Limitations of machine learning for building energy prediction: ASHRAE Great Energy Predictor III Kaggle competition error analysis

Clayton Miller 1∗, Bianca Picchetti 1, Chun Fu 1, Jovan Pantelic 2,3
1Department of the Built Environment, College of Design and Engineering, National University of Singapore (NUS), Singapore
2Department of Biosystems, KU Leuven, Belgium
3Well Living Lab, Delos Living LLC, USA
∗Corresponding Author: clayton@nus.edu.sg, +65 81602452

Abstract
Research is needed to explore the limitations and potential for improvement of machine learning for building energy prediction. With this aim, the ASHRAE Great Energy Predictor III (GEPIII) Kaggle competition was launched in 2019. This effort was the largest building energy meter machine learning competition of its kind, with 4,370 participants who submitted 39,403 predictions. The test data set included two years of hourly whole building readings from 2,380 meters in 1,448 buildings at 16 locations. This paper analyzes the various sources and types of residual model error from an aggregation of the competition’s top 50 solutions. This analysis reveals the limitations for machine learning using the standard model inputs of historical meter, weather, and basic building metadata. The errors are classified according to timeframe, behavior, magnitude, and incidence in single buildings or across a campus. The results show machine learning models have errors within a range of acceptability ($\text{RMSLE}_{\text{scaled}} \leq 0.1$) on 79.1% of the test data. Lower magnitude (in-range) model errors ($0.1 < \text{RMSLE}_{\text{scaled}} \leq 0.3$) occur in 16.1% of the test data. These errors could be remedied using innovative training data from onsite and web-based sources. Higher magnitude (out-of-range) errors ($\text{RMSLE}_{\text{scaled}} > 0.3$) occur in 4.8% of the test data and are unlikely to be accurately predicted.

Keywords: Building energy prediction, Energy model, Error analysis, Machine learning limitations, Kaggle competition, Artificial intelligence

1. Introduction

Building energy prediction using machine learning models has been around for decades and has become important for the evaluation of performance in the context of retrofits [1,2], measurement and verification [3,4], renewable energy integration [5], systems control [6,7], fault detection [8], residential energy use [9], and urban-scale energy modeling [10,11]. A recent text mining-driven review of 30,000 building energy-related data science publications shows that there has been rapid growth in the last ten years in techniques and applications [12]. This rapid expansion of the field is creating a myriad of prominent techniques using deep learning [13,14], embedded online systems [15], sequence learning [16], transfer learning [17,18], neural networks [19,20], bayesian probabilistic forecasting [20], and gradient boosting trees [21,22]. Despite this expansion and innovation, there has been a lack of firm understanding of which techniques, tools, and models are the most accurate, fastest, or easiest-to-use specifically for building energy consumption prediction. Benchmarking in the energy prediction context would help solve this problem by applying various methods on the same large data set, making the comparison of results meaningful [23]. Several initial studies have created environments to apply energy prediction techniques across a large group of buildings [24,25,26,27]. To further expand both the data set and talent pool of machine learning experts, the Great Energy Predictor III (GEPIII) competition was planned and executed in 2019 as a means of crowdsourcing thousands of prediction solutions and determining the most effective in a competitive environment [28]. GEPIII was the largest exercise in benchmarking machine learning methods for this context in terms of size of data set, number of compared techniques, and amount of open-source content shared with the community.

1.1. What are the limitations of machine learning for energy prediction when the best models, techniques, and experts are involved?

GEPIII brought together the combined efforts of 4,370 participants, and the top solutions can be considered near-optimal for the long-term prediction context used. These solutions have empirically found the best combination of preprocessing, feature selection, model types, and post-processing strategies for this scenario. Large ensembles of mainly gradient boosting trees with significant preprocessing of the training data were found to be the best solutions for this application [28]. The
areas of improvement that remain are those caused by the limitations due to the inability to predict behavior that was not observed previously and, therefore, not used to train the machine learning models. The key limitation in this sense is that machine learning models cannot predict behavior that is not contained in the training data. Therefore, the potential is to quantify and classify these situations to help pinpoint new sources of data that can supplement the process and improve performance in those scenarios.

1.2. Previous and related work

Few studies exist that explore the sources of error for building performance prediction. In the physics-based modeling realm, the quantification of error has been heavily focused on the capabilities of the simulation model specification process to capture reality accurately. This field explores uncertainty analysis to quantify the risk that assumptions made in this process are not precise or accurate enough [29]. In data-driven modeling, quantification of modeling error has been explored in the context of measurement and verification [30]. Granderson et al. empirically investigated the ten modeling frameworks against 537 building meters to show that error was within a reasonable range compared to commonly-used standards [24]. This study found that the median coefficient of variation of the root mean squared error (CVRSME) was less than 25% for various training data lengths (6 and 12 months).

Other domains have explored the use of extensive reviews or benchmarking exercises like modeling competitions to quantify error. For example, the medical field has characterized what causes errors for oncology [31] and cardiovascular medicine [32]. There is a comprehensive analysis of categories of limitation for machine learning applied to social objectives [33]. Several studies specifically analyzed the results of a Kaggle competition to characterize modeling error and its lessons [34, 35]. Other studies created their own large-scale benchmarking analysis similar to Kaggle [36].

1.3. Objectives and novelty

To build upon the momentum of the GEPIII competition, this research seeks to quantify and classify the error in behavior that the contestants could not predict even with the most advanced machine learning techniques at their disposal. The winning solutions were restrained from perfectly predicting the ground truth due to limitations in the training data and capabilities of machine learning. This paper builds upon this intent in the following ways:

- The method classifies sources of time-series prediction error for building energy prediction according to magnitude, shape, and number of coincidental occurrences across a collection of buildings.
- The breadth of data and prediction submissions create a scenario in which an error analysis from this competition is the largest of its kind. The results can show the percentage of the test data that belongs to the categories of easy to predict, prediction could be enhanced, or prediction likely impossible for state-of-the-art machine learning methods. These results are generalizable for the building and meter types with a critical mass in this data set (offices, classrooms, laboratories, dormitories, and municipal buildings).
- The error analysis results are used to make practical recommendations for improving general model performance.

Towards these objectives, the paper is structured as follows. First, Section 2 outlines the process of downloading and processing the prediction data from the top 50 contestants and calculating the error for each timestamp. Each prediction data point is then classified according to defined error categories to characterize the frequency and scope of various behaviors. Next in Section 3, these error categories are visualized straightforwardly from some representative sites, and then aggregations are created to show high-level trends. Next, Sections 4 and 5 provide insight into what innovative data sources could be helpful in further reducing the overall prediction error for each category. Finally, limitations to the analysis are discussed, and reproducible code/data are linked for further work.

2. Methodology

This paper aims to characterize the error of the GEPIII competition in a way that enables classification and inference of cause. These insights help understand under- and over-prediction, classify the most frequent and most important errors and identify possible remedies. Thus, the methodology first covers essential details of the competition itself, the extraction of the winning prediction data from Kaggle, and the replication of error calculation. These residuals are then classified according to the nature of the error. For example, did the error occur consistently across a period, or was it intermittent and more stochastic? These error definitions provide a foundation for investigating possible causes of error and the complementary means of addressing those errors using new data.

2.1. Overview of the competition

The Great Energy Predictor III competition was held on the Kaggle platform from October 5 to December 19, 2019. This competition was the next generation of the Great Energy Predictor Shootout I and II that were held in the 1990s [37, 38, 39, 40, 41]. A detailed overview of the GEPIII competition is found in a publication focused on the contestants, the winning solutions, and the fundamental insights [28]. The training data set for the competition included hourly meter data from the year 2016, while the public leaderboard test data set was the year 2017 and the private leaderboard was 2018. The 2016 and 2017 data sets were opened for use in building energy prediction studies as the Building Data Genome 2 (BDG2) project [42]. In addition to the meter training data, the contestants were provided with weather data from training and test data time ranges and various metadata about each building, such as primary use type, building age, system type, and the number of floors.
2.3. Model error categorization

Two types of error categories were created to characterize the reasons for the model error. These categories were created through exploratory analysis of error data as well as logical thresholds based on the usability of the model prediction outputs.

2.3.1. Magnitude and Reach Category

The first category is called the Magnitude and Reach (MR) type, and it relates to the magnitude of the error and its reach amongst the buildings within a site. The concept of reach describes whether the error impacts a single or small set of buildings or a more extensive or majority portion of the buildings from a common site. The MR types are:

- **Type A** - Single Building In-Range Error - Error that occurs on a single or small set of buildings on a site with an $\text{RMSLE}_{\text{scaled}}$ threshold between 0.1 and 0.3.
- **Type B** - Multiple Building In-Range Error - Error Type A that occurs on at least 33% of buildings on a site with an $\text{RMSLE}_{\text{scaled}}$ threshold between 0.1 and 0.3. This threshold was selected as a reasonable line between common behavior happening coincidentally or not.
- **Type C** - Single Building Out-of-Range Error - Error that occurs on a single or small set of buildings on a site with an $\text{RMSLE}_{\text{scaled}}$ threshold greater than 0.3.
- **Type D** - Multiple Building Out-of-Range Error - Error Type C that occurs on at least 33% of buildings on a site with an $\text{RMSLE}_{\text{scaled}}$ threshold greater than 0.3.

The $\text{RMSLE}_{\text{scaled}}$ threshold for the categories were selected to be consistent across all meter types and were based on the quartile and intraquartile ranges for the error distributions. These metrics for each meter type are found in Table 3. Predictions from the competition that falls below the thresholds for these types ($\text{RMSLE}_{\text{scaled}} < 0.1$) is considered a good fit as it falls within a range of relative accuracy.

2.3.2. Temporal Behavior Category

The second category is called the Temporal Behavior (TB) type, and it focuses on the temporal nature of the error, mostly based on the length of time the MR error persisted. The TB types are:

- **Type 1** - Long term error - Long-term error likely caused by some major consistent operational change or long-term meter failure. This error is detected when the $\text{RMSLE}_{\text{scaled}}$ is over a threshold (between 0.1 and 0.3 for in-range errors and over 0.3 for out-of-range errors) for a time period longer than three consecutive days.
- **Type 2** - Medium-term error - Medium-term error with a shorter span than long-term generally from shorter seasonal effects not characterized in the training data. This error is detected when the $\text{RMSLE}_{\text{scaled}}$ is over a threshold

### Table 1: Number of buildings included in the analysis per kind of meter and site.

| Site ID | Electricity | Chilled Water | Steam | Hot Water |
|---------|-------------|---------------|-------|-----------|
| 0       | 0           | 81            | 105   | 105       |
| 1       | 0           | 0             | 0     | 51        |
| 2       | 0           | 0             | 0     | 0         |
| 3       | 274         | 274           | 274   | 274       |
| 4       | 10          | 91            | 91    | 91        |
| 5       | 105         | 89            | 89    | 89        |
| 6       | 30          | 44            | 44    | 44        |
| 7       | 70          | 15            | 15    | 15        |
| 8       | 124         | 70            | 70    | 70        |
| 9       | 124         | 124           | 124   | 124       |
| 10      | 30          | 30            | 30    | 30        |
| 11      | 5           | 5             | 5     | 5         |
| 12      | 36          | 36            | 36    | 36        |
| 13      | 154         | 154           | 154   | 154       |
| 14      | 102         | 102           | 102   | 102       |
| 15      | 0           | 0             | 0     | 0         |

Where $n$ is the total number of observations in the data set, $p_i$ is the meter reading prediction submitted by the contestant, and $a_i$ is the actual meter reading.

To summarise the top fifty submissions, the RMSLE was calculated grouping by meter, building ID, and date: one data set per kind of meter (chilled water, electricity, hot water, and steam) was obtained, containing the RMSLE value calculated from the predictions of the top 50 submissions by building and date. All negative values were removed to calculate this metric. The RMSLE metric was scaled between 0 and 1 using the Min-Max scaler technique for a more appropriate comparison. Thus, $\text{RMSLE}_{\text{scaled}}$ is used for all the analyses presented here.
| Rank | Team                                | Score | Preprocess                           | Features                                                                 | Modeling                                                        | Postprocess                                                                 |
|------|-------------------------------------|-------|--------------------------------------|--------------------------------------------------------------------------|------------------------------------------------------------------|-----------------------------------------------------------------------------|
| 1    | Matthew Motoki and Isamu Yamashita (Isamu and Matt) | 1.231 | Removed anomalies in meter data and imputed missing values in weather data | 28 features. Extensively focused on feature engineering and selected | LightGBM, CatBoost, and MLP models trained on different subsets of the training and public data | Ensembled the model predictions using weighted generalized mean. |
| 2    | Rohan Rao, Anton Isakin, Yangguang Zang, and Oleg Knaub (c.Hats) | 1.232 | Visual analytics and manual inspection | Raw energy meter data, temporal features, building metadata, simple statistical features of weather data. | LightGBM, CatBoost, and Feed-forward Neural Network models trained on different subset of the training set | Weighted mean. (different weights were used for different meter types) |
| 3    | Xavier Capdevielle                   | 1.234 | Eliminated 0s in the same period in the same site | 23 features including raw data, weather, and various meta data | Keras CNN, LightGBM and Catboost | Weighted average. |
| 4    | Jun Yang                            | 1.235 | Deleted outliers during the training phase | 23 features including raw data, aggregate, weather lag features, and target encoding. Features are selected using sub-training set | XGBoost (2-fold, 5-fold) and Light GBM (5-fold) | Ensembled three models. Weights were determined using the leaked data. |
| 5    | Tatsuya Sano, Minoumi Tomioka, and Yuta Kobayashi (mna) | 1.237 | Dropped long streaks of constant values and zero target values. | Target encoding using percentile and proportion and used the weather data temporal features | LightGBM in two steps – identify model parameters on a subset and then train on the whole set for each building. | Weighted average. |
| 9    | MPWARE                              | 1.241 | Remove outlier, Imputation            | Timestamp feature, Holiday feature, Categorical statistic feature | LightGBM, CatBoost, LiteMOR, Neural Network | Ensemble Model. |
| 13   | Tim Yee                             | 1.243 | Remove outlier, Imputation            | Timestamp feature, Categorical statistic feature | LightGBM | Ensemble Model |
| 20   | [ods.ai]/PowerRangers               | 1.244 | Remove outlier, Imputation            | Timestamp feature | LightGBM, Neural Network | Ensemble Model |
| 25   | Georgi Pamakov                      | 1.245 | Remove outlier, Imputation            | Timestamp feature, Holiday feature | LightGBM, Neural Network, L1/L2 regression models | Ensemble Model |
| 46   | Fernando Weillmann                 | 1.248 |                       |                       | LightGBM | Ensemble Model |

Table 2: Overview of the publicly available solutions among the top 50 submissions. Only the top five winning teams were required to provide details of their solutions, but additional teams also publicly shared their solutions as Kaggle notebooks.

| Meter         | Q1     | Q2     | Q3     | IQR    | Q3 + 1.5 x IQR |
|---------------|--------|--------|--------|--------|---------------|
| Chilled Water | 0.03   | 0.06   | 0.13   | 0.10   | 0.30          |
| Electricity   | 0.01   | 0.02   | 0.04   | 0.03   | 0.21          |
| Hot Water     | 0.06   | 0.14   | 0.27   | 0.20   | 0.43          |
| Steam         | 0.03   | 0.06   | 0.14   | 0.11   | 0.31          |

Table 3: Quartiles and interquartile range for the scaled RMSLE in each kind of meter.

- **Type 3 - Short-term error** - Short-term error that spans during only one day and is usually related to non-routine events and momentary meter or system failures. This error is detected when the RMSLE \( \text{scaled} \) is over a threshold (between 0.1 and 0.3 for in-range errors and over 0.3 for out-of-range errors) for a time period longer than one day but shorter than three consecutive days.

- **Type 4 - Modulating error** - Error that modulates relatively rapidly to signify that there are dynamic causes of the error. Defined as error that occurs during a \( w \) (30-day window), and at least a proportion of them (0.1) is a short-term error (Type 3).

Figure 1 illustrates simple examples of these four TB categories from the data set. Not all prediction time stamps will have an MR or TB category assigned to them as much of the test data is considered a good fit if the RMSLE \( \text{scaled} \) is lower than 0.1. For the rest of this paper, the errors will be described with a combination of these two categories in the figures, i.e., \( A1 \) would indicate a prediction that has been assigned to MR Type A and TB Type 1.

### 3. Results

The implementation of the screening analysis resulted in the characterization of error types across the GEPIII competition sites from the test data set. This section outlines those results and creates heat maps that illustrate each error type category’s diversity and behavior patterns. Next, there is an aggregation of those errors at the site, building primary use type, and total error levels.

#### 3.1. Temporal error analysis

The initial implementation of these error type filters enables the characterization of the nature of the residual errors throughout the prediction time frame. The following subsection outlines the implementation of error type filtering from both categories. Each meter type (electricity, chilled water, hot water, and steam) is visualized using a heat map with a color palette of the error type categories for both Magnitude and Reach and Temporal Behavior. Detailed information about the sites and buildings can be found in the BDG2 project and its associated publication [42].

#### 3.1.1. Electrical meters

The first meter type is electricity, with Sites 5, 9, and 14 having an MR or TB category assigned to them as much of the test data is considered a good fit if the RMSLE \( \text{scaled} \) is lower than 0.1. For the rest of this paper, the errors will be described with a combination of these two categories in the figures, i.e.,
Figure 1: Example of the four error types created in this analysis to characterize the temporal behavior of the error: Type 1 is a longer-term error that abruptly starts and has an extended amount of time, Type 2 and 3 are similar, yet for shorter amounts of time, and Type 4 is a fluctuating error that alternates quickly in its goodness-of-fit.

Site 5 shows a significant amount of error that seems to be related to seasonal scheduling across the entire site. There are large clusters of consistent single and multiple building errors (Type A/B) with consistent cut-offs and seem to follow a particular phase-based schedule. Most of these errors are also perceived to be long-term (Type 1). In addition, there are several clusters of out-of-range (Type C/D) errors that overlap with the periods on several buildings. Site 5 is made up mostly of municipal buildings, and many of the buildings are primary and secondary schools that have set schedules where classes are likely in or out-of-session [42].

For sites 9 and 14, there are much fewer errors types found for most of the meters, with the majority being either individual buildings (Type A/C) that have frequent bursts of error (towards the top of the heat maps) or short time frames of errors occurring across most of the buildings on the site at the same time. These sites are university campuses, and their errors seem to show that most electrical consumption behavior is relatively easy to predict except for short and sometimes systematic exceptions [42]. A contribution to this ability to more accurately predict consumption might also be the higher quality of data produced by potentially revenue-generating electricity meters.

3.1.2. Chilled water meters

Figure 4 illustrates the screening heat maps for Sites 9, 13, and 14 for the public test/validation data set. Once again, these three sites were chosen based on their representation of this meter type. The visualization from these three sites quickly shows much more error occurring than with electricity. All three of these sites are university campuses with chilled water plants serving various buildings. Most errors are long-term (Type 1) and come from single buildings (Type A/C). There seems to be some heating and cooling seasonality in the errors, with some of the meters distinctly having more error in the winter and spring months, while other buildings have more error in the summer and fall months. There is less uniformity in the starting and stopping errors across buildings despite large groups of buildings seeming to experience similar errors simultaneously. There are variations in error magnitude, with most errors being in-range (Type A/B) with pockets of out-of-range (Type C/D) sometimes occurring in clusters.

3.1.3. Steam meters

Figure 5 shows the error type behavior for the same three sites (9, 13, and 14) as the chilled water. All three sites show approximately half of the meters have good predictions with few errors. The other half has a significant amount of long-term (Type 1) error, especially during the cooling season (for Sites 9 and 13). The steam system energy consumption is driven by behavior different from the previous year and not by weather differences. These sites are all university campuses, and the control of these systems seems to be more sporadic than electricity and chilled water meters.

3.1.4. Hot water meters

Finally, Figure 6 gives an overview of three representative sites of hot water meters. Sites 1, 10, and 14 are covered in this visualization and are all university campuses with some buildings on a centralized hot water supply network. Initially, it is evident that the number of hot water meters is lower than other types, and they seem to have much more error across a broader range than electricity or chilled water. An interesting situation in Site 10 occurs when the entire first month of the data is entirely out-of-range across all buildings. This situation could be an example of total system failure of either the data collection system or the meters themselves. Site 14 shows consistent error across most buildings, with bands of out-of-range errors across many buildings during similar periods.

3.2. Breakdown of error types across sites and meter types

Figure 7 illustrates the breakdown of Types 1-4 across the Magnitude and Reach categories (Types A-D) for all sites and meter types. The only error present in all sites for electricity meters is the single building in-range error (Type A). Multiple building errors (Types B/D) are only present in Sites 5 and 3, respectively. Out-of-range single building error (Type C) is present in Sites 9, 13, and 14 as the chilled water. All three sites show approximately half of the meters have good predictions with few errors. The other half has a significant amount of long-term (Type 1) error, especially during the cooling season (for Sites 9 and 13). The steam system energy consumption is driven by behavior different from the previous year and not by weather differences. These sites are all university campuses, and the control of these systems seems to be more sporadic than electricity and chilled water meters.
3.3. Aggregated error analysis across sites

Figure 7 shows the breakdown across all sites and meter types. This chart shows the diversity of error types across different locations. These aggregations are where it can first be seen that electricity is better predicted in general than the other meter types. The good fit range for electricity across all sites ranges from 60-99%. This chart also illustrates the RMSLE_{scaled} contribution of each error type. Steam meters have the next best level of performance, with the good fit category occurring 55-70% of the time out of the five sites compared in the graphic. Most of the errors were in the in-range categories (Types A/B), with several of the sites (Sites 6, 7, and 9) also having significant single building out-of-range error (Type C). Chilled water models perform worse on average than electricity and steam, with around 40-85% of the test data for each of the sites falling into the good fit range. The sites with higher error, like Site 10, tend to have much more multi-building and out-of-range errors. Site 9 has the lowest error due to its relatively low amount of out-of-range error. This insight is reinforced from the heat map in Figure 3. Hot water meters performed the worst out of the meter types. For all sites, most errors are in-range, except Site 7, which has out-of-range errors exclusively.

3.4. Aggregated error across meter types

The error analysis can be further aggregated according to the magnitude and site-related error type (Types A-D). Figure 8 shows this breakdown according to its impact on both the frequency of error occurrence and the RMSLE_{scaled} contribu-
Out-of-Range Error
Single Building
A1 A2 A3 A4
Multiple Building
B1 B2 B3 B4
C1 C2 C3 C4
D1 D2 D3 D4
In-Range Error
Good Fit
Building number

Figure 3: Heat map representation of the sixteen categories of error for chilled water prediction across the error type categories for Sites 9, 13, and 14. The x-axis is the time range of the public test/validation (2017) and the y-axis are the meters from that site sorted from top to bottom according to highest to lowest error prevalence.

3.5. Aggregated error across building primary use types

Figure 9 illustrates the Magnitude and Reach errors (Types A-D) as broken down by building primary use type. Most errors come from single buildings (Types A/C), and fewer errors come from cross-building events (Types B/D). In addition, except for some building types with only a few samples (e.g., Technology/Science, Food sales, and Service), almost all building types have good-fit categories covering 76.1 - 100.0% of error frequency.

Regarding the error contribution in terms of RMSLE scaled for the major building types, Entertainment/public Assembly, Office, and Lodging/residential all have a reasonably similar contribution structure, with the majority of the error coming from...
single compared to multiple buildings. Interestingly, Education and Technology/science buildings have a higher proportion of multi-building within-range errors (Type B), accounting for 11.9% and 16.6% of the RMSLE
\textit{scaled}. The possible reason for this situation could be that these buildings are primarily from university campuses and have different energy use behavior during semesters and vacations. In this competition, because the prediction model lacks site-specific schedule features, there are more simultaneous prediction errors for multiple buildings during holidays or break periods.

### 3.6. Overall error breakdown analysis

The highest level aggregation for this paper is shown in Figure 10, which illustrates the breakdown of Magnitude and Reach (MR) categories and a breakdown of each according to the Temporal Behavior categories (Types 1-4). This aggregation gives a high-level understanding of the proportion of different errors from the GEPIII competition. The in-range, single building error (Type A) accumulates to 11.8% of the time frame from the test data set while the multi-building version (Type B) is at 4.3%. This total of 16.1% is the errors from the meters that fall within a range that has the potential to be fixed through alternative data sources or other innovations. Out-of-range error (Types C/D) makes up 4.8% of the total time frame from the test data set. These errors can be considered more extreme in
Figure 6: Aggregations of the Temporal Behaviour (Types 1-4) across the Magnitude and Reach errors (Types A-D), sites, and meter categories. The number of buildings in each meter type category is in parentheses next to the x-axis label. If the bar is missing for a particular error, then that category wasn’t detected for that site.
Figure 7: Aggregations of error types across the meter types (from top to bottom: Electricity, Chilled Water, Hot Water, and Steam) and sites with percentages of in-range (Types A/B), out-of-range (Types C,D) and Good Fit (left) and across all four Magnitude and Reach categories (Types A-D) for each site (right). The number of buildings in each meter type category is in parentheses next to the x-axis label.
Figure 8: Error breakdown (Types A-D) across the meter types including the aggregation of frequency of error types as a percentage of all errors (left) and percentage of contribution to the scaled RMSLE error for the meter across the entire test data set. The number of buildings in each meter type category is in parentheses next to the x-axis label.

Figure 9: Error breakdown (Types A-D) from all meter types across the various primary use types of the buildings which the meters served. The number of buildings in each category is in parentheses next to the y-axis label.
their magnitude and are likely due to unpredictable non-routine events.

4. Discussion

The results of this study provide a foundation for suggested methods to increase the accuracy of building energy prediction. The following sub-sections outline strategies and technologies that address the error type definitions from the results.

4.1. Addressing in-range errors from a single building by incorporating sensors and other new data feeds

The single building in-range error classification (Type A) is an example of when a single or small subset of buildings have long, medium, short, or modulating error behavior that falls within a specific range of magnitude. These periods of less severe error theoretically are caused by phenomena not captured by previous training data related to systematic human behavior at the building level. Addressing these errors is likely a matter of finding other temporal data sources that could be harvested and used in the training/testing data.

For example, carbon dioxide sensors from the building management system could be used as a proxy for occupancy in a building whose energy consumption is sensitive to internal loads from people [43]. This signal could reduce error, especially for A3 and A4 errors that are shorter in length or more modulating in nature. Several previous studies have incorporated such data types for the purpose of subsystem optimization for cooling, heating, and systems optimization [44].

Long and medium-term errors (Types A1/A2) are likely more a result of a system or environmental changes that are not captured in previous training data and are not weather influence. These are situations when zone air temperature setpoints are changed or when systems are turned down or off compared to previous behavior. These errors could be mitigated by incorporating these signals from the building management system (BMS) into the training and test data sets. The characterization of these types of patterns has momentum for systems optimization [45] and model calibration [46] and further work in this direction would benefit multiple fields.

4.2. Addressing in-range errors for multiple buildings from the same location using site-wide data to detect systematic disruptions

The next category of errors to discuss is those where the residual is present across numerous buildings at the same site (Type B). This scenario indicates abnormal energy-consuming behavior at a larger scale, such as district-scale occupancy, climate, or other types of events or failures. These errors are likely also due to either change in human behavior or system control but at a different scale. Similar to the single building errors (Type A), there are short-term and modulating increases in residual due to site-wide human behavior (Types B3/B4). This scaled-up impact is likely due to scheduling scenarios where there is a site-wide change in building use. An example of this situation is when a university campus has a break period or holiday that only exists at that specific site (and did not happen precisely the same way in the training data). Several recent works have utilized various sensors at the campus level for the development of occupancy patterns [47], systems control [48], and indoor air quality prediction [49]. Mitigation of this error might be possible with additional data from online calendars regarding the seasonal changes in the use of buildings. Potential online or third-party data sources that include signals relevant to energy use might be used to mitigate this error, including from social media [50, 51]. For long and medium-term errors (Types B1/B2), these could also result from system-based changes that occur in those time ranges, but this time also at the systems level. Mitigation of these errors could include data from centralized chilled water plants or supervisory control systems that may make changes at the whole site level.

4.3. Non-routine event detection to find out-of-range errors that are unlikely to be fixed

The error categories considered out-of-range (Type C/D) mirror the in-range (Type A/B) in all ways except the magnitude. The magnitude of these categories is such that the predictions diverge from the actual signal in these situations so much that it is unlikely to assume that just adding another training data stream or modifying machine learning parameters would be adequate. A prominent source of this type of error would be building or system sensor failures that cause the meter reading to go to zero or flat-line at a specific value. Other causes of these errors could be drastic changes in the way buildings operate due to renovation, retrofit, or significant system change. A good strategy for dealing with these errors might be to create a change-point detection trigger that essentially shows that the model is not helpful anymore [52]. There is little work in change-point event detection for building performance models, which could be a fruitful direction for enhancing modeling efforts.

4.4. Using occupant-centric operations data in building energy prediction

The subsections in this discussion have explored various sensors, but there are emerging opportunities to include human-related data streams into the energy prediction flow. In the literature, there are only a few instances of training data using occupants counts [53]. Previous work has shown that the way occupants use buildings differs from expected [54, 55] and data from Wifi signals of occupants [56, 57], Bluetooth localization [58, 59, 60], and even text mining [61] can capture signals that could be used as inputs in the modeling process. This type of modeling approach may also influence occupant-centric controls [62]. There is significant potential in characterizing human activity in buildings for numerous research communities, and this direction should be further pursued.

4.5. Limitations

While this analysis is the largest of its kind for building energy prediction, it does not include a critical mass of all building
Figure 10: Highest level overview of errors with the Magnitude and Reach categories (Types A-D) breakdown shown in the middle and the temporal behaviour type breakdowns (Types 1-4) according to in-range (top) and out-of-range (bottom) errors.
types from all climate and geographic contexts worldwide. It can be seen from the open data set released from this competition that five primary use types are included in the data set: Education (classrooms), Offices, Entertainment/Public Assembly, Lodging/Residential, and Public Services. In addition, these buildings are all from North America, the UK, or Ireland. Each of these building use types has enough of a population of buildings so that the results can be considered generalizable for these use types and contexts, but further benchmarking work needs to be done to create a broader analysis from a more extensive and more comprehensive data set. The real challenge in addressing these limitations is the data availability from actual buildings, and building owners now have a low incentive to capture or share data with the research community. This work is a first step in gathering ever-larger and more diverse data sets that the community can use to apply emerging tools and techniques going forward.

Another limitation is the sole use of the RMSLE metric in characterizing the errors in the competition. This metric is used exclusively in this analysis to make a solid link to the competition and the broader machine learning community. The building performance community relies heavily on the coefficient of variation of the root mean square error (CVRMSE) and the mean bias error (MBE). These metrics will be applied in future work using these error data focused on practical applications unrelated to the competition analysis.

5. Conclusions

This paper analyzes the residual errors from the largest machine learning competition ever held for the building energy prediction domain. The results show that 79.1% of the test data set in the competition was predicted with reasonable levels of error (good fit). Still, there was a sizeable limitation in prediction ability for steam and hot water meters and buildings primarily used for technology, science, and food sales. Most of the regions of error were classified as being in-range (16.1% of the test data set) and therefore potentially addressable using alternative data sources as training data beyond just weather and metadata factors. A minority of the test data (4.9%) were found to be from the out-of-range error category, which is considered to be errors greater than what is reasonable to assume can be fixed within the model. These error regions likely need to be addressed using unsupervised change-point models to detect which identify when the signal is so far out of range as reasonable.

In addition to describing the limitations of models in capturing behavior, this analysis also suggested various methods to assist the machine learning community in improving errors in this domain. For example, incorporating alternative data sources from occupancy-related sensors could reduce the error caused by on-demand energy uses in buildings influenced by the variability of occupancy. The addition of district-scale data such as site-specific schedules could be helpful in reducing systematic errors across collections of related buildings. And the act of detecting simply when a model is not relevant any longer due is useful in detecting fundamental shifts in the way a building uses energy, such as major retrofit, disruption, or natural disaster. These insights provide a foundation for future work in building energy prediction using data-driven methods. The limitations of this analysis can help facilitate future machine learning competitions that a) include an even larger and more diverse data set, b) provide a broader range of building use types and locations and c) incentivize building owners to donate further open data sets for community use and benchmarking.

5.1. Reproducibility

Segments of the raw data and analysis code used for this study are available in an open-access Github repository that includes further documentation: https://github.com/buds-lab/ashrae-great-energy-predictor-3-error-analysis

CRediT author statement

Clayton Miller: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Supervision, Project administration, Funding acquisition; Bianca Picchetti: Methodology, Software, Formal analysis, Investigation, Writing - Review & Editing, Visualization; Chun Fu: Formal analysis, Data Curation, Writing - Review & Editing; Jovan Pantelic: Methodology, Writing - Review & Editing.

Funding

The Singapore Ministry of Education (MOE) provided support for the development and implementation of this research through the Temporal Mining of Energy and Indoor Environmental Quality Data from Buildings (R296000181133) Project.

Acknowledgements

This analysis is possible due to the GEPIII competition planning and operations committees. The technical committee to be acknowledged includes (alphabetical order) Anjukan Kathirgamanathan, June Young Park, Pandarasan Arjunan, and Zoltan Nagy. Planning committee members to be acknowledged are Anthony Fontanini, Chris Balbach, Jeff Haberl, and Krishnan Gowri. The ASHRAE organization is recognized for providing support for the competition prize money and the Kaggle platform for hosting GEPIII as a non-profit competition. In addition, the authors would like to thank those who assisted in collecting and releasing the BDG2 data set, including Brodie Hobson, Forrest Meggers, Paul Raftery, and Zixiao Shi.

References

[1] C. Deb, A. Schlueter, Review of data-driven energy modelling techniques for building retrofit. Renewable Sustainable Energy Rev. 144 (2021) 110990.

[2] B. Grillone, S. Danov, A. Sumper, J. Cipriano, G. Mor, A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings, Renewable Sustainable Energy Rev. 131 (2020) 110027.
resolutions: Building energy performance and model calibration, Appl. Energy 286 (2021) 116492.

[47] Y. Ding, W. Chen, S. Wei, F. Yang, An occupancy prediction model for campus buildings based on the diversity of occupancy patterns, Sustainable Cities and Society 64 (2021) 102533.

[48] X. Liu, S. Lee, I. Bilionis, P. Karava, J. Joe, S. A. Sadeghi, A user-interactive system for smart thermal environment control in office buildings, Appl. Energy 298 (2021) 117005.

[49] L. C. Tagliazucchi, F. Re Ceconi, S. Rinaldi, A. L. C. Ciribini, Data driven indoor air quality prediction in educational facilities based on IoT network, Energy Build. 236 (2021) 110782.

[50] X. Liu, F. Feng, Z. Pang, T. Yang, Z. O’Neill, Extracting typical occupancy schedules from social media (TOSSM) and its integration with building energy modeling, Build. Simul. (2021).

[51] C. Fu, C. Miller, Using google trends as a proxy for occupant behavior to predict building energy consumption, Appl. Energy 310 (2022) 118343.

[52] L. C. Tagliazucchi, F. Re Ceconi, S. Rinaldi, A. L. C. Ciribini, Data driven indoor air quality prediction in educational facilities based on IoT network, Energy Build. 236 (2021) 110782.

[53] J. Y. Park, C. Miller, Z. Nagy, A Data-Driven load shape profile based building benchmarking: Comparing doe reference buildings with a large metering dataset, in: Building Simulation Conference Proceedings, pp. Pages 4282–4289.

[54] M. Quintana, P. Arjunan, C. Miller, Islands of misfit buildings: Detecting uncharacteristic electricity use behavior using load shape clustering, Build. Simul. 14 (2021) 119–130.

[55] K. Nweye, Z. Nagy, HVAC scheduling based on Wi-Fi derived occupancy, in: Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys ’20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 340–341.

[56] S. Zhan, A. Chong, Building occupancy and energy consumption: Case studies across building types, Energy and Built Environment 2 (2021) 167–174.

[57] Z. D. Tekler, R. Low, B. Gunay, R. K. Andersen, L. Blessing, A scalable bluetooth low energy approach to identify occupancy patterns and profiles in office spaces, Build. Environ. 171 (2020) 106681.

[58] P. Jayathissa, M. Quintana, M. Abdelrahman, C. Miller, Humans-as-a-Sensor for Buildings—Intensive longitudinal indoor comfort models, Buildings 10 (2020) 174.

[59] M. S. Rahaman, H. Pare, J. Lione, F. D. Salim, Y. Ren, J. Chan, S. Kudo, T. Rowling, A. Sinickas, OccuSpace: Towards a robust occupancy prediction system for activity based workplace, in: 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pp. 415–418.

[60] C. Miller, M. Quintana, J. Glazer, Twenty years of building performance analysis trends: A topic modeling analysis of the Bldg-Sim email list archive, in: Building Simulation Conference Proceedings, pp. Pages 1522–1529.

[61] B. Gunay, Z. Nagy, C. Miller, M. Ouf, B. Dong, Using Occupant-Centric control for commercial HVAC systems, ASHRAE Journal 63 (2021) 30–32.

[62] P. Raftery, M. Keane, A. Costa, Calibrating whole building energy models: Detailed case study using hourly measured data, Energy Build. 43 (2011) 3666–3679.