# Parametric Probability Density Function Characterization of Single Household Peak Loads

Alexander Jordan[1], Yuxiang Chen[1]
[1]University of Alberta, Edmonton, Canada

## Abstract

Residential peak load is an important aspect of residential energy consumption, as the highest volumes of greenhouse gas emissions are released during peak load hours. It's beneficial for decision-makers, designers, and distributors alike to have information and estimates regarding future peak loads that go beyond a point-load prediction. Therefore, this study used Probability Density Functions (PDFs) to characterize residential electrical peak load consumption data to provide a probabilistic outlook for both the frequency and magnitude of future peak loads. Data for PDF fitting was used from an energy monitoring project in Edmonton, Canada that collected high-resolution electrical consumption data sub-metered to different channels. Probability distribution fitting was carried out for six typical PDFs and one mixture model. The goodness of fit for each distribution was visually inspected using quantile-quantile plots (QQ-Plots) and numerically evaluated using the 2-sample Kolmogorov-Smirnov test. The results of this work show that Gaussian Mixture Modelling (GMM) was the best performing model for characterising the probability distribution of peak load behaviour of a single household. It was also shown that while some typical parametric PDFs may capture peak load behaviour at hourly-resolution data, the performance of these models was drastically hindered when the resolution of consumption data was refined to minutely data.

## 1. Introduction

Through the 21st century, energy consumption has remained a key issue on the global stage, with the race to cut greenhouse gas (GHG) emissions consistently driving innovations and research topics. In Canada, the residential sector accounted for 13% of the total GHG emissions released in 2018, corresponding to 17% of the country's energy consumption (NRC, 2018). Within the Canadian power generation sector, many new alternative energy sources are entering the market, notably photovoltaic (PV) technology. Despite new clean energy sources like PV generation, GHG emissions remain relatively high due to the timings of max PV generation and residential peak load not being well aligned (Richardson and Thomson 2013). This study defined residential peak load as the maximum electrical demand that can either be the single maximum power demand at a single moment, a subset of data exclusive to maximum loads, or a set of regularly occurring maximums, such as daily peak loads. Peak load demand within the residential sector is a unique field of study concerned with developing a better understanding of the drivers and consequences of regular peak loads. While some studies focus on methods for reducing peak loads, such as Time-of-Use pricing (Xu, Wang, and Xiao 2019), it's also beneficial to both governments and distributors to have accurate methods for estimating future peak loads for load flow modelling, unit commitment, and reliability planning (Hong and Fan 2016).

Gneiting and Katzfuss (2014) define probabilistic forecasting as predictive probability distributions that are calibrated to describe future values or events. For the long-term prediction of future residential loads, probabilistic load forecasting (PLF) is the overarching key term that encompasses several different techniques that quantify the probabilities of future loads. Within PLF, methods for future load prediction can typically be categorized into either artificial intelligence (AI) based models or statistical-based models (Hong and Fan 2016) (Weron 2006). AI-based predictions typically use models such as artificial neural networks (ANNs), fuzzy logic, and support vector regression (SVR), while some common statistical modelling approaches include semi-parametric additive models and moving average models (Hong and Fan 2016). Statistical models are appealing due to their well-defined mathematical infrastructure and physical interpretations but are limited in their ability to model non-linear behaviours such as load volatility (Weron 2006). While PLF is used for forecasting future values, peak load forecasting requires a unique approach through extreme value statistics, as it only concerns the maximum power consumption. Another categorization of PLF models is the basis of the model, which can be categorized as either data-driven approaches (black-box), theoretically driven (white-box) approaches, or hybrid approaches driven through both theoretical equations and data (grey-box) (Dong et al. 2016). With the rise of residential metering technology, data-driven approaches for forecasting residential consumption are growing in popularity due to the growing accessibility to large volumes of well-refined data. For these newly collected sets of big data, the development of advanced statistical models requires a comprehensive understanding of the underlying

equations and distributions to quantify the probabilities and uncertainties of different outcomes.

A common form of probabilistic characterization of a univariate dataset is the 2D probability density function (PDF), which models the frequency and magnitudes of different values occurring within the set. PDFs are visualized through histograms, showing the frequency on the y-axis for different spans of data on the x-axis. Parametric PDFs represent a family of density functions that use parameters within their equation which are adjustable so that the function can change scale and shape to fit the data distribution. Previous studies have used Weibull, Normal, Lognormal, Gamma, Beta, and Generalized Extreme Value (GEV) functions for fitting electrical consumption data (Xu, Wang, and Tang 2019) (Munkhammar, Widén, and Rydén 2014) (Chesser et al. 2021) (Carpaneto and Chicco 2008) (Jang et al. 2020). Xu, Wang, and Tang (2019) used Lognormal, Normal, Gamma, and Weibull distributions to fit peak electrical abnormal differential loads and found that the Lognormal test had the best performance. A separate study fitted Weibull and Lognormal distributions to average electricity load data for a set of houses in Sweden for each hour of the day, each day of the week, and each month of the year resulting in 4032 PDFs fit, with the Weibull distribution being accepted as a fitting distribution more often than Lognormal (Munkhammar, Widén, and Rydén 2014). This somewhat contrasted Carpaneto and Chicco (2018), who used collected survey data to generate load profiles for several numbers of consumers and found that for 20 modelled homes, only Gamma and Lognormal distributions exhibited acceptable fits. PDF fitting has also been used to assess the impact of retrofitting old homes with electric heat pumps (Chesser et al. 2021). Specifically for peak load forecasting. Jang et al. (2020) used a non-homogenous GEV function to model daily peak demands where the functions were fit to future predictions for temperature and population within a power system incorporating the effects of climate change. Another parametric modelling technique is Gaussian mixture modelling (GMM), which uses a weighted summation of multiple Gaussian distributions. It was shown that a more complex Tracy-Widom distribution was better for modelling daily peak power values for generated electricity load profiles developed using an online tool that incorporates different socioeconomic and occupant descriptive inputs (Shiwen et al. 2020), however, both models exhibited acceptable results. A summarized flowchart of the approach can be seen below in Figure 1, where blue boxes represent approaches outside the scope of this study.

For the PDF fitting of electrical data, there is variation between studies for which PDFs performed the best for fitting residential peak load data. This could be due to the

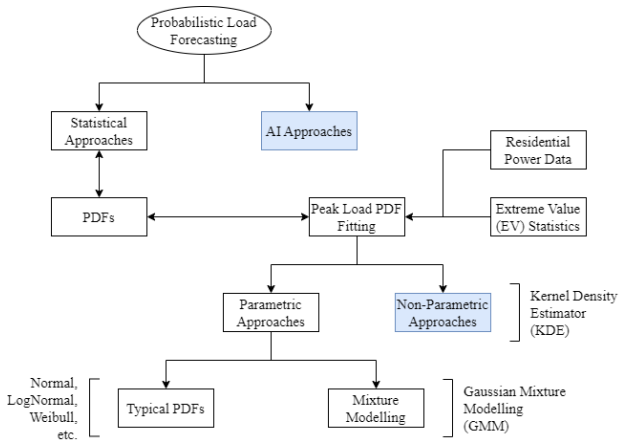


*Figure 1 Summarized Approach Flowchart*

differences that arise from dataset collection schemes and from the differences that arise between regions across the globe. Another gap existing in literature for the PDF fitting of peak residential load is the lack of high-resolution data. It's been shown that for peak load analysis, many of the occupant activities leading to residential peak loads are high-intensity and short duration, such that data resolutions larger than 1-2 minutes lose some of the characteristics exhibited by peak load (Wright and Firth 2007). Additionally, not many studies have compared the differences between the PDF fitting of the total consumption of a household compared to sub-metered consumptions, such as investigating only the heating consumption. Therefore, the research objectives of this paper are to (1) determine the suitability of several parametric PDFs identified from literature for characterizing peak load behaviour for different time resolutions (minutely and hourly), (2) test different methods for segmenting data to increase the compatibility of peak load data to distribution fitting, and (3) identify the ability of different distributions to model total peak loads, heating and ventilation loads, and occupant behaviour peak loads.

The remainder of this paper is organized as follows: Section 2 details the methods used for data collection, segmentation, fitting, and evaluation. Section 3 shows the results of the distribution fitting. Section 4 explores the findings of the study and offers insights into the performance of different distributions. Section 5 summarizes the results of the study and provides limitations and possible directions for future studies.

## 2. Methods

This section details the methods used for the fitting of different parametric probability density functions (PDFs) and Gaussian mixture models (GMMs) to residential electricity consumption profiles. The first step was to retrieve data, which was taken from an energy monitoring project covered in Section 2.1. To allow for focus on peak load behaviour, alternative datasets dissected to only include peak load behaviour were formed in Section 2.2, and

fitting was conducted using the algorithms discussed in Section 2.3. To evaluate the goodness of each fit, validation techniques and error metrics were selected and described in Section 2.4.

### 2.1 Data Collection

To properly fit a PDF to the peak load behaviour of a household, enough power consumption data is required such that the dataset fully describes the patterns exhibited through extreme load periods. For the dataset to achieve this representation, it needs to include samples comprised of all unique external factors relevant (weather conditions, weekdays vs weekends, etc.), but also needs to be resolved enough to properly represent peak values. Having high-resolution data is essential to the modelling of peak load behaviour, as many electrical appliances in a household are high-intensity, low-frequency devices that may only be used for a few minutes within a day (Wright and Firth 2007). The length of time monitored should at the least encompass all regularly occurring patterns that the household is subjected to, mainly seasonality. While this mainly helps to describe the yearly heating and cooling profiles for the household, larger datasets also provide the advantage of providing diversity to the aggregated load profile, as individual household electricity consumption has been shown to be non-normal and nonlinear (Knowles, Hostetler, and Liebovitch 2018). For generalization purposes, having datasets collected from different building types, styles, and ages would aid in explaining the variation between different designs, while having data collected from multiple of the same building design offers insights into variations that arise within the same system design. This study focuses on the latter, which creates an opportunity to focus on the comparison of fitting models for the same design, allowing for direct interpretation of the results of each household.

The project case study data used in this paper is from a newly constructed Multi-Unit Residential Building (MURB) consisting of 6 conjoined units in a single building block. The household units monitored had three levels on the interior and can be categorized as townhome style in design. The project units were located in Edmonton, Canada, which is categorized as a cold climate. The data collection period lasted for three years, starting in March 2020, and ending in February 2023.

The sensor equipment used to collect data from the households included sensors for measuring both electric power consumption from the electrical panel and natural gas consumption from the gas meter. The electrical panel was sub-metered to include channels for the total, air-source heat pump (ASHP), air handling unit (AHU), and mechanical room power (mainly a heat recovery ventilator and a water pump) draws measured in Watts. From the sub-metered channels, a heating, ventilation, and air conditioning (HVAC) consumption profile was developed by summing the ASHP, AHU, and mechanical room values at each time step. This system also indirectly monitored the plug load consumption from the occupants, defined as the total load vector subtracting the HVAC vector. The data collection resolution was once every minute, and the online cloud converted this data to be available in both hourly and daily data files as well.

### 2.2 Extreme Value (Peak) Segmentation

For the characterization of the peak load behaviour of a single household, multiple approaches can be taken for fitting PDFs to this specific section of data. Based on previous literature, it was unlikely that typical parametric PDFs would accurately describe the entire dataset, however, this was still attempted to provide a comprehensive analysis. The main weakness of fitting the entire dataset with a single PDF is that accurately capturing peak load behaviour requires the right tail-end of the distribution to be highly accurate. While this is possible, it adds additional complexities that provide more information than needed. A simpler, more accurate method of modelling peak load behaviour is to only fit a subset of the entire set that captures the behaviour of the peak. One method used for generating subsets of data is to only investigate the single peak values observed each day (Jang et al. 2020) (Shiwen et al. 2020). Another approach involves utilizing the concept that power demand histograms of a single household exhibit multimodal behaviour (Shiwen et al. 2020), so if a cutoff point could be determined to simplify the data set to a singularly occurring highest frequency of power consumption (a single mode), this would create a dataset that could be more compatible with typical single mode PDFs. Since the underlying data distributions vary between units, a generalized top 1% cutoff was selected for analysis, where the generated subset of data removed all values in the bottom 99% of consumption values. These subsets of data were generated for the total, HVAC, and plug load consumption profiles at both minutely and hourly resolutions.

### 2.3 Parametric PDF Fitting

For fitting distribution functions to each dataset identified in the previous section, a statistics coding package was used. From literature, it was found that for electricity modelling, Normal, Lognormal, Weibull, and Gamma distributions are commonly used in the field, so these four PDFs were primarily selected. To test normalized data sets, the Beta distribution was also added to the equation pool due to its convenient fitting between 0 and 1. Based on the nature of this study regarding peak loads, the GEV distribution was also selected as a potential distribution for fitting. For PDF fitting of these typical single-mode functions, the coding package used maximum likelihood estimation (MLE) for each distribution to solve for the unknown parameters that resulted in the best fit to the input dataset.

Another type of fitting used in this study is GMM which utilizes a combination of normal distributions to fit a complex PDF. The modelling equation for GMM is shown below in Equation 1 (Van Dat, Van Toan, and Thanh 2022):

$$p(x) = \sum_{i=1}^{k} \alpha_i N(x|\mu_i, \sigma_i) \quad (1)$$

Where p(x) is the probability density at x, α is the weight of each normal distribution, and μ and σ are the scaling parameters of each normal distribution.

An example of GMM fitting is the use of two normal distributions to fit a dataset that has two identifiable modes, which is known as a bimodal distribution. For a dataset with an unknown number of nodes, an iterative loop can be used to increase the number of modes used until the function fits the data. Typically, GMM is widely used in clustering algorithms as the clusters are represented through individual normal distributions within the summation. As a fitting technique, the fit of the final function is the only metric evaluated, so the makeup composition of the internal distributions is considered irrelevant. The benefit of GMM is that it still qualifies as a parametric function but allows for the ability to model more complex PDFs without the need for the complexities involved with non-parametric approaches. GMM fitting was carried out using pre-defined functions within the coding package that allowed for k distributions to be fit to the input dataset. The GMM function used Expectation-Maximization (EM) to solve for the parameters using a random sample initialization. Due to the random samples causing variation between initializations, 25 functions were fit for each k number of distributions, except for k = 25, where only 10 distributions were fit to relieve computational costs.

### 2.4 Evaluation and Validation Criteria

Two evaluation methods were used for determining the PDF fitting quality to the load datasets: quantile-quantile plots (QQ plots) and the Kolmogorov-Smirnov test statistic (KS-test). The first method, QQ plotting, provides a visual tool for comparing two separate sets of univariate data to estimate if they share a probability distribution (Das and Resnick 2008). This is done through plotting the quantiles of one set of data on the x-axis, and the quantiles of the other distribution on the y-axis. If the datasets had a shared density function, the expected output would be a linear 45-degree line. For example, if the x-axis dataset and y-axis dataset came from similar distributions, it's expected when the 10% quantile of one dataset is reached, the corresponding comparison dataset will have also reached its 10% quantile. A sample QQ-Plot with a poorly fit distribution is shown in Figure 2.

While QQ plots provide a quick, simplistic method for visually rejecting probability functions that have a poor fit to the dataset, a measurement statistic is still needed to assess the goodness of fit. The 2 sample Kolmogorov-Smirnov test (KS-test) for goodness of fit is a widely used statistical test for measuring the fit of a univariate dataset to a PDF dataset that uses the maximum absolute difference between two observed cumulative step functions (Massey Jr 1951). KS-testing was carried out using a function from the coding package that inputs two univariate data vectors and outputs the null hypothesis test result, which compares the p-value test statistic against a significance level of alpha, which by default is 0.05 (which was used as the level of significance for this study). This test algorithm is very useful as it's able to compare different sets of univariate data of different lengths through a nonparametric test, which means that no assumptions or information is required regarding the underlying true distribution of the samples. Since the KS-test function requires two sets of data, inverse transform sampling was used to generate a dummy dataset of n = 1000000 points that was representative of each fitted PDF which could then be used for comparison within the KS-test function.

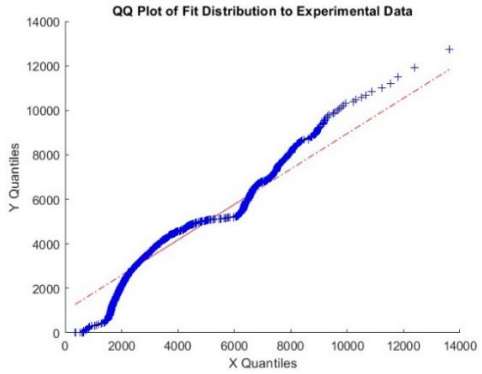


*Figure 2 Sample QQ-Plot of a poor fit between distributions.*

## 3. Results

The objective of this study was to determine the suitability of different parametric PDFs for the characterisation of different categories of electrical loads during peak consumption hours. This was carried out for a variety of typical PDFs as well as through GMM for minutely and hourly data resolutions. Correspondingly, the results of distribution fitting for hourly data are presented in Section 3.1, while those for minutely resolution data are in Section 3.2.

### 3.1 Hourly Data Fitting Results

For hourly data fitting, the input data consisted of energy consumed during each hourly interval measured in kWh. The hourly results for typical parametric distribution fitting can be seen in Table 1 for total consumption, Table 2 for HVAC consumption, and Table 3 for plug consumption. As seen in Table 1 for total consumption fitting, the GEV distribution accurately captured the top 1% of total consumption for five of the six units and also had good performance for the daily peak consumption for three of the six units. The Weibull function performed the second best, which well described the daily peaks of four of six units. Except for the Gamma function that successfully fit the daily peaks of unit 3, the rest of the tested parametric PDFs failed to model the hourly total datasets. For HVAC fitting in Table 2, only the top 1% segmented data were

successfully fit for units 2 and 6 using the Lognormal and GEV equations. The corresponding average p-values from the KS-test for unit 6 were 0.052 and 0.090 for the Lognormal and GEV functions, respectively, indicating these functions did not pass the test with a large margin. The fit for the top 1% of HVAC for unit 2 had a p-value of 0.562, which showed a surprisingly good fit. Table 3 shows for plug behaviour that the GEV function consistently fit the top 1% data for five of the six units, and unit 3's daily peaks dataset was successfully fit by both Weibull and GEV distributions. For all the hourly fit tests, none of the full datasets were fit through the tested PDFs, and the Normal and Beta functions failed to fit any of the datasets.

The other type of parametric PDFs fitted were GMMs, which consisted of k number of summed weighted Gaussian functions. Table 4, Table 5, and Table 6 show the percentage of KS-tests passed for k number of distributions fit for total, HVAC, and plug consumptions, respectively. Bold values in tables indicate only one fit was performed instead of taking an average. This was done in cases where the EM algorithm struggled to converge within 100 iterations, and blank cells indicate 3 successive convergence failures. Immediately, it's observed that the GMMs are more successful at fitting a wider variety of consumption data compared to typical PDFs. For the total consumption data in Table 4, GMMs reached a 100% pass rate at k=25

*Table 1 Percentage (%) of KS-Tests passed for Typical PDFs for **Total** consumption datasets (Hourly).*

| Unit # (Dataset) | Weibull | Gamma | Normal | LogNorm | GEV | Beta |
|---|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 1 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 1 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 2 (dailypeaks) | 100 | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 3 (dailypeaks) | 100 | 100 | 0 | 0 | 100 | 0 |
| Unit 4 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 4 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 4 (dailypeaks) | 100 | 0 | 0 | 0 | 100 | 0 |
| Unit 5 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 5 (dailypeaks) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 6 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (top 1%) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (dailypeaks) | 100 | 0 | 0 | 0 | 100 | 0 |

*Table 2 Percentage (%) of KS-Tests passed for Typical PDFs for **HVAC** consumption datasets (Hourly).*

| Unit (Dataset) | Weibull | Gamma | Normal | LogNorm | GEV | Beta |
|---|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 1 (top 1%) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 1 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 2 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (top 1%) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 4 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 4 (top 1%) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 4 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (top 1%) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (top 1%) | 0 | 0 | 0 | 100 | 100 | 0 |
| Unit 6 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |

*Table 3 Percentage (%) of KS-Tests passed for Typical PDFs for **Plug** consumption datasets (Hourly).*

| Unit (Dataset) | Weibull | Gamma | Normal | LogNorm | GEV | Beta |
|---|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 1 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 1 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 2 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 3 (dailypeaks) | 100 | 0 | 0 | 0 | 100 | 0 |
| Unit 4 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 4 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 4 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (top 1%) | 0 | 0 | 0 | 0 | 100 | 0 |
| Unit 5 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (full) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (top 1%) | 0 | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (dailypeaks) | 0 | 0 | 0 | 0 | 0 | 0 |

*Table 4 Percentage (%) of KS-Tests passed for k distributions for GMM fits of **Total** consumption (Hourly)*

| Unit (Dataset) | k = 1 | k = 2 | k = 5 | k = 10 | k = 25 |
|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 64 | 100 | 100 |
| Unit 1 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 1 (dailypeaks) | 0 | 96 | 100 | 100 | 100 |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 2 (dailypeaks) | 0 | 96 | 100 | 100 | 100 |
| Unit 3 (full) | 0 | 0 | 0 | 0 | 40 |
| Unit 3 (top 1%) | 0 | 100 | **100** | | |
| Unit 3 (dailypeaks) | 0 | 100 | 100 | 100 | 100 |
| Unit 4 (full) | 0 | 0 | 0 | 96 | 100 |
| Unit 4 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 4 (dailypeaks) | 0 | 100 | 100 | 100 | 100 |
| Unit 5 (full) | 0 | 0 | 0 | 24 | 80 |
| Unit 5 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 5 (dailypeaks) | 0 | 96 | 100 | 100 | 100 |
| Unit 6 (full) | 0 | 0 | 16 | 100 | 100 |
| Unit 6 (top 1%) | 0 | 100 | **100** | **100** | |
| Unit 6 (dailypeaks) | 0 | 100 | 100 | 100 | 100 |

for each unit's top 1% and daily peak consumption values. For the full datasets, there was some success except for unit 2, where given a higher k-value it's expected that most of the GMMs would reach a 100% pass rate. For HVAC and plug consumption, it's seen that GMMs fit each unit's top 1% and daily peaks segmented datasets from both categories. Interestingly, for plug consumption, the full unit datasets were fit for four of the six units, while in no cases did the GMMs fit the full HVAC consumption datasets. From Table 4 and Table 6, GMM convergence to the dataset occurs typically after k=2 for the top 1% and daily peaks, which can be interpreted as the data being well described as having 2 main nodes.

*Table 5 Percentage (%) of KS-Tests passed for k distributions for GMM fits of **HVAC** consumption (Hourly)*

| Unit (Dataset) | k = 1 | k = 2 | k = 5 | k = 10 | k = 25 |
|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 1 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 1 (dailypeaks) | 0 | 100 | 100 | 100 | 100 |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 2 (dailypeaks) | 0 | 44 | 16 | 96 | 100 |
| Unit 3 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (top 1%) | 0 | 0 | 96 | **100** | **100** |
| Unit 3 (dailypeaks) | 0 | 0 | 24 | 80 | 100 |
| Unit 4 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 4 (top 1%) | 0 | 80 | 100 | 100 | 100 |
| Unit 4 (dailypeaks) | 0 | 0 | 100 | 100 | 100 |
| Unit 5 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (top 1%) | 0 | 92 | 100 | 100 | 100 |
| Unit 5 (dailypeaks) | 0 | 0 | 12 | 76 | 100 |
| Unit 6 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 6 (dailypeaks) | 0 | 0 | 4 | 64 | 100 |

*Table 6 Percentage (%) of KS-Tests passed for k distributions for GMM fits of **Plug** consumption (Hourly)*

| Unit (Dataset) | k = 1 | k = 2 | k = 5 | k = 10 | k = 25 |
|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 0 | 20 | 92 |
| Unit 1 (top 1%) | 0 | 100 | 100 | 100 | |
| Unit 1 (dailypeaks) | 0 | 100 | 100 | 100 | 100 |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 2 (dailypeaks) | 0 | 100 | 100 | 100 | 100 |
| Unit 3 (full) | 0 | 0 | 0 | 24 | 100 |
| Unit 3 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 3 (dailypeaks) | 0 | 100 | 100 | 100 | 100 |
| Unit 4 (full) | 0 | 0 | 0 | 0 | 40 |
| Unit 4 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 4 (dailypeaks) | 0 | 96 | 100 | 100 | 100 |
| Unit 5 (full) | 0 | 0 | 68 | 100 | 100 |
| Unit 5 (top 1%) | 0 | 100 | 100 | 100 | 100 |
| Unit 5 (dailypeaks) | 0 | 100 | 100 | 100 | 100 |
| Unit 6 (full) | 0 | 0 | 20 | 96 | 100 |
| Unit 6 (top 1%) | 0 | 0 | 100 | 100 | 100 |
| Unit 6 (dailypeaks) | 0 | 0 | 100 | 100 | 100 |

### 3.2 Minutely Data Fitting Results

Regarding the fitting of minutely data through typical PDFs, the same process was carried out as done for hourly data, however, minutely data represented the recorded power draw in Watts each minute, instead of recording the energy consumed in kWh. For typical distribution fitting of minutely data, it was found that Weibull, Gamma, Normal, Lognormal, GEV, and Beta distributions were unsuccessful at characterising any of the minutely datasets. Zero percent of KS-tests were passed for any segmentation of data for each category of total, HVAC, and plug consumptions. The highest average p-value recorded from KS-testing was for unit 3's plug load daily peaks at p = 0.02, which shows that there was a minor resemblance outside statistical significance.

Contrasting the results of typical PDF functions, GMM-derived functions showed promise of fitting for several minutely data categories. The percentage of KS-tests passed from the minutely GMM fitting process of k distributions are shown in Table 7, Table 8, and Table 9 for total, HVAC, and plug consumptions, respectively. For each of the three categories of consumption data, it was found that GMMs were able to successfully capture the behaviour of each unit's top 1% and daily peak consumption data. Each of these segmented datasets was accurately modelled within 5 to 25 distributions used, with none of them exhibiting modal or bimodal behaviour. It was noted that for HVAC GMM fitting in Table 8, there was difficulty in obtaining convergence for multiple units, but for each difficult case, at least one number of k distributions reached a 100% KS-test pass rate before difficulties arose. Despite the positive results obtained for segmented peak load data, GMM distributions were unable to characterize the full consumption data sets for any of the categories.

*Table 7 Percentage (%) of KS-Tests passed for k distribution GMM fits of **Total** consumption (Minutely)*

| Unit (Dataset) | k = 1 | k = 2 | k = 5 | k = 10 | k = 25 |
|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 1 (top 1%) | 0 | 0 | 30 | 100 | 100 |
| Unit 1 (dailypeaks) | 0 | 0 | 100 | 100 | 100 |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 0 | 40 | 80 | 100 |
| Unit 2 (dailypeaks) | 0 | 0 | 90 | 100 | 100 |
| Unit 3 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (top 1%) | 0 | 0 | 0 | 100 | 100 |
| Unit 3 (dailypeaks) | 0 | 90 | 100 | 100 | 100 |
| Unit 4 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 4 (top 1%) | 0 | 0 | 0 | 100 | 100 |
| Unit 4 (dailypeaks) | 0 | 0 | 100 | 100 | 100 |
| Unit 5 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (top 1%) | 0 | 0 | 0 | 70 | 100 |
| Unit 5 (dailypeaks) | 0 | 0 | 60 | 100 | 100 |
| Unit 6 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (top 1%) | 0 | 0 | 50 | 100 | 100 |
| Unit 6 (dailypeaks) | 0 | 30 | 100 | 100 | 100 |

*Table 8 Percentage (%) of KS-Tests passed for k distribution GMM fits of **HVAC** consumption (Minutely)*

| Unit (Dataset) | k = 1 | k = 2 | k = 5 | k = 10 | k = 25 |
|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 0 | | **0** |
| Unit 1 (top 1%) | 0 | 0 | 50 | 30 | 100 |
| Unit 1 (dailypeaks) | 0 | 0 | 40 | **100** | **100** |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 0 | 50 | 100 | 100 |
| Unit 2 (dailypeaks) | 0 | 0 | 100 | **100** | **100** |
| Unit 3 (full) | 0 | 0 | **0** | **0** | **0** |
| Unit 3 (top 1%) | 0 | 0 | 0 | 90 | 100 |
| Unit 3 (dailypeaks) | 0 | 0 | 50 | 100 | 100 |
| Unit 4 (full) | 0 | 0 | **0** | | |
| Unit 4 (top 1%) | 0 | 0 | 90 | 100 | 100 |
| Unit 4 (dailypeaks) | 0 | 0 | 10 | 100 | **100** |
| Unit 5 (full) | 0 | 0 | 0 | 0 | **0** |
| Unit 5 (top 1%) | 0 | 0 | 0 | 80 | 100 |
| Unit 5 (dailypeaks) | 0 | 0 | 0 | 100 | **100** |
| Unit 6 (full) | 0 | 0 | 0 | 0 | **0** |
| Unit 6 (top 1%) | 0 | 0 | 30 | 100 | 100 |
| Unit 6 (dailypeaks) | 0 | 0 | 100 | 100 | 100 |

*Table 9 Percentage (%) of KS-Tests passed for k distribution GMM fits of **Plug** consumption (Minutely)*

| Unit (Dataset) | k = 1 | k = 2 | k = 5 | k = 10 | k = 25 |
|---|---|---|---|---|---|
| Unit 1 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 1 (top 1%) | 0 | 0 | 0 | 100 | 100 |
| Unit 1 (dailypeaks) | 0 | 0 | 70 | 100 | 100 |
| Unit 2 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 2 (top 1%) | 0 | 0 | 100 | 100 | 100 |
| Unit 2 (dailypeaks) | 0 | 0 | 100 | 100 | 100 |
| Unit 3 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 3 (top 1%) | 0 | 0 | 10 | 100 | 100 |
| Unit 3 (dailypeaks) | 0 | 0 | 10 | 30 | 100 |
| Unit 4 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 4 (top 1%) | 0 | 0 | 100 | 100 | 100 |
| Unit 4 (dailypeaks) | 0 | 0 | 100 | 100 | 100 |
| Unit 5 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 5 (top 1%) | 0 | 0 | 0 | 80 | 100 |
| Unit 5 (dailypeaks) | 0 | 0 | 60 | 100 | 100 |
| Unit 6 (full) | 0 | 0 | 0 | 0 | 0 |
| Unit 6 (top 1%) | 0 | 0 | 60 | 100 | 100 |
| Unit 6 (dailypeaks) | 0 | 0 | 100 | 100 | 100 |

## 4. Discussions

From each of the fitting methods performed in this study, it was found that the GEV distribution was the best performing typical parametric PDF tested, closely followed by the Weibull distribution. For hourly data fitting, GEV had good fitting results for the top 1% of the data for total and plug load consumptions. There was also success seen for GEV fitting of half of the units' total consumption daily peaks. Alternatively, the Weibull function was able to capture the density functions of peak load data, which aligns with the literature review, where the Weibull function was noted as having a good performance at the characterisation of peak load consumption for multiple houses. Therefore, for the modelling of a single household, the performance is expected to decrease, which was observed by only four of the six unit's total daily peak consumptions being fit by the Weibull distribution. There were also some surprise fits for select data files observed throughout the study. This was seen for the total consumption daily peaks of unit 3 by the Gamma function and the top 1% of HVAC data in unit 6. These fits were likely the results of the random samples happening to take shapes characterised through these functions, but these results were not consistent. Therefore, from the pool of selected typical parametric PDF functions tested, it was determined that the GEV and Weibull distributions exhibited the best performance.

For mixture modelling, GMM outperformed any of the typical PDF functions tested for hourly data and was able to characterise many of the full datasets from total and plug load consumptions from each unit. This extended to fitting for each unit's top 1% and daily peak consumptions in each consumption category. Regarding the fitting of minutely data, it was found that none of the typical parametric PDFs were able to fit the minutely datasets for power consumption. This was expected from literature, as load volatility is nonlinear and multimodal, as reinforced by the results of this study. Despite this, GMM distribution fitting exhibited good fitting for the segmented datasets introduced in this study. The GMM functions successfully fit all units for each of the electrical consumption categories tested with 100% of the distributions passing the KS-test. This shows the power of GMM as a tool for modelling the behaviour of residential electrical peak load data over typical PDFs.

It's also highlighted in this study that there is a stark difference in the fitting results between hourly and minutely resolution data. Despite good results from some typical distributions at an hourly resolution, these results did not follow through for more refined data. Notably, the performance of GEV and Weibull distributions from the typical PDF pool tested drastically decreased for minutely consumption data compared to hourly in all forms of segmented data tested. The performance of the GMM fitting was also impacted, as mixture models were no longer able to obtain a good fit for the full datasets of each unit when the resolution of the data increased. This did not impact the ability of GMM distributions to fit peak load data subsets.

## 5. Conclusion

The research objective of this study was to investigate the best performing parametric functions or models that could accurately characterize different types of residential electricity consumption peak loads at different time resolutions. This was done by using electrical consumption data collected from homes in Canada that were dissected into segments that consisted only of peak load behaviour, where data was available for total, HVAC, and plug load electrical consumptions. For PDF fitting, six typical parametric PDFs were selected based on literature, as well as a mixture model (GMM). It was found that the GEV and Weibull distributions had an adequate performance for the hourly top 1% and daily peak datasets but failed to fit minutely data profiles. This highlights that for higher resolution data fitting, alternative options are needed, which

in the case of this study was fulfilled through mixture modelling. GMM functions were able to consistently fit peak load segmented datasets for all units at a minutely resolution and were also able to characterize both full and segmented datasets at the hourly level. Therefore, for the modelling of accurate PDFs for peak residential energy consumption, GMM was found to be the best model.

The limitations of this study revolved around the shortage of data available. While having minutely data is still relatively uncommon in literature, being limited to one style of six houses provided a limited sample size for analysis. More sizes and designs of households would be useful for further generalizability of GMM's advantages over typical PDFs. Additionally, this study only focused on a limited sample of parametric forms of density functions. Other parametric distributions or nonparametric approaches such as kernel density estimation (KDE) are more complex, but methods such as KDE could have better performance due to their data-driven nature. Despite these limitations, the findings of this study can be applied to cases where a probabilistic estimation is required for either the magnitude or frequency of residential electrical power consumption.

## Acknowledgements

The project was mainly funded by Natural Sciences and Engineering Research Council of Canada (NSERC) Collaborative Research and Development Grants, CRDPJ 530932-18 entitled "Intelligent net-zero energy (ready) modular homes for cold regions". The authors would like to express appreciation to Landmarks Homes who provide research fields, resources, and financial support for this work, without their sponsor, this project is not possible.

## References

Carpaneto, E., and G. Chicco. 2008. 'Probabilistic characterisation of the aggregated residential load patterns', *IET Generation, Transmission and Distribution*, 2: 373-82-82.

Chesser, M., P. Carroll, P. O'Reilly, and P. Lyons. 2021. 'The impact of extreme weather on peak electricity demand from homes heated by air source heat pumps', *Energy Sources, Part B: Economics, Planning and Policy*, 16: 707-18-18.

Das, B., and S. I. Resnick. 2008. 'QQ Plots, Random Sets and Data from a Heavy Tailed Distribution', *Stochastic Models*, 24: 103-32.

Dong, Bing, Zhaoxuan Li, S. M. Mahbobur Rahman, and Rolando Vega. 2016. 'A hybrid model approach for forecasting future residential electricity consumption', *Energy and Buildings*, 117: 341-51.

Gneiting, Tilmann, and Matthias Katzfuss. 2014. 'Probabilistic Forecasting', *Annual Review of Statistics and Its Application*, 1: 125-51.

Hong, Tao, and Shu Fan. 2016. 'Probabilistic electric load forecasting: A tutorial review', *International Journal of Forecasting*, 32: 914-38.

Jang, Youngchan, Eunshin Byon, Elham Jahani, and Kristen Cetin. 2020. 'On the long-term density prediction of peak electricity load with demand side management in buildings', *Energy and Buildings*, 228: 110450.

Knowles, H. S., M. E. Hostetler, and L. S. Liebovitch. 2018. 'Describing the dynamics, distributions, and multiscale relationships in the time evolution of residential building energy consumption', *Energy and Buildings*, 158: 310-25-25.

Massey Jr, Frank J. 1951. 'The Kolmogorov-Smirnov Test for Goodness of Fit', *Journal of the American Statistical Association*, 46: 68-78.

Munkhammar, J., J. Widén, and J. Rydén. 2014. 'Characterizing probability density distributions for household electricity load profiles from high-resolution electricity use data', *Applied Energy*, 135: 382-90-90.

Natural Resources Canada (2018). *Energy Efficiency Trends in Canada.* Accessed November 5, 2023. [Online]. https://oee.nrcan.gc.ca/publications/statistics/trends/2018/residential.cfm.

Richardson, I., and M. Thomson. 2013. 'Integrated simulation of photovoltaic micro-generation and domestic electricity demand: A one-minute resolution open-source model', *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, 227: 73-81-81.

Shiwen, Liao, Wei Lu, Kim Taehyung, and Su Wencong. 2020. 'Modeling and Analysis of Residential Electricity Consumption Statistics: A Tracy-Widom Mixture Density Approximation', *IEEE Access*, 8: 163558-67.

Van Dat, Nguyen, Pham Van Toan, and Ta Minh Thanh. 2022. 'Solving distribution problems in content-based recommendation system with gaussian mixture model', *Applied Intelligence*, 52: 1602-14.

Weron, Rafał. 2006. *Modeling and forecasting electricity loads and prices : a statistical approach* (John Wiley & Sons).

Wright, A., and S. Firth. 2007. 'The nature of domestic electricity-loads and effects of time averaging on statistics and on-site generation calculations', *Applied Energy*, 84: 389-403-03.

Xu, Lei, Shengwei Wang, and Rui Tang. 2019. 'Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load', *Applied Energy*, 237: 180-95.

Xu, Lei, Shengwei Wang, and Fu Xiao. 2019. 'An adaptive optimal monthly peak building demand limiting strategy considering load uncertainty', *Applied Energy*, 253: 113582.