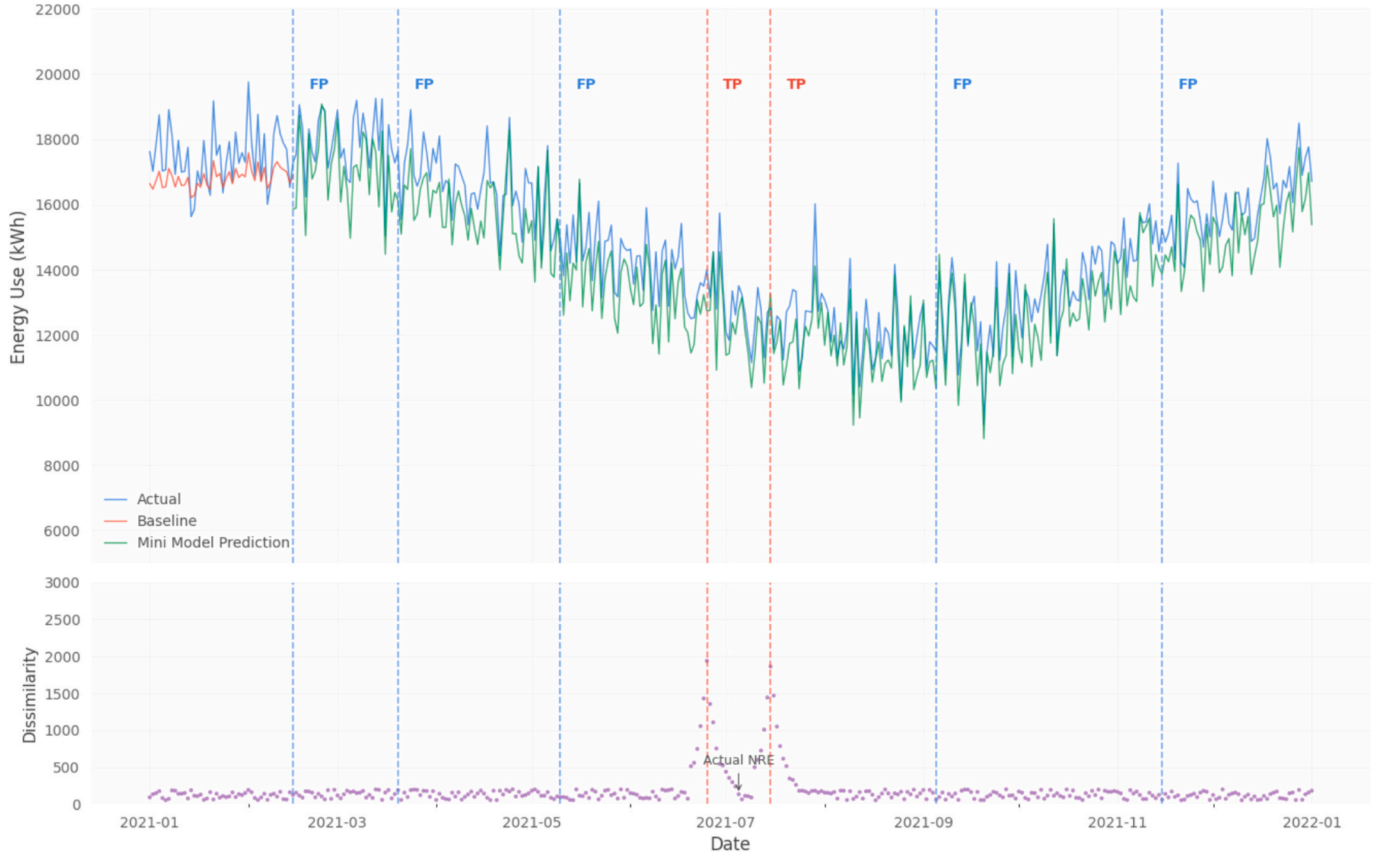


Fig. 4. Top chart: mini model fit (red), actual energy use (blue), mini model predictions (green), identified changepoints by base detector (brown). Bottom chart: dissimilarity (purple), identified changepoints by base detector (brown). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

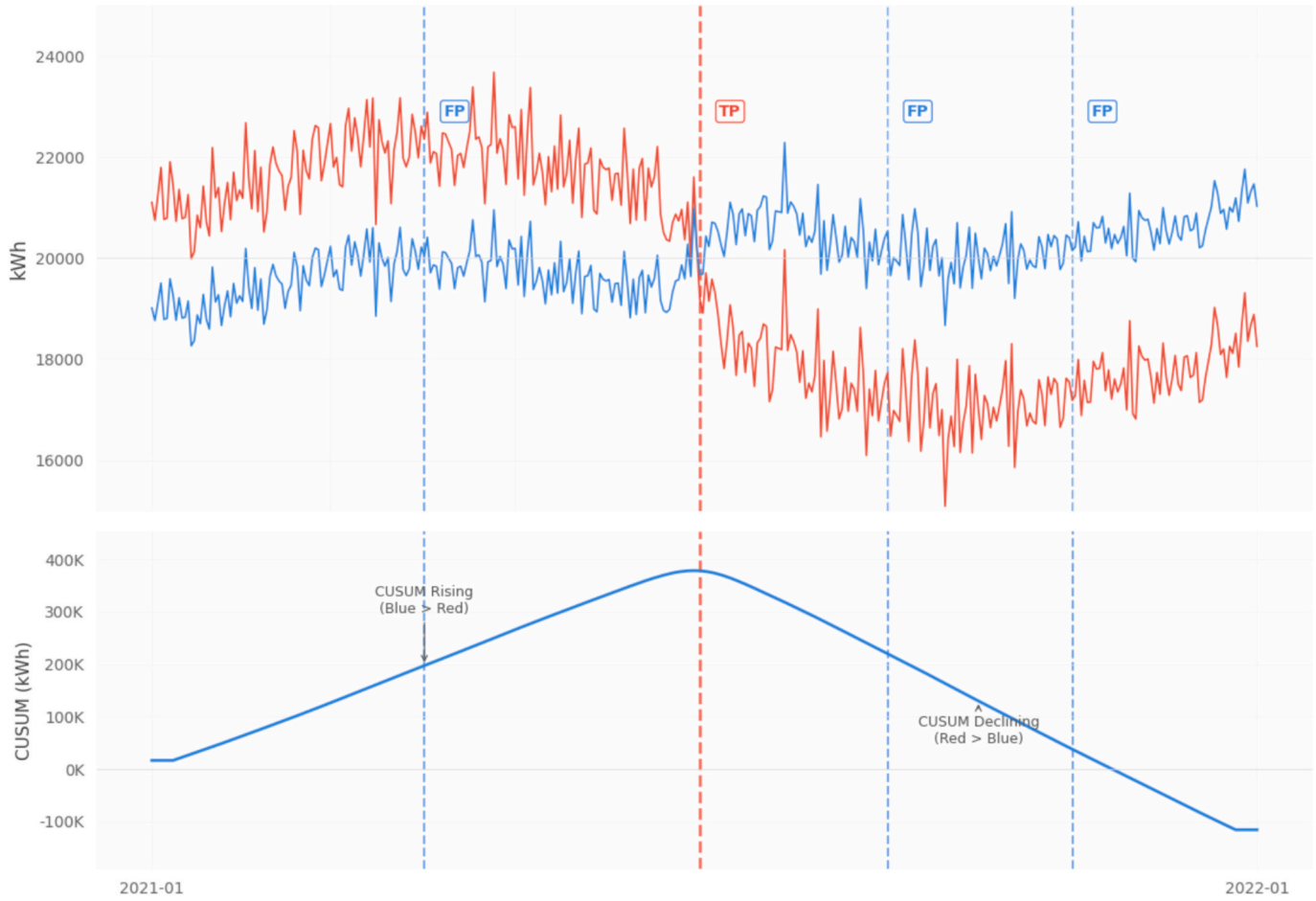


Fig. 5. Top chart: baseline model predictions (red), actual energy use (blue), identified changepoints by base detector (TP and FP dashed lines). Bottom chart: Cumulative residuals between the baseline model's predictions and actual energy use in the post period. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

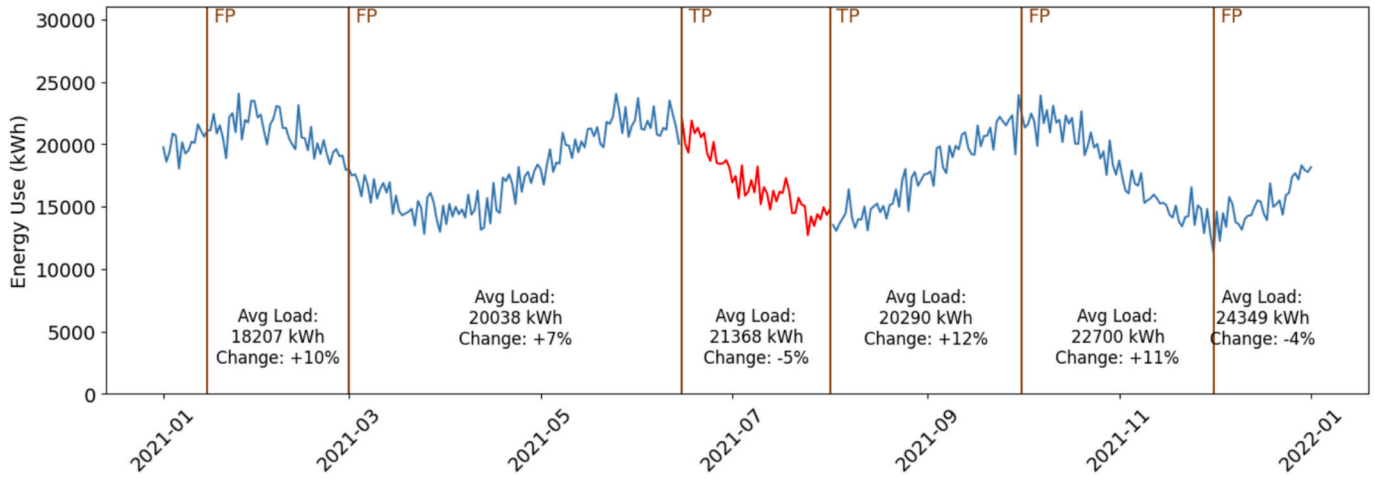


Fig. 6. Figure depicting energy usage (kWh) highlighting FP and TP with corresponding average load changes between segments. Red segment indicates period of true NRE. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

vertical brown lines. The time series is segmented based on these changepoints, and the average energy use for each segment is labeled along with the relative percentage change compared to the previous segment. Notably, the segment highlighted in red indicates the period corresponding to a true NRE. The change in average load between segments ranges from −5% to +12%, reflecting varying operational or occupancy-related shifts throughout the year.

While conceptually similar to Rule 1, this approach focuses exclusively on energy consumption values rather than the dissimilarity series that incorporates both value proximity and behavioral patterns. Consequently, this method excels at identifying constant, large-magnitude NREs but demonstrates reduced sensitivity to small-

magnitude or variable-load NREs that produce gradual consumption changes.

2.3. Internal parameter adjustments

In addition to the supplementary rules described above, a parameter tuning was implemented. This study focused on tuning parameter k, which governs the weighting between temporal behavior and value proximity in the dissimilarity calculation between adjusted baseline and post-period data. k is a tuning parameter that modulates the weight of the behavioral component. A higher k value forces the score to be more sensitive to changes in temporal shape (correlation) rather than just magnitude. In the original 2019 study, k was set to 1.0. For this study, we performed a sensitivity analysis (grid search) on a training subset of the data, testing values from 0.5 to 5.0. The value k = 2.5 was selected as it maximized the harmonic mean of sensitivity and specificity across the aggregate dataset.

2.4. Simulated dataset

This study analyzed 180 simulated test cases to evaluate changepoint detection algorithms under diverse NRE scenarios in building energy consumption. The Large Office building type was chosen due to its consistent year-round operational profile—relative to buildings with more seasonal operational variation such as schools—and its prevalence in the national building stock. All simulations were designed to capture performance variation across NRE types, durations, magnitudes, and climate conditions. The dataset incorporated broad variations in simulation parameters to generate more generalizable conclusions. Table 2 provides a summary of the simulation parameters used, with more details provided in the Appendix.

Table 2
Simulation parameters.

Parameter	Description
Climate Zones	All 15 ASHRAE-defined climate zones were represented.
Retrofit Types	– Lighting Retrofit: Reduced Lighting Power Density from 16.89 W/m^2 to 8 W/m^2. – Cooling Efficiency Upgrade: Increased Chiller COP from 5.2 to 7.8.
NRE Types	– Change in Occupancy Start/End Times – Change in Occupancy Density – Change in Base Plug Load
NRE Durations	– Temporary–Short – Temporary–Long – Permanent
NRE Magnitudes	Events were labeled as small or large based on magnitude of load variation relative to total consumption
NRE Implementation Methodology	Each simulation contained exactly one NRE to simplify evaluation of detection accuracy. Temporary events introduced two changepoints; permanent events introduced one. Single-day events were excluded due to negligible energy impact.
Occupancy Schedule Changes	– Baseline: 8:00 am–5:00 pm – Modifications drawn from truncated normal distributions: - Start Time: μ = 8:00 am, σ = 2h, bounds = - End Time: μ = 5:00 pm, σ = 2h, bounds = – Magnitude threshold: - Small: <3-hour shift - Large: ≥3-hour shift
Occupancy Density and Plug Load Changes	– Multiplier drawn from truncated normal distribution – Direction randomly assigned – Magnitude classification: - Small: multiplier ≤ 1.7 - Large: multiplier > 1.7

2.5. Accuracy evaluation metrics

The change point detection algorithm was assessed using two key performance metrics: true positive rate (TPR) and false discovery rate (FDR). The TPR quantifies the percentage of actual NREs correctly identified by the detector, while the FDR measures the percentage of flagged change points that do not correspond to true NREs.

$$TPR = 100 \times \frac{\sum_{i=1}^{n_{NRE}} I(min_j\{|NRE_i - \delta_j|\} \le th)}{n_{NRE}}$$

$$FDR = 100 \times [1 - \frac{n_{NRE} \times TPR}{n}]$$

where NRE_i is the actual location of ith NRE, δ_j is the location of jth detected potential NRE, n_{NRE} is the actual number of locations (i.e., dates) where there is an NRE, n is the number of potential NREs, and th is the threshold that defines an actual NRE as detected by the algorithm if the estimated position is within ± th number of days.

Both metrics incorporate a temporal tolerance threshold (±th days) to account for slight variations in the precise timing of detected versus actual NRE locations. The TPR uses the total number of true NREs as its denominator, indicating detection sensitivity, whereas the FP rate uses the total number of flagged change points as its denominator, reflecting detection specificity. This evaluation framework provides a comprehensive assessment of the detector's ability to accurately identify genuine events while minimizing false alarms, with the threshold parameter allowing for realistic temporal matching between predicted and actual NRE occurrences.

The 2019 study used a threshold of ±2 days to classify detection success. Although this level of temporal precision may exceed the requirements for annual energy savings estimation, it offers a rigorous benchmark. Recognizing that if the primary M&V metric is annual energy savings, it may be overly restrictive to have a 2-day precision window, so this study extended the evaluation to include ±5-day and ±7-day thresholds. This multi-threshold approach allows a more comprehensive assessment of algorithm sensitivity under different temporal precision requirements.

3. Results

In this results section, we first evaluate the detector's baseline performance across various NRE categories, examining TPR and FDR for different event characteristics including temporary versus permanent, short-duration versus long-duration, large versus small magnitude, and variable versus constant load shifts. Next, we assess potential performance improvements through two optimization approaches: (1) application of supplementary engineering rules, particularly Rule 1 with its 14-day minimum segment length and 0.75 slope threshold parameters, and (2) tuning of the algorithm's internal k parameter to emphasize behavioral proximity in the dissimilarity computation. We then conduct a detailed analysis of how detection effectiveness varies across specific retrofit contexts (lighting versus cooling) and NRE characteristics (constant versus variable load, temporary versus permanent duration), identifying the conditions under which each detection method demonstrates optimal performance. This comprehensive evaluation provides actionable insights for tailoring detection strategies to specific building operational contexts while balancing the critical trade-off between sensitivity and specificity in NRE detection.

3.1. Characteristics of the detector's accuracy for various types of NRE

Fig. 7 compares the average TPR and FDR rates across different NRE categories. (See Table 8 in the Appendix for the number of test cases contributing to each category average). Most categories have an average TPR around 70–80%, indicating that the model or test correctly

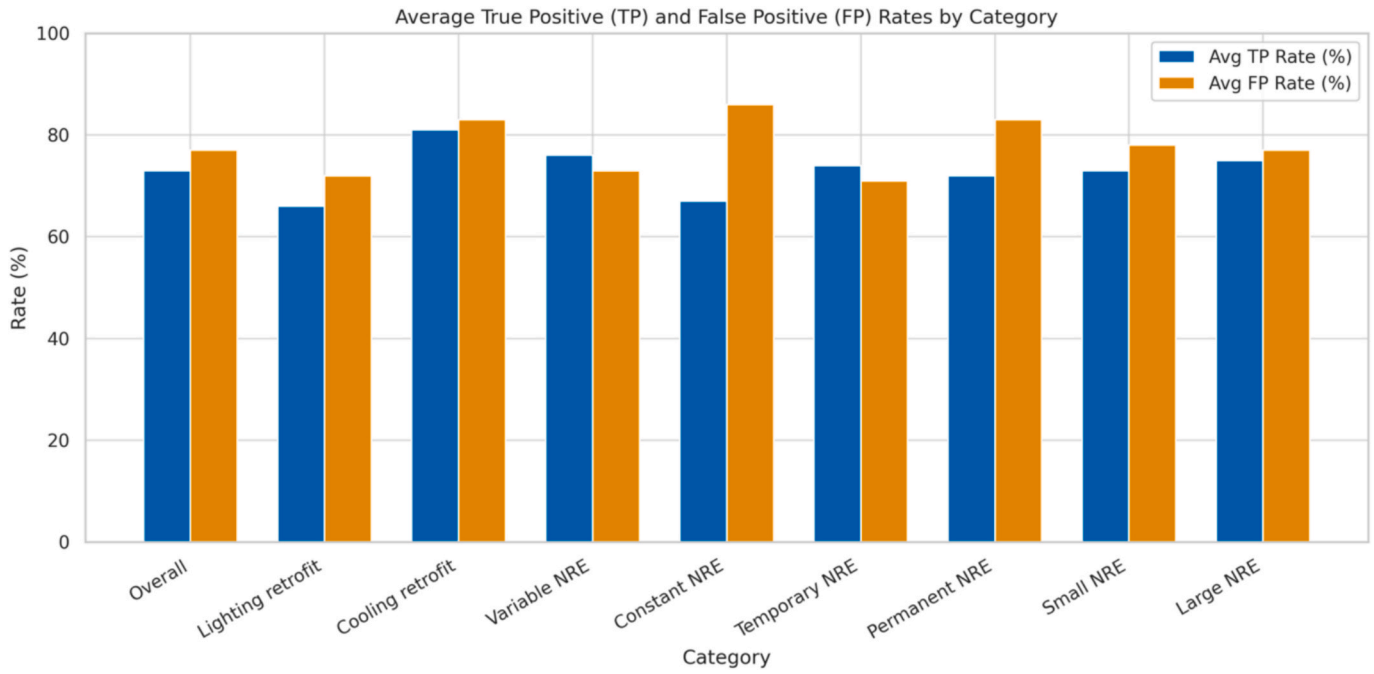


Fig. 7. TP and FP rates by NRE Category.

identifies 70–80% of the NREs in these categories. The "Overall" category has a TPR slightly above 80%, suggesting a relatively high sensitivity across all data. The orange bars represent the FDR, which is consistently around 80% across all categories. This indicates that the detector has a high rate of false positives, meaning it incorrectly labels 80% of negative cases as positive (i.e., for every 10 changepoints flagged by the detector, 8 of them are not true NREs). The consistently high FDR across all categories suggests that while the model is good at detecting true positives (high TPR), it is over-sensitive, leading to a large number of false positives. The TPR and FDR values are relatively consistent across the categories, with minor variations. For example, the "Variable" category has a slightly lower TPR (around 70%) compared to "Overall" (above 80%), but its FDR remains around 80%, similar to other categories.

Fig. 8 illustrates the trade-off between average TPR and FDR for various energy retrofit and NRE detection strategies. Each point represents a category—such as lighting or cooling retrofits, or types of NREs like "Temporary" or "Permanent"—with the x-axis showing TPR and the y-axis showing FDR, both expressed as percentages. Categories like Lighting retrofit and Constant NRE demonstrate lower TPRs but also exhibit variation in FDRs, suggesting potential inefficiency or inconsistency in detecting true positives. In contrast, Temporary NRE and Large NRE show higher TPRs with relatively moderate FDRs, indicating more reliable performance in identifying actual events without excessive false alarms. Overall, variable NREs and lighting retrofits demonstrate superior detection precision through reduced FP incidence, whereas constant NREs and cooling retrofits remain challenging due to persistent and ambiguous load patterns. To enhance interpretability, the plot

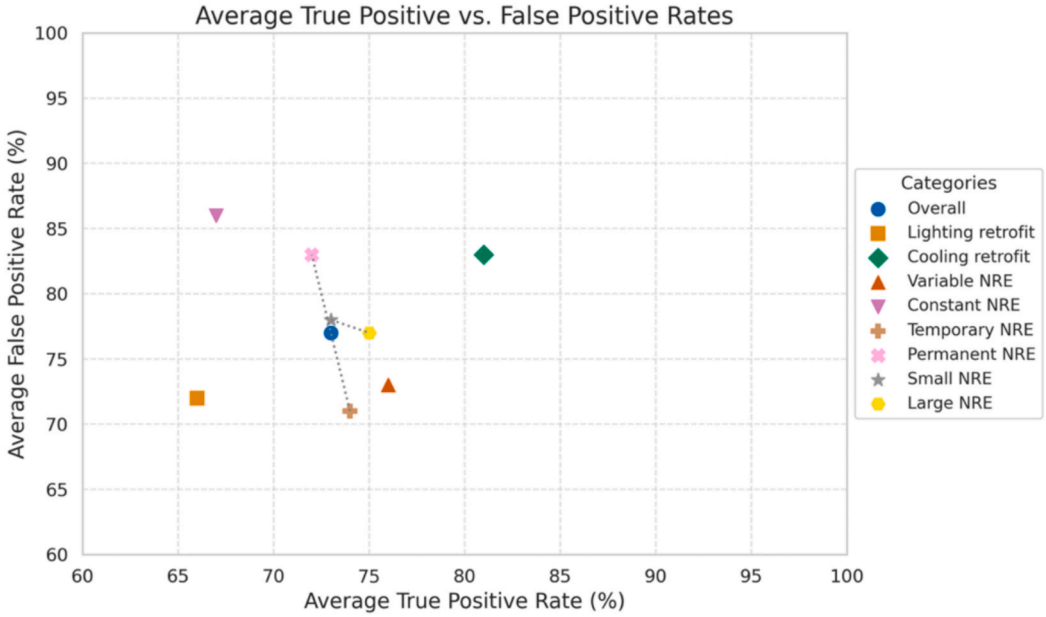


Fig. 8. Average TP and FP rates across different NRE categories.

directly labels each data point and includes dotted lines connecting conceptually related categories (Small vs. Large NRE and Temporary vs. Permanent NRE). These connections help visualize how detection performance varies within related groupings. The visual layout underscores that improving NRE detection involves balancing the ability to catch true positives while minimizing false alarms, a crucial consideration for practical deployment in NRE. Notably, categories such as Cooling retrofit and Constant NRE exhibit relatively high FDR, exceeding their corresponding TPR. This suggests that while these systems may be sensitive, they could produce a higher volume of incorrect detections, which could lead to unnecessary interventions or misinterpretation of system behavior. On the other hand, categories like Large NRE, Small NRE, and Temporary NRE show a more balanced or favorable TP-to-FP ratio, indicating a more effective detection mechanism in terms of both sensitivity and specificity. The Overall category displays moderate performance, reflecting the aggregate effect of its subcategories. These differences highlight the potential benefit of tailoring detection strategies based on the type of retrofit to optimize operational accuracy and efficiency. For example, buildings with certain retrofit types may benefit from adjusted sensitivity thresholds or different detection parameters. Also, since detection accuracy does not vary significantly by NRE magnitude, we do not look at the detailed impact in the results.

3.2. Accuracy results of applying the supplementary rules and parameter tuning

Among the supplementary rules, only Rule 1—which confirms or rejects changepoints based on shifts in the rolling average of dissimilarity values—demonstrated reliable performance, significantly reducing false positives across all test cases, though with a modest decrease in true positives for small NREs. The other rules (mini-models, residual directionality, and average segment consumption) were found to be insufficiently robust for automated use and are better suited to support visual review rather than standalone implementation. Additionally, tuning the k parameter from 1.0 to 2.5 increased the TPR while only marginally raising the FDR, particularly benefiting detection in variable-load contexts. Consequently, only Supplementary Rule 1, k parameter tuning, and their combination were selected for further analysis, as they offered the best trade-off between sensitivity and specificity within an automated framework. Results highlight the baseline detector, an enhanced detector using Rule 1 (14-day minimum segment length, 0.75 slope threshold), a variant with parameter k tuned to 2.5 to emphasize behavioral proximity, and a combined approach applying both enhancements.

As shown in Fig. 9, the average TPR exhibits notable variation across detection configurations. The base detector achieves a mean rate of 73%, while application of Rule 1 reduces this to 61%, indicating a 16% decline in sensitivity. Tuning the k parameter to 2.5 yields the highest mean TPR at 83%, improving performance by 10% over the base detector. However, combining Rule 1 with $k = 2.5$ results in the lowest mean TPR of 50%, suggesting that the enhancements interact sub optimally. These findings indicate that while behavioral weighting improves true positive detection, Rule 1's filtering mechanism may suppress valid changepoints, and their combination further compromises detection efficacy.

The average FDR reveals that as TPR decreases, FDR also tends to decrease, though the relationship varies by method. The base detector exhibits a high FDR of approximately 80% (meaning 8 out of every 10 flagged changepoints are not true NREs), which is significantly reduced to 50% with the application of Rule 1—meaning that 38% of the false positives identified by the base detector are filtered out through the application of Rule 1. In contrast, tuning the k parameter to 2.5 increases the FDR to 88%, indicating diminished precision. Notably, the combined approach (Rule 1 + $k = 2.5$) achieves the lowest FDR at 33%, suggesting strong suppression of spurious detections. These results highlight that while k tuning enhances sensitivity, Rule 1 is more effective for false positive mitigation, and their combination maximizes specificity at the cost of reduced sensitivity.

3.3. Detection performance by retrofit type and NRE characteristics

The study examined how key NRE characteristics influence the performance of changepoint detectors, with particular attention to retrofit type (lighting vs. cooling), NRE type (constant vs. variable load), NRE magnitude (small vs. large), and NRE duration (temporary vs. permanent). These variables were assessed in the context of their interaction with two enhancement strategies: Rule 1 and k parameter tuning.

3.4. Detection performance by retrofit type

Fig. 10 presents the average TPR and FDR for different detection methods disaggregated by retrofit type: cooling retrofits (left panel) and lighting retrofits (right panel). Each panel compares four detection strategies: the base detector, Rule 1, a behavioral distance weighting method ($k = 2.5$), and a combination of all rules. In the cooling retrofit

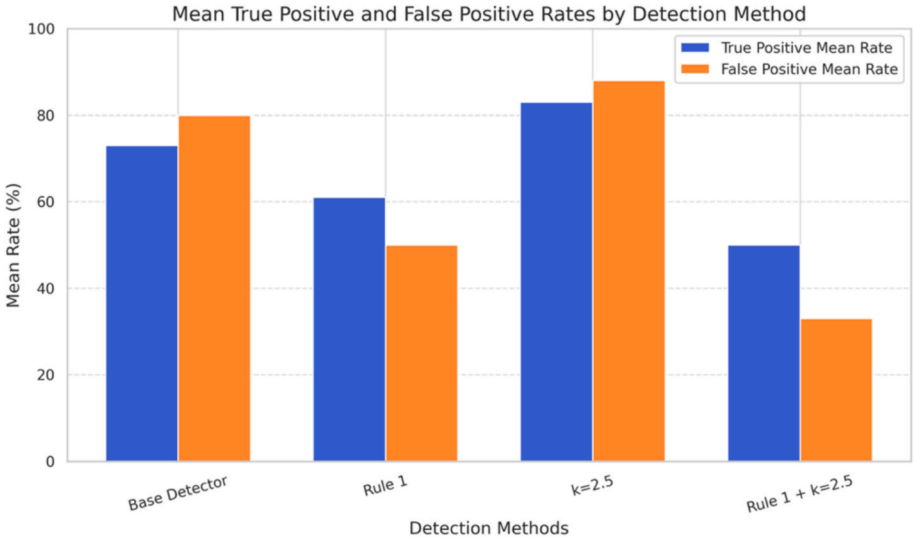


Fig. 9. Comparative analysis of different detection methods for identifying changepoints in energy consumption data.

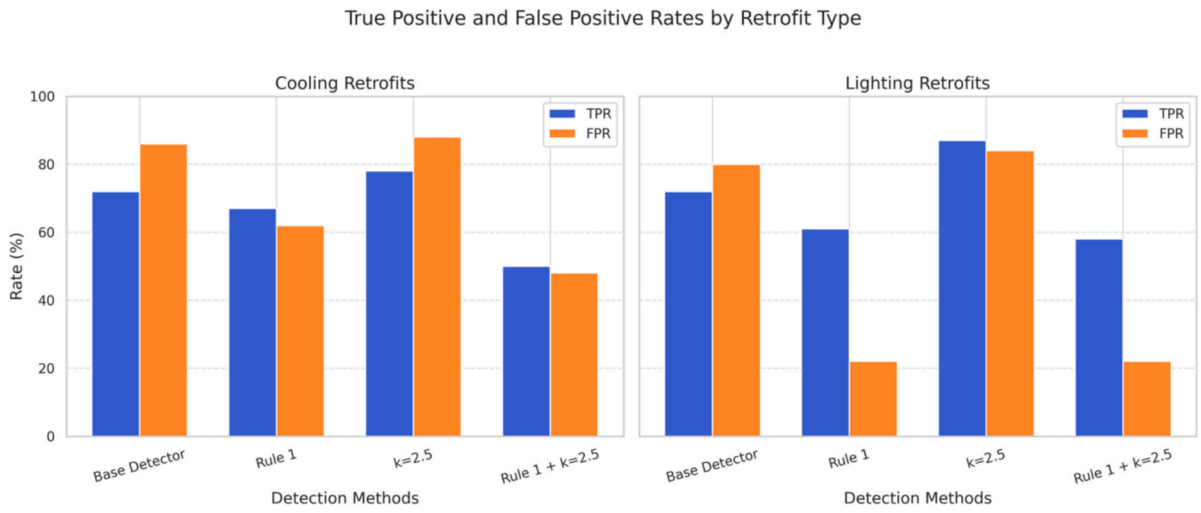


Fig. 10. TP and FP rate by detection method and retrofit type.

panel (left), the base detector achieves a relatively strong TPR of approximately 70%, but this comes with a very high FDR of around 85%, indicating that it frequently misclassifies normal behavior as NREs. Applying Rule 1 significantly reduces the FDR to about 60%, thereby improving specificity; however, it also lowers the TPR to approximately 60%, reflecting a loss in sensitivity. When behavioral weighting is increased (k = 2.5), the method yields the highest TPR of around 85%, demonstrating improved ability to detect true changes. Yet, this gain comes at a cost—FDR rises further to about 90%, underscoring a trade-off between sensitivity and false alarms. The combined approach, which integrates both Rule 1 and the k-weighted method, achieves the lowest FDR of approximately 45%, but this is accompanied by the lowest TPR of around 50%, illustrating a substantial reduction in detection capability in exchange for greater specificity.

In the lighting retrofit panel (right), the base detector again shows similar performance, with a TPR near 75% and an FDR also around 85%. However, Rule 1 proves especially effective in this context, sharply reducing the FDR to about 20%, while only slightly reducing the TPR to approximately 65%, indicating strong precision with minimal sensitivity loss. The k = 2.5 method continues to produce high TPR values of approximately 90%, but again exhibits a high FDR of about 80%, consistent with its tendency to prioritize detection at the cost of false positives. Notably, the combined method offers the most balanced performance, lowering the FDR to about 25% while maintaining a respectable TPR of approximately 60%, thereby delivering a more optimal trade-off between sensitivity and specificity for lighting-related changes.

Interestingly, tuning the k parameter to 2.5—which gives greater weight to behavioral proximity—led to a larger improvement in TPR for lighting retrofits (19%) than for cooling retrofits (4%). This finding was counter to expectations, as behavioral weighting was presumed to better support detection in behavior-altering retrofits like cooling. The results also underscore the importance of retrofit-specific tuning in detection systems. Rule 1 is especially effective for retrofits characterized by sustained, less variable shifts (e.g., lighting), while parameter tuning with k improves sensitivity but may require additional filtering to mitigate false positives.

3.5. *Detection performance by NRE type*

Fig. 11 compares the average TPR and FDR across four detection strategies—Base Detector, Rule 1, k = 2.5, and the combined Rule 1 +

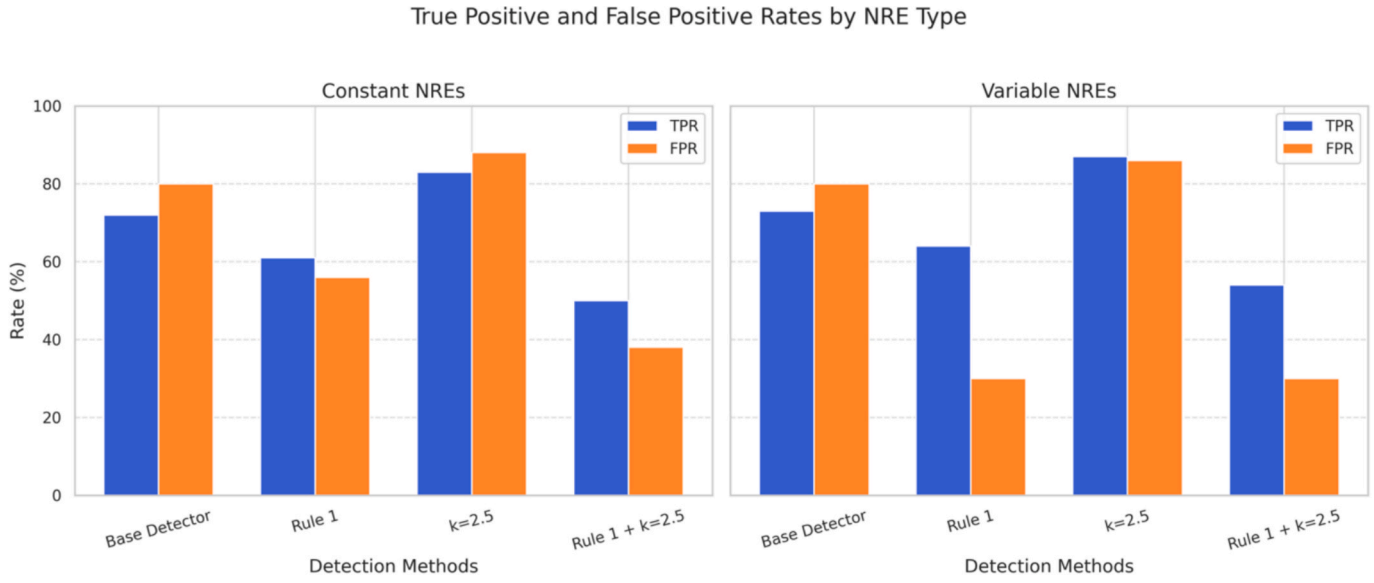


Fig. 11. TP and FP rate by detection method and NRE type.

k—for two categories of NREs: constant and variable, shown in the left and right panels respectively. In the constant NRE category (left panel), which includes events with persistent changes in energy consumption, the base detector achieves a relatively high TPR of 75% but also a high FDR of 80%, indicating many false alarms. Rule 1 substantially reduces FDR to 55%) while slightly decreasing TPR. The k = 2.5 method yields the highest TPR of 90% but at the cost of increased FDR of 90%). The combined method reduces both TPR and FDR significantly to 50% and 40%, respectively, prioritizing specificity over sensitivity. For variable NREs (right panel), which exhibit more temporally dynamic changes, all methods show improved overall performance. The base detector and k = 2.5 both achieve high TPRs of 80%– and 95%, but only Rule 1 and the combined method effectively reduce FDR. Notably, Rule 1 achieves the best FDR reduction of 30% while maintaining a TPR close to 65%. The combined method achieves the lowest FDR of 30% but at the expense of TPR 55%. These findings suggest that Rule 1 is particularly well-suited for detecting variable NREs with fewer false positives, while the k threshold enhances sensitivity across both NRE types. However, the combination of Rule 1 and offers the most conservative strategy, minimizing false positives but sacrificing event detection capability.

This disparity reflects how variable load NREs produce both value and behavioral changes, which in turn generate higher dissimilarity values that Rule 1 can more reliably detect. Constant load NREs, which only impact value proximity, generate subtler dissimilarities that are harder to distinguish from background variation.

3.6. Detection performance by NRE duration

Fig. 12 compares the TPR and FDR of four detection methods—Base Detector, Rule 1, k = 2.5, and the combined Rule 1 + k—across NREs classified by their temporal duration: temporary and permanent, shown in the left and right panels respectively. In the temporary NRE category (left panel), where deviations are short-lived, the base detector achieves moderate TPR of 75% but suffers from a high FDR of 80%). Applying Rule 1 reduces FDR substantially to 35% while maintaining a TPR above 60%, demonstrating effective filtering of spurious detections. The k = 2.5 method yields high TPR of 90% but retains a high FDR of 90%, emphasizing sensitivity at the cost of specificity. The combined method further reduces FDR to 20% but at the expense of TPR, which drops to 50%.

In the permanent NRE category (right panel), characterized by sustained shifts in energy use, the base detector again exhibits high TPR of 80% with elevated FDR of 85%. The effect of Rule 1 is less pronounced here: while it reduces FDR slightly to 70%, it also diminishes TPR. The k = 2.5 method again shows high TPR of 90%) but elevated FDR. The combined strategy produces the lowest FDR of 50%) but correspondingly lowers TPR to around 60%. These results indicate that temporary NREs benefit more from Rule 1, which filters out short-term fluctuations. In contrast, permanent NREs are more challenging, as the rolling dissimilarity metric used in Rule 1 may plateau, limiting its discriminatory power.

In summary, detector performance is highly sensitive to retrofit and NRE characteristics. Across all test cases and detection methods, TPR ranged from as low as 50% (Rule 1 + k = 2.5 for cooling retrofits) to as high as 95% (k = 2.5 for variable NREs). FDR varied from as low as 20% (Rule 1 + k = 2.5 for temporary NREs) to as high as 90% (k = 2.5 for constant NREs and cooling retrofits). Importantly, when compared with the base detector, there was no test case or method combination that achieved both an increase in TPR and a decrease in FDR; in all cases there was a clear trade-off between the two metrics. Rule 1 is particularly effective in improving specificity in lighting retrofit scenarios and temporary or variable load NREs but can compromise sensitivity in constant or permanent cases. Meanwhile, k parameter tuning improves sensitivity, especially in lighting retrofit datasets, but may increase FDR. These findings underscore the importance of tailoring detection strategies to the specific operational and load characteristics of a building and its retrofit context.

4. Discussion

The results of this study demonstrate the significant potential of hybrid approaches that integrate statistical change point detection with engineering rules to improve the detection of NREs in building energy consumption data. The application of tailored enhancements, particularly Supplementary Rule 1 combined with parameter tuning, yielded notable improvements in reducing false positives, a known weakness of automated statistical detectors [7,15]. These findings are consistent with prior work advocating hybrid, human-in-the-loop methods that balance data-driven adaptability with domain expertise [17,18].

Among the strategies evaluated, Supplementary Rule 1—which leverages rolling slope analysis of the dissimilarity time series—consistently improved specificity. It proved especially effective for lighting retrofits, temporary NREs, and variable load changes, where its filtering mechanism significantly reduced spurious detections. However,

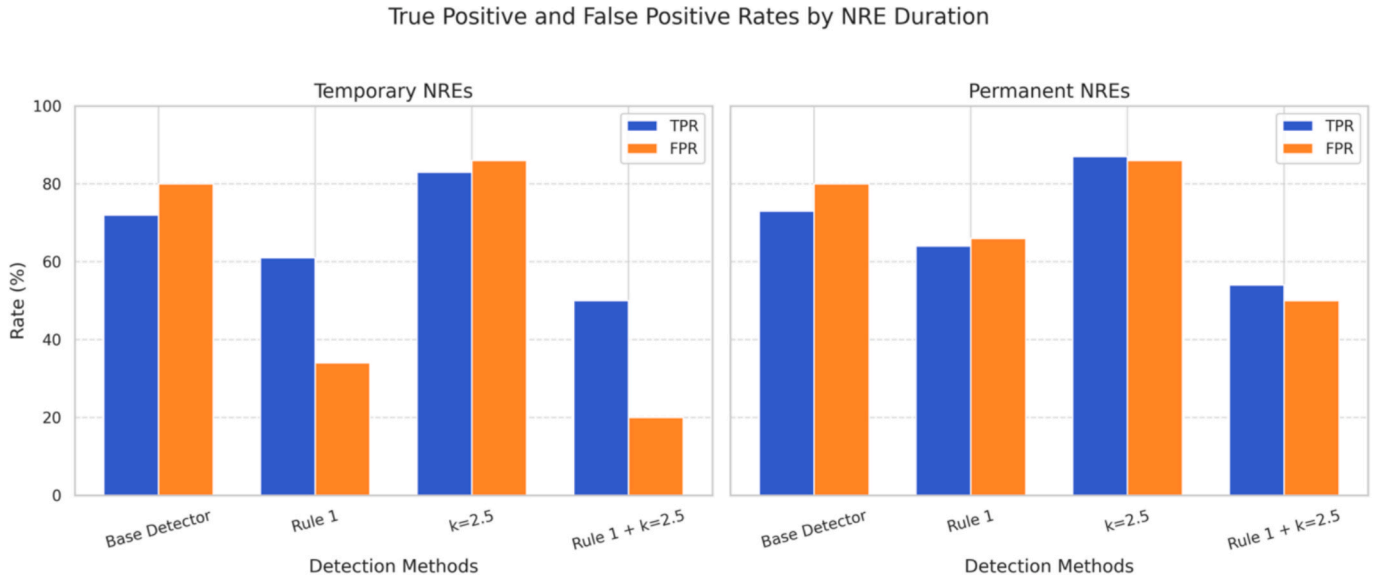


Fig. 12. TP and FP rates by detection method and NRE duration.

this benefit came at the cost of reduced sensitivity. For example, when Rule 1 was applied, the FP rate dropped from 80% to 50%, but the TPR also declined from 73% to 61%. This approximately 2:1 trade-off between false positive reduction and true positive loss underscores the need to align detection configurations with operational priorities—whether the goal is minimizing false alarms or ensuring no events are missed.

Tuning the k parameter to emphasize behavioral proximity ($k = 2.5$) yielded the highest overall TPR, particularly benefiting lighting retrofits—with detection sensitivity rising from 65% to 90%. This was unexpected, as behavioral tuning was initially assumed to be more helpful for cooling retrofits, where changes affect both values and usage patterns. Instead, lighting scenarios saw the greatest gains, suggesting the need to revisit theoretical assumptions about how behavioral weighting affects dissimilarity-based detection. This finding aligns with broader observations in the literature [14,16] that emphasize the importance of tuning algorithms to context-specific dynamics. While initially counterintuitive, a post-hoc analysis of the load shapes suggests a technical explanation for this result. Lighting retrofits often manifest as distinct 'square wave' changes in the daily load profile—sharp, consistent step-changes in start/stop times that are strictly schedule-dependent. This creates a high-fidelity signal in the temporal correlation (CORT) metric.

In contrast, cooling retrofits are weather-dependent and inherently 'noisier,' with load shapes that fluctuate significantly even without NREs. When k was increased to 2.5, the algorithm amplified the weight of the 'clean' behavioral shifts found in lighting data more effectively than the noisy shifts in cooling data. Essentially, the lighting NREs provided a stronger behavioral signal-to-noise ratio for the tuned algorithm to latch onto.

Despite these improvements, no method proved universally superior. The combined approach (Rule $1 + k = 2.5$) achieved the lowest FDR (33%), but also the lowest TPR (50%), indicating an overly conservative trade-off. This again reinforces the central theme: optimizing NRE detection requires case-specific strategies. Detection performance varied substantially by retrofit type and NRE characteristics. For example, lighting retrofits, which shift value more than behavior, saw strong FDR reductions with minimal TPR loss using Rule 1. Conversely, cooling retrofits—introducing complex behavioral dynamics—proved more difficult to detect cleanly, with higher false positives and lower net gains. Similarly, variable load NREs were more easily detected than constant ones, likely due to the stronger dissimilarity signals they produce. Temporary NREs were also more detectable than permanent events, as Rule 1's rolling slope method is inherently more responsive to short-term fluctuations.

While the combined detection method achieved a False Discovery Rate (FDR) of 33%—meaning one in three flagged events may still be a false alarm—this represents a substantial operational improvement over the baseline detector's 80% FDR. In a practical 'M&V 2.0′ workflow, this tool is not intended to function as a fully autonomous decision maker, but as a prioritization filter for practitioners. By effectively suppressing the vast majority of spurious signals, the algorithm allows engineers to focus their limited time on verifying a smaller, high-probability set of potential events, thereby solving the scalability barrier that currently hinders population-level analysis.

The proposed framework addresses this by acting as a triage tool rather than an autonomous decision-maker. By achieving a reduced FDR of 33%, the algorithm allows engineers to focus on a high-probability subset of potential NREs. The detector outputs flagged time periods with associated confidence signals, dissimilarity scores, and context indicators—such as segment energy changes—to assist in expert review. While this study utilized controlled simulation data, real-world variability—including weather, occupancy, and operational noise—may affect detection reliability in practice. Moreover, classifications of NRE magnitude and duration were necessarily heuristic due to the absence of field-verified ground truth. Future research using larger, real-world datasets is required to improve threshold robustness and quantify human-in-the-loop effectiveness in actual deployments.

For operational deployment across large building portfolios, algorithmic efficiency is critical. The baseline detector utilizes the Pruned Exact Linear Time (PELT) algorithm, which offers an exact solution with a linear computational cost, O(n) [21]. The recommended enhancement, Supplementary Rule 1, utilizes a rolling average and slope calculation which also operates in linear time. Consequently, the combined approach maintains the scalability of the baseline detector, capable of processing years of hourly data in fractions of a second. In contrast, the rejected Rule 2 (Mini-Model) required training a new regression model for every potential change point identified. This iterative model fitting introduces significant computational latency. The high computational burden of Rule 2, combined with its poor accuracy in early-period detection, confirms that it is ill-suited for automated, large-scale applications, though it may remain useful for offline, deep-dive forensic analysis.

Finally, the broader implications of this work mirror trend in artificial intelligence research toward hybrid solutions. Combining data-driven changepoint detection with knowledge-based filters offers the benefits of automation while retaining interpretability and contextual sensitivity. These findings support a shift toward modular, adaptive detection systems in M&V—one's that can be tuned to building-specific dynamics and retrofits, enabling more accurate energy savings attribution with minimal manual oversight.

5. Conclusion

Detecting NREs in building energy consumption remains a persistent challenge in advanced M&V practices. As highlighted in prior studies [7,15,19], statistical approaches can automate detection but often struggle with high false positive rates and limited interpretability. To address these limitations, this study builds on the model-agnostic framework developed by Touzani et al. and investigates how supplementary engineering rules and parameter tuning can refine detection performance across a broader and more diverse simulated dataset.

The analysis reveals important trade-offs between TPR and FDR across different detection methods. The choice of method depends on the specific application requirements:

- If minimizing false positives is critical, the combined approach (Rule $1 + k = 2.5$) offers the best performance with a 33% FDR.
- If maximizing true positive detection is the priority, the $k = 2.5$ parameter tuning alone provides the highest rate at 83%.
- For a balanced approach, Rule 1 alone offers a significant reduction in false positives (to 50%) with a moderate impact on true positives (reduced to 61%).

NRE and retrofit characteristics significantly influence detector performance, with lighting retrofits and variable load NREs generally showing better results with Rule 1 enhancement. The unexpected performance of the k parameter tuning suggests that the theoretical understanding of how this parameter affects detection needs refinement. Notably, increasing the behavioral proximity weighting from $k = 1.0$ to $k = 2.5$ was expected to primarily enhance performance in cooling retrofits—where operational behaviors change post-retrofit. However, the largest TPR improvement (from 65% to 90%, a 25-point gain) occurred instead in lighting retrofits, which primarily induce static, value-based shifts. In contrast, cooling retrofits saw only a marginal increase (from 70% to 85%). This counterintuitive result suggests that the current theoretical framework for k's role in distinguishing behavioral vs. value-based NREs may be incomplete and warrants further investigation. These enhancements were evaluated using a significantly expanded and more diverse test dataset, enabling a deeper assessment of algorithm robustness and variability in performance across NRE types, magnitudes, durations, and retrofit contexts. Results show a marked improvement in specificity—most notably, a substantial reduction in

false positives—without fully compromising sensitivity, particularly when optimized parameter-rule combinations were applied. However, while the improvements are meaningful, the current approach still falls short of delivering a fully automated, universally robust solution for NRE detection.

Nevertheless, the refined detection method presents a practical advancement for M&V professionals. With limited human intervention, the optimized approach offers a more interpretable and reliable tool to support advanced M&V analyses. Future research should explore adaptive parameter tuning, real-time deployment in operational settings, and the integration of external contextual data (e.g., weather, occupancy) to further improve detection accuracy and generalizability. Future work could also explore adaptive approaches that select the appropriate detection method based on building characteristics or combine these methods in more sophisticated ways.

CRediT authorship contribution statement

Jessica Granderson: Writing – original draft, Methodology, Investigation, Funding acquisition, Conceptualization. **Samuel Fernandes:** Writing – review & editing, Writing – original draft, Visualization. **Eliot Crowe:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

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.

Acknowledgements

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Program, of the U. S. Department of Energy under Contract no. DE-AC02- 05CH11231. The authors would like to thank Mrinalini Sharma, David Jump and Devan Johnson from kW Engineering for their role in supporting the quantitative analyses.

Appendix

1] Retrofit Type Classification

The NRE detector's accuracy is evaluated across two types of retrofits in this analysis (See Table 3): constant and variable, i.e., constant change in load (independent of outside weather conditions and other factors), and variable change in load (temperature dependent). The detector detects the start and/or the end of an NRE by identifying statistical changepoints in the dissimilarity time series, which is based on the proximity in behavior and values of the adjusted baseline and the post-period energy use time series. Since the dissimilarity time series is directly dependent on the post period's energy consumption, which is affected by the implemented retrofit, we designed this study to examine the effect of the retrofit type on the detector's accuracy.

Table 3
Retrofit type classification key.

Retrofit type	Classification
Reducing Lights Power Density	Constant
Increasing Cooling Equipment Efficiency	Variable

2] NRE Type Classification

The post period time series is also dependent on the presence of NREs, their type (variable or constant), their duration, and their magnitude. NREs were classified into variable and constant based on their impact on energy usage. If the change in energy usage was constant (changes not weather-dependent) as a result of the NRE, the NRE was categorized as constant. Table 4 summarizes NRE types and classification, with further detail below.

Table 4
NRE Type Classification Key.

NRE Type	Classification
Change in Occupancy Start and End Times	Variable
Change in Occupancy Density	Variable
Change in Base Electrical Plug Load	Constant

The test dataset includes three types of NREs, selected from a large pool of parametric simulations that were developed using default parameters, statistical variation inputs, and threshold limits for each parameter:

Change in Occupancy start and end time:

- Default start and end times are 8 am to 5 pm
- Start and end times are picked independently from a truncated normal distribution to simulate this NRE

Table 5
Distribution of Occupancy Start & End Times.

	Start Time	End Time
Mean	8:00 am	5:00 pm
Standard Deviation	2 h	2 h
Minimum	3:00 am	2:00 pm
Maximum	11:00 am	8:00 pm

- If the draw from the normal distribution results in the exact times of 8:00 am – 5:00 pm, another draw is done

Change in Occupancy density:

- The nominal value is either multiplied or divided by a random coefficient picked from truncated normal distribution
- Normal Distribution: Mean of 1.7, standard deviation of 0.3, minimum of 1.2, and no maximum
- Whether the occupant density is increased or decreased is a uniform random parameter

Change in Base Electrical Plug load.

- The nominal value for the baseload is either multiplied or divided by a random coefficient picked from truncated normal distribution
- Normal Distribution: Mean of 1.7 standard deviation of 0.3, minimum of 1.2, and no maximum
- Whether the baseload is increased or decreased is a uniform random parameter

3] NRE Magnitude and Duration Classification

To answer the first research sub-question, the NREs were classified into small and large magnitude NREs based on the parameter values for the specific NREs (See Table 6). All NRE parameters considered in this test dataset are derived from a normal distribution.

The magnitude for change in occupancy start and end time is computed as change in the number of hours from the initial 8am – 5 pm. We have a range of 0–5 for the change in number of hours. A change of 0 means that occupied hours are still 9 h out of the 24, but the start and end times have shifted.

Table 6
Classification of NREs in small and large magnitudes.

	NRE Types			
NRE Magnitude	Change in Occupancy Density	Change in Base Electrical Plug Load	Change in Occupancy Start and End Times (# hours change $\neq 0$)	Change in Occupancy Start and End Times (# hours change $= 0$)
Small	Less than or equal to 1.7	Less than or equal to 1.7	Less than 3 h	Less than 3 h
Large	Greater than 1.7	Greater than 1.7	More than or equal to 3 h	More than or equal to 3 h

Lastly, the NREs were categorized based on their duration in the post-period (See Table 6).

Table 7
Classification of NRE Durations.

NRE Duration Category	NRE Duration
Permanent	Permanent
Temporary – Short	Temporary and lasting less than or equal to 21 days
Temporary – Long	Temporary and lasting more than 21 days

4] Frequency of NREs.

The dataset used in this study has one NRE per simulation. This translates to two changepoints in datasets where the NRE is temporary and one changepoint where the NRE is permanent. The number of changepoints in the simulation datasets has a direct bearing on the TPR of the main and improved versions of the NRE detector: the TPR can only be one of three values: 0%, 50%, or 100%. The 2019 research included two single-day NREs, to simulate one-off unoccupied days; these were not included in the current round of research. While a single, random, unoccupied day is technically an NRE, in practical terms its energy impact would be small and not introduce significant error into an annual energy savings estimate. As such, accurately detecting a single unoccupied day was not considered a suitable measure for the algorithm.

5] Performance summary.

Table 8 summarizes the performance of the detection across the entire set of results.

Table 8
Detection method performance summary.

Detection Method	Mean TPR (%)	Mean FDR (%)	Cooling TPR (%)	Cooling FDR (%)	Lighting TPR (%)	Lighting FDR (%)	Constant TPR (%)	Constant FDR (%)	Variable TPR (%)	Variable FDR (%)	Temporary TPR (%)	Temporary FDR (%)	Permanent TPR (%)	Permanent FDR (%)
Base Detector	75	85	70	85	75	85	75	80	80	85	80	85	85	85
Rule 1	65	60	60	60	65	20	65	55	65	30	70	35	75	65
k = 2.5	90	90	85	90	90	80	90	90	95	80	90	90	95	80
Rule 1 + k = 2.5	55	45	50	45	60	25	50	40	55	30	60	20	65	55

6] Distribution of test cases by Category.

Table 9
Distribution of test cases (N) by category of NRE.

Category group	Sub-Category	Number of cases (N)	Percent of total dataset
Total Dataset	All Cases	180	100%
Retrofit Type	Lighting Retrofit	90	50%
	Cooling Retrofit	90	50%
NRE Magnitude	Small NRE	90	50%
	Large NRE	90	50%
NRE Duration	Temporary (Short + Long)	120	67%
	Permanent	60	33%
NRE Load Type	Variable Load (Occupancy-Driven)	120	67%
	Constant Load (Plug Load)	60	33%

Data availability

Data will be made available on request.

References

[1] California Legislature. (2015). Assembly Bill No. 802, chaptered October 8, 2015 (Chapter 590, Statutes of 2015): Energy efficiency. https://www.leginfo.ca.gov/pub/15-16/bill/asm/ab_0801-0850/ab_802_bill_20151008_chaptered.html.

[2] Missouri Public Service Commission (MPSC). (2025). Order approving EM&V parameters under MEEIA Cycle 4 (DSIM funding). https://efis.psc.mo.gov/Document/Display/77521.

[3] New York Public Service Commission (NYPS). (2018). Order authorizing utility administered energy efficiency portfolio budgets and targets for 2019–2020 (Case 15-M-0255).

[4] Connecticut Department of Energy and Environmental Protection (CDEEP). (2018). Advanced measurement and verification (M&V) pilot in Connecticut. https://portal.ct.gov/DEEP/Energy/A-Pilot—Advanced-Measurement-and-Verification-in-Connecticut.

[5] Lawrence Berkeley National Laboratory (LBNL) (2019). RMV2.0: Open-source tools for measurement and verification. https://lbnl-eta.github.io/RMV2.0/.

[6] k.W. Labs. (2019). nmecr: An R package for normalized metered energy consumption (NMEC) analysis. https://github.com/kW-Labs/nmecr.

[7] S. Touzani, B. Ravache, E. Crowe, J. Granderson, Statistical change detection of building energy consumption: applications to savings estimation, Energy Build. 185 (2019) 123–133, https://doi.org/10.1016/j.enbuild.2018.12.020.

[8] S. Yuan, Z. Hu, J. Lin, Y. Zhang, A framework for the automatic integration and diagnosis of building energy consumption data, Sensors 21 (4) (2021) 1395, https://doi.org/10.3390/s21041395.

[9] Y. Tonooka, J. Liu, Y. Kondou, Y. Ning, O. Fukasawa, Using intelligent data analysis to detect abnormal energy consumption in buildings, Energy Build. 39 (1) (2006) 52–58, https://doi.org/10.1016/j.enbuild.2006.04.011.

[10] G. Fenza, M. Gallo, V. Loia, Drift-aware methodology for anomaly detection in smart grid, IEEE Access 7 (2019) 9645–9657, https://doi.org/10.1109/ACCESS.2019.2891315.

[11] M.-M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, A. Gómez-Expósito, Detection of non-technical losses using smart meter data and supervised learning, IEEE Trans. Smart Grid 10 (3) (2018) 2661–2669, https://doi.org/10.1109/TSG.2018.2807925.

[12] S. Oprea, A. Bâră, F. Puican, I.C. Radu, Anomaly detection with machine learning algorithms and big data in electricity consumption, Sustainability 13 (19) (2021) 10963, https://doi.org/10.3390/su131910963.

[13] Y. Weng, N. Zhang, C. Xia, Multi-agent-based unsupervised detection of energy consumption anomalies on smart campus, IEEE Access 7 (2018) 2169–2181, https://doi.org/10.1109/ACCESS.2018.2886583.

[14] Y. Himeur, K. Ghanem, A. Alsalemi, F. Bensaali, A. Amira, Artificial intelligence-based anomaly detection of energy consumption in buildings: a review, current trends, and new perspectives, Appl. Energy 287 (2021) 116601, https://doi.org/10.1016/j.apenergy.2021.116601.

[15] M.L. Goldberg. (2019). Can the non-routine become routine? Dealing with non-routine events in normalized metered energy consumption analysis. Proceedings of the International Energy Program Evaluation Conference (IEPEC). Denver, CO.

[16] L. Pereira, N. Nunes, Performance evaluation in non-intrusive load monitoring: datasets, metrics, and tools—a review, Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery 8 (6) (2018) e1265.

[17] A. Kovačević, A. Dehghan, M. Filannino, J. Keane, G. Nenadić, Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives, J. Am. Med. Inform. Assoc. 20 (5) (2013) 859–866, https://doi.org/10.1136/amiajnl-2013-001625.

[18] S. Ge, Z. Yu, F. Yang, J. Liu, L. Wang, Human–machine collaboration-based sound event detection, CCF Trans. Pervasive Comput. Interact. 4 (2) (2022) 158–172, https://doi.org/10.1007/s42486-022-00091-9.

[19] H. Fanaee-T, J. Gama, Event labeling combining ensemble detectors and background knowledge, Prog. Artif. Intell. 2 (2) (2013) 113–127, https://doi.org/10.1007/s13748-013-0040-3.

[20] P. Montero, J.A. Vilar, TSclust: an R package for time series clustering, J. Stat. Softw. 62 (1) (2015) 1–29.

[21] Rebecca Killick, Paul Fearnhead, A. Eckley Idris, Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association 107 (500) (2012) 1590–1598, https://doi.org/10.1080/01621459.2012.737745.