Sustainable Energy, Grids and Networks

Modeling Electricity Consumption Patterns in Buildings Using Probability Distributions

--Manuscript Draft--

Manuscript Number:	SEGAN-D-25-01740
Article Type:	Research Paper
Keywords:	Building electricity consumption; Statistical modeling; Probability distributions; Probability density functions; Goodness-of-Fit; energy efficiency
Corresponding Author:	Mathias de Schietere de Lophem UCLouvain BELGIUM
First Author:	Mathias de Schietere de Lophem
Order of Authors:	Mathias de Schietere de Lophem
	Lambert Misselyn
	Rosana Veroneze
	Hélène Verhaeghe
	Axel Legay
Abstract:	Building electricity consumption highly varies due factors such as occupancy behavior, weather conditions, and operational practices. In this study, we develop a novel statistical framework to model these complex consumption patterns using an extensive set of probability distributions. Specifically, we systematically evaluate twenty candidate models -- including fourteen canonical continuous probability distributions, Gaussian mixture models, 2-parameter Weibull mixture models, and kernel-density estimation using Gaussian kernels -- across three distinct temporal clustering granularities. Our analysis is conducted on three sets of datasets that cover UK households and two Renewable Energy Communities from Brussels, Belgium, encompassing both residential and business energy consumption profiles. For each data cluster, probability distributions are fitted using maximum likelihood estimation or the Expectation-Maximization algorithm. We assess the goodness-of-fit of each model by combining multiple metrics: log-likelihood, Akaike information criterion, likelihood-ratio tests, and a parametric bootstrap-based Kolmogorov–Smirnov test that adjusts for parameter estimation uncertainty. Our findings indicate that clustering granularity plays a crucial role in the effectiveness of probability distribution fitting for electricity consumption data. At coarser granularities, mixture models (MMs) consistently outperformed other methods. However, as granularity became finer, the performance landscape became more varied. While MMs maintained strong performance, other models (such as the log-normal) also demonstrated competitive results in specific scenarios. Our approach enhances electricity consumption modeling accuracy while offering valuable insights for demand management and energy policy. By integrating diverse datasets and a rigorous evaluation framework, we set a new benchmark for analyzing building energy use and supporting more efficient energy systems.

Modeling Electricity Consumption Patterns in Buildings Using Probability Distributions

Mathias de Schietere de Lophem[a,*], Lambert Misselyn[a], Rosana Veroneze[a], Hélène Verhaeghe[a], Axel Legay[b]

[a] *Louvain School of Engineering, Université catholique de Louvain, Louvain-La-Neuve, Belgium*
[b] *Legay Consulting, Belgium*

Abstract

Building electricity consumption highly varies due factors such as occupancy behavior, weather conditions, and operational practices. In this study, we develop a novel statistical framework to model these complex consumption patterns using an extensive set of probability distributions. Specifically, we systematically evaluate twenty candidate models – including fourteen canonical continuous probability distributions, Gaussian mixture models, 2-parameter Weibull mixture models, and kernel-density estimation using Gaussian kernels – across three distinct temporal clustering granularities. Our analysis is conducted on three sets of datasets that cover UK households and two Renewable Energy Communities from Brussels, Belgium, encompassing both residential and business energy consumption profiles. For each data cluster, probability distributions are fitted using maximum likelihood estimation or the Expectation-Maximization algorithm. We assess the goodness-of-fit of each model by combining multiple metrics: log-likelihood, Akaike information criterion, likelihood-ratio tests, and a parametric bootstrap-based Kolmogorov–Smirnov test that adjusts for parameter estimation uncertainty. Our findings indicate that clustering granularity plays a crucial role in the effectiveness of probability distribution fitting for electricity consumption data. At coarser granularities, mixture models (MMs) consistently outperformed other methods. However, as granularity became finer, the performance landscape became more varied. While MMs maintained strong performance, other models (such as the log-normal) also demonstrated competitive results in specific scenarios. Our approach enhances electricity consumption modeling accuracy while offering valuable insights for demand management and energy policy. By integrating diverse datasets and a rigorous evaluation framework, we set a new benchmark for analyzing building energy use and supporting more efficient energy systems.

Keywords:
Building electricity consumption, Statistical modeling, Probability distributions, Probability density functions, Goodness-of-Fit, Energy efficiency

1. Introduction

Energy consumed by the building sector, encompassing energy used for construction, heating, cooling, lighting, and the operation of appliances and equipment, accounts for 30% of global final energy consumption and 26% global energy-related emissions[1]. In Europe, buildings are the largest

*Corresponding author
Email address: mathias.deschietere@uclouvain.be (Mathias de Schietere de Lophem)

[1] https://www.iea.org/energy-system/buildings

energy consumer, making this sector pivotal in achieving EU's ambitious energy and climate goals[2]. Given this critical role, optimizing energy efficiency in buildings is increasingly recognized as a cornerstone for sustainable development. A new trend emerging within this context is the promotion of Renewable Energy Communities (RECs), where groups of users collaboratively produce, share, and consume renewable energy. RECs are a way to decentralize the energy network management, distribute and use energy more efficiently and reduce energy costs. They enhance local energy autonomy and facilitate the integration of renewable energy sources into the grid, contributing to decarbonization efforts. As energy efficiency and sustainability demand intensifies, advancing data-driven methods for understanding and optimizing local building energy consumption becomes imperative. This approach offers potential pathways to reduce emissions and foster greener, more resilient communities.

Research on modeling electricity consumption patterns in buildings is essential for advancing energy efficiency and sustainability. By accurately capturing consumption patterns, models enable better demand-side management, inform energy policy, and guide the design of energy-efficient systems and appliances. These models are particularly critical in RECs, where localized production, sharing, and renewable energy consumption depend on precise forecasts and adaptive strategies to match supply with demand. In RECs, understanding consumption patterns is key to optimizing energy distribution and maximizing the use of renewable sources such as solar and wind, which are inherently variable.

Modeling electricity consumption patterns in buildings using probability distributions provides a statistical framework to analyze and predict energy usage behavior. Each probability distribution is characterized by a probability density function (PDF), which describes the likelihood of a continuous random variable taking on specific values. These functions enable the characterization and quantification of variability, trends and uncertainties in electricity consumption, which are critical for effective energy management. Probabilistic models account for variability and uncertainty in energy use, supporting tasks such as load shedding, planning for peak demands, and simulating various scenarios to assess their impact on energy use. In general, the estimation of PDFs serves as an intermediate step for other tasks, including machine learning tasks, such as clustering, regression, forecasting, and anomaly detection [1, 2, 3]. For instance, PDFs are valuable for detecting anomalies such as equipment malfunctions or unauthorized energy use, by highlighting deviations from expected patterns. In the context of RECs, probabilistic modeling helps to optimize the balance between electricity production and consumption over different time scales, enabling strategies such as maximizing self-consumption and minimizing losses. For example, these models can be used to notify community members when surplus production is expected, encouraging efficient use of energy. By leveraging the descriptive and predictive power of PDFs, we can develop smarter, data-driven strategies to enhance sustainability, resilience, and energy efficiency in the building sector.

This paper extends the current body of literature on modeling electricity consumption patterns in buildings at the individual level, particularly by building upon the influential work of Munkhammar et al. [4, 5]. Their study [4] modeled household electricity load using only two probability distributions (2-parameter Weibull and 2-parameter log-normal) and a single temporal resolution. While impactful, it left several avenues unexplored. Following directions suggested in their future work, such as evaluating a broader range of probability distributions, we significantly

[2] https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en

expand the scope of analysis by assessing 20 probability models, including canonical, mixture, and non-parametric distributions, across three temporal clustering granularities and three diverse sets of datasets (UK households and two Belgian RECs). In addition, their evaluation method was primarily based on the percentage of times the fitted models passed the Kolmogorov–Smirnov (KS) test without accounting for parameter estimation uncertainty. Although they complemented the KS test with visual assessments, their reliance on this single statistical metric led them to conclude that the 2-parameter Weibull distribution was the best-fitting model. As a result, their subsequent work [5] focused exclusively on this distribution. Relying on a more robust evaluation framework, we provide a more comprehensive, data-driven reassessment of established practices in electricity consumption modeling. This systematic extension provides practical insights for model selection and energy management, particularly in the context of RECs.

Our study systematically reassesses established practices in building energy consumption modeling to evaluate how well commonly used probability distributions perform under varying conditions and whether more flexible alternatives should be considered. While prior works have primarily relied on simpler versions of common distributions (e.g., 2-parameter Weibull), we extend this analysis by testing both their more complex variants (e.g., 3-parameter Weibull) and additional distributions, such as exponentially modified normal and inverse Gaussian, to better capture right-skewed consumption patterns. We further explore Gaussian and Weibull mixture models (MMs) to accommodate multimodal behaviors and employ kernel density estimation (KDE) as a non-parametric alternative. Through this comprehensive assessment, we refine best practices for energy consumption modeling, ensuring that model selection is guided by empirical performance rather than convention, ultimately leading to more accurate and reliable statistical representations of electricity consumption patterns.

Our study also stands out for its consideration of diverse datasets. We analyzed three distinct sets of data, encompassing electricity consumption profiles from households in the UK Power Networks' Low Carbon London project [6] and two active RECs in Brussels, Belgium. These RECs include profiles from households and workshops (e.g., catering kitchen, garage, atelier, etc.). Unlike prior studies that focused solely on residential or commercial contexts, our inclusion of both household and business profiles introduces a broader spectrum of consumption dynamics. Households tend, for instance, to display peak usage in the morning and evening, while small businesses exhibit irregular patterns influenced by their specific operations. This diversity makes our modeling approach more representative of varied real-world consumption behavior. Additionally, our analysis of RECs introduces an unexplored perspective, as probability distribution fitting for energy consumption patterns in RECs has, to our knowledge, not been previously studied.

In this study, we focus on electricity consumption rather than instantaneous load, as considered in Munkhammar et al. [4, 5]. While both perspectives offer valuable insights, modeling consumption, defined as energy use aggregated over time, aligns more closely with the objectives of our analysis. Consumption data is increasingly central to modern energy systems, particularly in the context of RECs, where aligning demand with variable renewable generation requires an understanding of temporal usage patterns. This choice enables us to capture behavioral trends, support billing and efficiency analyses, and reflect the type of data most commonly available through smart metering infrastructure. By adopting this perspective, our work complements and extends prior studies, offering a broader and more application-oriented evaluation of probabilistic models for energy use in buildings.

Another contribution is our approach to clustering consumption profiles. Inspired by the temporal resolution-based strategy used by Munkhammar et al. [4, 5], we expanded the scope to include three categories of temporal resolution, providing different clustering granularities. This

approach addresses a key challenge for new RECs, which lack sufficient data to implement finer-grained clustering. Our work facilitates the broader adoption of probabilistic modeling in diverse REC contexts by offering a flexible framework adaptable to varying data availability.

Finally, our evaluation framework establishes a new standard for robustness in assessing model performance. We integrated multiple metrics, including the KS test (with parameter estimation adjustments), log-likelihood, Akaike information criterion (AIC), likelihood-ratio tests, and visual assessments of data histograms and fitted PDFs. Many prior studies relied on the KS test without accounting for its sensitivity to sample size or the need for adjustments when parameters are estimated [4, 5, 7]. By addressing these methodological oversights, we ensure a more rigorous and reliable evaluation of the models. Our findings further highlight the importance of reassessing established practices through comprehensive assessments. For instance, despite the popularity of the 2-parameter Weibull distribution, our results suggest that it may not be a reliable choice for general use, emphasizing the need for a more data-driven approach to model selection.

To sum up, this work's contributions include:

- **Expanded and flexible modeling approaches**: we evaluate both simpler and more complex forms of common parametric distributions (e.g., 2- and 3-parameter Weibull), incorporate additional right-skewed models (e.g., exponentially modified normal and inverse Gaussian), use Gaussian and Weibull MMs to capture multimodal behaviors, and apply KDE as a non-parametric alternative.

- **Refined clustering approach**: we introduce three different temporal clustering granularities, offering a flexible framework that accounts for data availability constraints, making probabilistic modeling more accessible to new RECs.

- **Diverse dataset analysis**: our study includes electricity consumption profiles from UK households and two RECs in Belgium, spanning both residential and business consumption patterns to ensure broader applicability.

- **Rigorous and multi-metric evaluation framework**: We integrate log-likelihood, AIC, likelihood-ratio tests, and a bootstrap-adjusted KS test to assess model fit, addressing methodological gaps in prior studies that used unadjusted or single-metric evaluations.

- **Comprehensive, data-driven evaluation of modeling practices**: We systematically reassess commonly used probability distributions under diverse conditions, revealing the limitations of convention-driven choices (e.g., 2-parameter Weibull), and offer practical guidance for selecting appropriate models – balancing complexity, performance, and clustering granularity.

The remaining sections of this paper are organized as follows. Section 2 introduces some basic concepts. Section 3 presents the related works. Section 4 describes our temporal-based strategies for clustering energy consumption profiles, the models tested in our experiments, and our evaluation framework for goodness-of-fit (GoF). Section 5 presents the experimental setup. Section 6 presents the experimental results. Section 7 discusses our results, and Section 8 concludes our work.

2. Background

Probability distributions provide a statistical framework for analyzing and predicting electricity consumption patterns in buildings. Each probability distribution is characterized by a

probability density function (PDF), which describes the likelihood of a continuous random variable taking on specific values. In our work, the continuous random variable is electricity consumption (energy usage in kilowatt-hours, kWh) at a given time interval. The PDF describes the likelihood of different electricity consumption levels occurring within those intervals, allowing for the characterization of consumption patterns over time.

To estimate the parameters of our chosen distributions, we employ **maximum likelihood estimation** (MLE) [8], a statistical method that identifies the parameter values that maximize the likelihood of observing the given sample data. This approach ensures that our models are the best possible fit to the observed data, based on the likelihood principle [9].

In scenarios where the data exhibits complex patterns that cannot be adequately described by a single probability distribution, **mixture models** (MMs) [9] offer a versatile and powerful solution. An MM represents a combination of multiple probability distributions, each corresponding to a different subpopulation within the overall data set [10]. These subpopulations, or components, may each follow their own unique distribution, allowing the MM to capture the underlying heterogeneity in the data [9].

The **Expectation-Maximization** (EM) algorithm [9, 10] is a key technique we utilize for parameter estimation in MMs. By iteratively alternating between the expectation (E) step, which estimates the distribution of latent variables, and the maximization (M) step, which updates the parameters, the EM algorithm efficiently converges to a local maximum of the likelihood function.

Assessing the **goodness-of-fit** (GoF) [11, 12] of our models is crucial to validate their accuracy and reliability. GoF tests provide a quantitative measure of how well the model's predictions align with the observed data [11, 12]. This alignment is essential for identifying any discrepancies and refining our models accordingly. In our methodology, we employed various GoF metrics to ensure the robustness of our models.

3. Related Work

In this section, we focus on related work that addresses probability distribution fitting and its evaluation in the context of modeling building electricity use. These studies are particularly relevant as they explore the suitability of various statistical distributions to represent electricity consumption patterns and their validation methodologies. For an in-depth review of the current state and future challenges of profile models for building electricity use, we recommend Kang et al.'s work [13]. Their paper differentiates between electricity use profile models (to which our contribution belongs) and electricity use prediction models. It then examines research papers, focusing on four aspects: (*i*) the temporal and spatial characteristics of the used datasets, (*ii*) the distinction between statistical and bottom-up modeling approaches, (*iii*) the validation metrics, and (*iv*) the practical applications of these models in building energy system design and energy policy development. Bottom-up modeling approaches refer to models that typically apply physics-based or elemental methods to analyze the composition of building electricity use profiles. Among the most traditional bottom-up models are those built using building modeling programs, such as EnergyPlus [14]. On the other hand, feature analysis, clustering, regression, classification, and probability distribution fitting are named statistical models by the authors. Clearly, our proposal belongs to their category of statistical models.

Our study's most closely related work is that of Munkhammar et al. [4]. They used 2-parameter Weibull and 2-parameter log-normal distributions to model individual household electricity load profiles. Each electricity load profile were clustered and analyzed considering month, weekday, and hour of the day, resulting in $12 \times 7 \times 24 = 2016$ groups (datasets for the distribution fitting)

per profile. The analysis was also extended to multiple households via convolution of individual electricity use profiles. Household electricity load profiles from Sweden [15] were used in the experiments, consisting of 200 detached houses (20 households were measured for a whole year and the rest for approximately one month only). Measurements were made on a 10-minute resolution for each household. The distributions were fitted using the entire dataset rather than, for instance, splitting it into training and test data. In machine learning, dividing the dataset like this is common. However, their goal was simply to fit a distribution to the data, so using the entire dataset was more appropriate. The parameters of the probability distributions were estimated using maximum likelihood estimation (MLE). The fitted distributions were analyzed in terms of relative variation estimates of electricity use and standard deviation, as well as success in the KS test. Visual assessment of some solutions were also provided. Apparently, the authors did not adjust the KS test to account for the fact that the parameters were estimated from the data [16]. Overall, the Weibull distribution provided a better fit than log-normal, with 74% and 67% pass rates in the KS tests, respectively.

In a follow-up paper, Munkhammar et al. [5] extended their earlier paper by presenting a probabilistic distribution model that integrates household electricity consumption, electric vehicle home-charging, and photovoltaic power production. The model employs a convolution approach to combine three distinct probability distribution models: a 2-parameter Weibull distribution for household electricity use, a Bernoulli distribution for electric vehicle home-charging, and a bimodal normal distribution clear-sky index model for photovoltaic power production. This integrated model was analyzed at two system levels: the individual household level and the aggregate level across multiple households.

Michalková et al. [7] modeled the electricity consumption of a manufacturing company that provides hot and cold bending. The data covers the consumption of electric energy (in kW) during working hours throughout one year, consisting of $3,984$ observations. Five probability distributions were fitted to these observations: normal, 2-parameter Weibull, 1-parameter Rayleigh, logistic, and 2-parameter Gamma. The parameters of the probability distributions were estimated using MLE. The evaluation of the models included the Akaike information criterion (AIC), the coefficient of determination (R^2), the root mean square error (RMSE), and the KS test (both R^2 and RMSE were computed based on the empirical and estimated cumulative distributions). Apparently, the authors did not adjust the KS test either. The authors also included a plot showing the PDFs of the five fitted probability distributions alongside a histogram of their data. The Weibull distribution provided the best fit to their data.

Xu et al.'s work [17] proposed a method to simulate individual-level electricity use curves in residential buildings. Using monthly electricity consumption data from residential sectors of the Jiangsu Province in China, the authors developed a stochastic model to simulate synthetic electricity use curves. The modeling process consisted of three main steps: feature extraction, two-step cluster analysis, and distribution fitting. Three feature types were extracted to characterize household energy use: monthly averaged electricity consumption (μ), coefficient of variance of monthly electricity consumption (CV), and monthly electricity consumption $(a_1, a_2, ..., a_{12})$. The first-step clustering grouped all samples into m clusters based on μ and CV. The second-step clustering further divided each cluster from the first step into n sub-clusters based on the monthly electricity consumption. Thus, after the two-step clustering, data samples were split into $m \times n$ sub-clusters. For each sub-cluster, 14 distributions were fitted: 12 normal distributions to model the monthly electricity consumption (one for each month); and 2 log-normal distributions to model the mean (μ) and CV. The quality of the distribution fitting was evaluated using the Student's t-test. Visual assessment of some distribution fittings was also provided. Their two-step

clustering-based grouping strategy contrasts with the time-based grouping approach adopted by Munkhammar's [4, 5] and our proposal, highlighting different methodologies for structuring data in building electricity use modeling.

A study by E. Carpaneto et al. [18] modeled aggregated residential load power using the following distributions: 1-parameter exponential, 1-parameter Rayleigh, 2-parameter gamma, 2-parameter Gumbel, 2-parameter Weibull, normal, and 3-parameter beta (with the third parameter set to the maximum value of the data). Their experiments used synthetic data generated via Monte Carlo simulation based on a previous study [19]. The simulation inputs included number of customers, hour of the day, season, and working/weekend day (e.g., 100 customers, 10:00, winter, working day). χ^2, KS, and geometrical adaptation statistical tests were used to investigate the GoF of the distributions. The χ^2 test is intended for comparing theoretical and observed *categorical* distributions. The authors thus used discretization to assess the (continuous) distributions. The χ^2 test outcomes are sensitive to the chosen binning (discretization) strategy and can become invalid if the observed or expected frequencies in any category are too small. It is unclear how the authors addressed these challenges. However, the authors did account for the estimation of distribution parameters when performing the KS tests. Additionally, visual assessments of some results were provided. Their findings indicated that the gamma distribution achieved the best results.

Gonzalez-Longatt et al [20] explored the application of the mean-variance mapping optimization (MVMO) algorithm [21] to estimate the parameters of GMMs representing the variability of power system loads. Real-world load data measurements from two substations in the Venezuelan power system (Punto Fijo and Judibana) were used to investigate the suitability of the proposed MVMO approach in modeling complex load patterns. The study found that the GMMs obtained using MVMO and the standard Expectation-Maximization (EM) algorithm produced similar results. However, based on the chi-squared test, the authors identified a slight advantage for the model trained with MVMO. Additionally, simpler models, including normal, gamma, Weibull, exponential, and Rayleigh distributions, were also tested but failed to pass the KS test. Visual inspection of the resulting CDFs further confirmed that these simpler models were not suitable for accurately capturing the complexity of the load data.

Jordan et al [22] conducted a comprehensive study on parametric PDF characterization of single household peak electrical loads. They tested six typical PDFs (Weibull, gamma, normal, log-normal, generalized extreme value (GEV), and beta) and one Gaussian mixture model (GMM). The study utilized high-resolution electrical consumption data from an energy monitoring project in Edmonton, Canada, with data sub-metered into different channels including total, HVAC, and plug load consumptions. The researchers evaluated the GoF using visual inspection through quantile-quantile plots and numerical evaluation via the 2-sample KS test. The results showed that GMM consistently outperformed the typical PDFs, especially for minutely resolution data. While GEV and Weibull distributions performed adequately for hourly data, particularly for top 1% and daily peak datasets, they failed to fit minutely data profiles. The study highlighted the stark difference in fitting results between hourly and minutely resolution data, emphasizing the importance of high-resolution data in capturing peak load behavior accurately.

Bandória et al [23] conducted a comprehensive statistical analysis of electricity consumption patterns across 128 buildings within the University of Campinas (UNICAMP) smart campus, Brazil. The study employed multiple statistical models and tests to characterize Electricity Use Profiles. Specifically, they utilized five hypothesis tests for normality (Shapiro-Wilk, Anderson-Darling, D'Agostino's K-squared, Monte Carlo and Lilliefors' tests), four tests for stationarity (Kwiatkowski-Phillips-Schmidt-Shin, Augmented Dickey-Fuller, Osborn-Chui-Smith-Birchenhall and Canova-Hansen tests), and two tests for autocorrelation (Durbin-Watson and Ljung-Box tests).

The authors also fitted the data to five parametric probability distributions (beta, normal, skew-normal, log-normal and Weibull) and used KDE. The dataset comprised 28 months of electricity consumption data from January 1, 2022, to April 1, 2024, resampled at 15-minute intervals. Results were evaluated by comparing the performance of different models across various time intervals and building types, considering both working and non-working days. The study found that no single distribution model could accurately represent electricity consumption patterns over time for all buildings, highlighting the need for tailored assessments in smart campus energy management.

Existing literature on modeling electricity consumption patterns in buildings typically exhibits four key limitations:

- **Limited model diversity**: Most studies evaluate fewer than ten probability distribution functions (PDFs) [4, 5, 7, 17, 18, 20, 22, 23], often focusing on standard choices like the 2-parameter Weibull or 2-parameter log-normal.
- **Restricted temporal resolution**: Prior works, such as Munkhammar's [4, 5], generally apply a single clustering granularity, limiting insights into how temporal resolution affects model performance.
- **Narrow dataset scope**: Many studies rely on data from a single region or consumer type (e.g., households or offices) [4, 7, 17, 18, 20, 22, 23], which restricts the generalizability of their findings.
- **Simplistic evaluation methodology**: Prior studies often rely on few GoF metrics [4, 20]. Also, it is common to use statistical tests without accounting for their sensitivity to sample size or the need for adjustments when parameters are estimated[4, 7, 22]. This can lead to misleading conclusions about model performance and hinders a robust comparison across different PDFs.

Our study addresses these limitations through a systematic and comprehensive approach. We evaluate 20 different probability models – including canonical, mixture-based, and non-parametric distributions – across three temporal clustering granularities. We apply this framework to three diverse sets of datasets, including UK households and Belgian RECs, which cover both residential and business profiles. Furthermore, we integrate a robust set of GoF metrics – such as the KS test (adjusted for parameter estimation uncertainty), log-likelihood, AIC, likelihood-ratio tests, and visual assessments – into a unified evaluation framework. While some of these metrics have been used individually in prior studies, our work systematically combines them to provide a more rigorous and holistic analysis. These contributions collectively fill important methodological and practical gaps in the literature on electricity consumption modeling.

4. Methodology

This section outlines our methodological approach for modeling building electricity consumption. We begin by describing our clustering strategy, which groups energy consumption data into distinct temporal-based clusters to capture inherent usage patterns. Next, we detail the suite of probability distributions – including canonical distributions, Gaussian mixture models (GMMs), 2-parameter Weibull mixture models (WMMs), and kernel density estimation (KDE) – applied to each data cluster. Finally, we present our comprehensive evaluation framework that employs multiple GoF metrics to assess model performance. This structured methodology enables us to rigorously analyze consumption behavior while providing practical insights for energy efficiency improvements.

4.1. Clustering granularity

Clustering is applied before fitting probability distributions to group time periods with potentially similar energy consumption patterns, enhancing the accuracy and interpretability of the distribution fitting process. Electricity consumption may vary based on factors such as time of day, weekday, month, and seasonality, leading to distinct consumption behaviors. Without clustering, a single probability distribution would have to capture multiple overlapping patterns, potentially resulting in poor fits. Thus, clustering helps mitigate the impact of extreme variations, allowing for more robust parameter estimation and facilitating the identification of well-suited probability models for different temporal contexts.

Previous studies that inspired our work [4, 5] clustered data considering month, weekday, and hour of the day, resulting in $12 \times 7 \times 24 = 2016$ groups per energy consumption profile. A possible limitation of this strategy is that it can lead to clusters with very few samples, potentially compromising the reliability of the evaluation. To address this limitation, we introduce two coarser clustering granularities. The coarser one considers only weekdays and hours of the day. The intermediate one considers the season instead of the month. Thus, the data of each dataset (i.e., energy consumption profile) is clustered and analyzed considering three different categories of temporal resolution:

1. *DH*: week**d**ay × **h**our of the day. This results in $7 \times 24 = 168$ groups.
2. *SDH*: **s**eason × week**d**ay × **h**our of the day. This results in $4 \times 7 \times 24 = 672$ groups.
3. *MDH*: **m**onth × week**d**ay × **h**our of the day. This results in $12 \times 7 \times 24 = 2016$ groups.

This means that, for instance, for a single energy consumption profile under the DH category, each model (such as a normal distribution) will be fitted 168 times (each time in a different data cluster, such as Monday and hour of the day between 14h and 15h).

The three clustering categories offer different trade-offs between temporal resolution and the number of samples per cluster. The MDH category provides the finest granularity, segmenting the data into $2,016$ clusters per energy consumption profile, which allows for highly detailed analysis but may result in clusters with very few data points, potentially compromising the reliability of distribution fitting. In contrast, the SDH category, with 672 clusters per profile, serves as an intermediate approach, striking a balance between capturing seasonal variations and maintaining a sufficient number of samples per cluster for robust modeling. The DH category, with only 168 clusters per profile, benefits from even larger cluster sizes, reducing the risk of insufficient data in each cluster. However, this coarser granularity increases the variability within clusters, which may make probability distribution fitting more challenging.

4.2. Models

In our experiments, we considered: fourteen *canonical* continuous probability distributions; Gaussian mixture models (GMMs); 2-parameter Weibull mixture models (WMMs); and kernel-density estimation (KDE) using Gaussian kernels. Table 1 presents the PDFs for the models used in our experiments, along with their parameters. Each probability distribution is associated with a concise label for reference in the Experimental results section.

Observe that regarding the PDF of a mixture model (MM), the mixture components are normal distributions in the case of GMMs, whereas for WMMs, the components are 2-parameter Weibull distributions.

Table 1: Models used in the experiments.

Label	**Model**	**Probability density function (PDF)**	**Parameters**
norm	normal	$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mu \in \mathbb{R}$, $\sigma > 0$
lognorm2	2-parameter log-normal	$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}$, $\sigma > 0$
lognorm	3-parameter log-normal	$f(x;\mu,\sigma,\theta) = \frac{1}{(x-\theta)\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln(x-\theta)-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}$, $\sigma > 0$, $x > \theta$
exponnorm	exponentially modified normal	$f(x;\mu,\sigma,\lambda) = \frac{\lambda}{2} \exp\left(\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x)\right) * \operatorname{erfc}\left(\frac{\mu+\lambda\sigma^2-x}{\sqrt{2}\sigma}\right)$, where erfc(.) is the *complementary error function*	$\mu \in \mathbb{R}$, $\sigma > 0$, $\lambda > 0$
invgauss2	2-parameter inverse Gaussian	$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right)$	$\mu > 0$, $\lambda > 0$
invgauss	3-parameter inverse Gaussian	$f(x;\mu,\lambda,\theta) = \sqrt{\frac{\lambda}{2\pi(x-\theta)^3}} \exp\left(-\frac{\lambda(x-\theta-\mu)^2}{2\mu^2(x-\theta)}\right)$	$\mu > 0$, $\lambda > 0$, $x > \theta$
beta	beta	$f(x;\alpha,\beta,a,b) = \frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}\mathrm{B}(\alpha,\beta)}$, where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(.)$ is the gamma function	$\alpha > 0$, $\beta > 0$, $a < x < b$
gamma2	2-parameter gamma	$f(x;k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k\,\Gamma(k)}$	$k > 0$, $\theta > 0$
gamma	3-parameter gamma	$f(x;k,\theta,\mu) = \frac{(x-\mu)^{k-1}e^{-\frac{x-\mu}{\theta}}}{\theta^k\,\Gamma(k)}$	$k > 0$, $\theta > 0$, $x > \mu$
Gumbel	right-skewed Gumbel	$f(x;\mu,\beta) = \frac{1}{\beta} e^{-\left(z+e^{-z}\right)}$, where $z = \frac{x-\mu}{\beta}$	$\mu \in \mathbb{R}$, $\beta > 0$
Rayleigh1	1-parameter Rayleigh	$f(x;\sigma) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$	$\sigma > 0$
Rayleigh	2-parameter Rayleigh	$f(x;\sigma,\theta) = \frac{x-\theta}{\sigma^2} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)$	$\sigma > 0$, $x \geq \theta$
Weibull2	2-parameter Weibull minimum	$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$	$\lambda > 0$, $k > 0$
Weibull	3-parameter Weibull minimum	$f(x;\lambda,k,\theta) = \begin{cases} \frac{k}{\lambda}\left(\frac{x-\theta}{\lambda}\right)^{k-1} \exp\left(-\left(\frac{x-\theta}{\lambda}\right)^k\right) & \text{if } x \geq \theta \\ 0 & \text{if } x < \theta \end{cases}$	$\lambda > 0$, $k > 0$, $\theta \in \mathbb{R}$
MM	mixture model	$f(x) = \sum_{i=1}^{k} \pi_i f_i(x;\theta_i)$, where $f_i(.)$ is the i-th mixture component	$\pi_i \geq 0$, $\sum_{i=1}^{k} \pi_i = 1$, θ_i parameters of $f_i(.)$
KDE	kernel-density estimation	$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$, where $K(.)$ is the kernel function, h is the bandwidth, and $\{x_1, x_2, \cdots, x_n\}$ are the data used for the KDE	

The selection of probability distributions in our study is motivated by the need to comprehensively model the complex and highly variable nature of building energy consumption. Previous works have explored a subset of these distributions – normal, log-normal, beta, gamma, Gumbel, Rayleigh, and Weibull – primarily in their simpler forms (e.g., 2-parameter Weibull). We expand on this by evaluating these distributions across diverse datasets and clustering granularities, using a more rigorous evaluation framework. By systematically reassessing established practices, we aim to determine how well commonly used models perform under varying conditions and whether alternative or more flexible models should be considered. This comprehensive assessment helps refine best practices for energy consumption modeling, ensuring that the selected distributions align with the underlying data characteristics rather than being chosen solely based on convention.

To provide greater flexibility in capturing skewness, heavy tails, and overall distributional shape, we also include more complex variants (e.g., 3-parameter Weibull). Additionally, we incorporate the exponentially modified normal and inverse Gaussian distributions, which are well-suited for modeling right-skewed data – an expected characteristic of energy consumption patterns.

Beyond single distributions, we explore Gaussian and Weibull mixture models, which offer greater adaptability by accommodating multimodal distributions. This is particularly relevant when energy usage exhibits multiple distinct behavioral patterns. Finally, we consider KDE as a non-parametric alternative, allowing data-driven distribution modeling without imposing rigid parametric assumptions.

By systematically evaluating this diverse set of models, our study establishes a more robust statistical foundation for modeling electricity consumption in buildings, helping to identify the most appropriate models for different scenarios.

4.3. Goodness of fit

In this study, we do not apply a training and test data split, cross-validation, or similar methods commonly used in machine learning. This is because our focus is not on predictive modeling but on evaluating the goodness-of-fit (GoF) of probability distributions to the data. Instead, we focus on GoF measures to rigorously assess how well the fitted distributions represent the observed data. This approach is more appropriate for the goal of distribution fitting, where the emphasis lies in capturing the underlying data characteristics rather than generalizing to unseen data.

The GoF of a model describes how well it fits a set of observations. It is typically assessed by quantifying the discrepancy between the observed values and those predicted by the model. We used the following tools to evaluate the GoF of a model: log-likelihood; information criteria, in particular the Akaike information criterion (AIC) [24]; likelihood-ratio test when a model is a special case of another one (for instance, 2- and 3-parameter Weibull distributions); Kolmogorov–Smirnov (KS) test (we performed a parametric bootstrap KS test [16, 25] since we are dealing with the case of fitted parameters); and visual assessment.

Log-likelihood is a natural choice for evaluating models trained using MLE and EM, as it measures how well the model explains the observed data. Given n observations $\{x_1, x_2, \cdots, x_n\}$ and a model with parameters Θ, the log-likelihood $\hat{L}$ is given by:

$$\hat{L} = \sum_{i=1}^{n} \ln(f(x_i; \Theta)), \tag{1}$$

where $f(x_i; \Theta)$ is the PDF of the model evaluated at x_i with parameters Θ. The higher the value of log-likelihood, the better. In addition to the log-likelihood, we considered the AIC, which provides

a measure of model quality by balancing the fit of the model with its complexity. The AIC is calculated as:

$$AIC = 2k - 2\hat{L}, \tag{2}$$

where k represents the number of estimated parameters in the model. The AIC introduces a penalty for models with more parameters, which helps prevent overfitting by favoring models that achieve a good fit with fewer parameters. The preferred model is the one with the smallest AIC value.

Using the log-likelihood and AIC metrics provides a balanced approach to model evaluation, where the log-likelihood assesses the fit, and the AIC introduces a complexity trade-off, encouraging parsimonious models. However, these two metrics do not test whether the data follows the assumed distribution shape across all regions (e.g., tails vs. center). For instance, a high log-likelihood could occur even if the model fails to capture specific features of the data distribution. On the other hand, the KS test examines whether the model distribution is consistent with the data. Using the KS test ensures that the chosen model not only fits well (log-likelihood) and balances complexity (AIC) but also adheres to the overall distributional shape of the data. Combining these measures provides a more robust GoF evaluation framework.

The KS test is a widely used GoF method for evaluating how well a sample matches a specified distribution [16]. Let $\{x_1, x_2, \cdots, x_n\}$ be a random sample of size n from a continuous distribution. The null hypothesis H_0 is that x_i's follow a reference distribution. The KS statistic D_n quantifies a distance between the empirical cumulative distribution function (eCDF) of the sample, denoted here as $F_n(.)$, and the CDF of the reference distribution, denoted here as $F(.)$:

$$D_n = \sqrt{n} \sup_x |F_n(x) - F(x)|. \tag{3}$$

However, when the reference distribution has unknown parameters (common in GoF statistical tests), the standard KS test cannot be directly used. Replacing true parameters with estimated ones in the KS test is known to produce inaccurate results [16]. This issue can be addressed using a parametric bootstrap method, where bootstrap samples of the test statistics are constructed from samples generated from the fitted hypothesized distribution [16, 25]. Accordingly, we performed parametric bootstrap KS tests to evaluate the GoF of our models.

The likelihood-ratio (LLR) test is a hypothesis test that involves comparing the GoF of two competing models, where one of them is a particular case of the other. The test statistic, denoted by D_{LLR}, is given by:

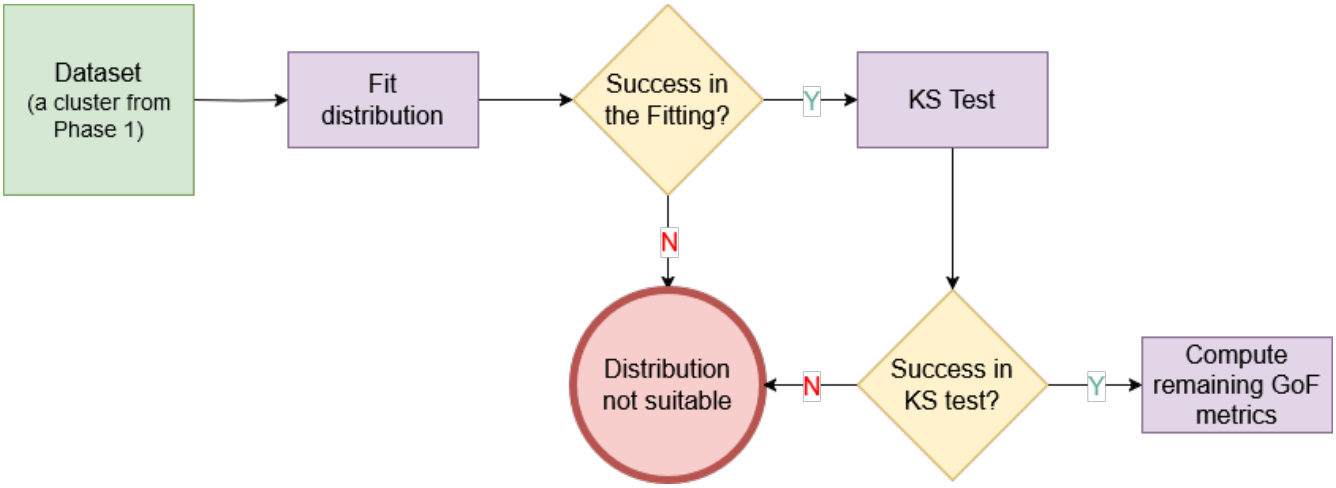
$$D_{LLR} = 2\left(\hat{L}_{alt} - \hat{L}_{null}\right), \tag{4}$$

where $\hat{L}_{alt}$ is the log-likelihood of the alternative model, and $\hat{L}_{null}$ is the log-likelihood of the null model. The model with more parameters (here, alternative) will always fit at least as well (i.e., have the same or greater log-likelihood) than the model with fewer parameters (here, null). To determine whether the more flexible model provides a significantly better fit, we calculate the test p-value, which tells us how likely it is to observe the difference in fit, D_{LLR}, purely by chance if the simpler model (null hypothesis) is true. When the simpler model is a special case of the more complex one, the test statistic approximately follows a χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters between the two models: $df = nparams_{alt} - nparams_{null}$. If the p-value is small, it suggests that the observed difference is unlikely to have occurred by chance, favoring the more complex model.

Evaluation procedure. The evaluation of our methodology follows a step-by-step process. First, we assess whether the model fitting was successful. This requires satisfying the following conditions: (1) the fitting procedure completes without errors, (2) the estimated parameter values are valid, i.e., they do not include *not a number* (NaN) or *infinity* values, and (3) the log-likelihood value is also valid (it is finite and it is not a NaN). Naturally, if the fitting process fails, it would not be significant to calculate any subsequent metrics. Second, we assess whether the p-value of the KS test is greater than the significance level, indicating insufficient evidence to reject the hypothesized model. Thus, the hypothesized model is accepted as a good candidate to represent the data. Given these two conditions (success in the fitting procedure and in the KS test), we analyze the log-likelihood and AIC values of the tested models. Additionally, we analyze the LLR test results for models of the same family.

Fig. 1 presents flowcharts summarizing the experimental process, which is divided into two phases. In the first phase, the input datasets (i.e., energy consumption profiles) are clustered according to DH, SDH, or SDH. In Phase 2, each cluster generated in Phase 1 is an input dataset for the distribution fitting and evaluation process.

(a) Phase 1: clustering of the energy consumption profiles.



(b) Phase 2: distribution fitting and evaluation.

Figure 1: Flowcharts summarizing the experimental process.

5. Experimental setup

5.1. Datasets

Three different sets of datasets were considered in the experiments. The first set, named **UK**, is public and comprises a selection of 10 energy consumption profiles from London households participating in the UK Power Networks' Low Carbon London project, conducted from November 2011 to February 2014 [6]. These 10 profiles were selected because they contain the highest number

of data points and do not include negative values. The other two sets are from active RECs in Brussels, Belgium. They were supplied by the partner company WeSmart[3], which collects them for the day-to-day management of the RECs. For privacy reasons, these datasets are anonymized, and we will refer to them as **REC1** and **REC2** datasets. REC1 comprehends consumption profiles of 10 households collected from April 2023 to March 2024. REC2 includes consumption profiles from 5 workshops (including, for instance, catering kitchen, garage, atelier, etc.) collected from April 2021 to September 2024.

We highlight that our selection of energy consumption profiles was guided by the need for datasets with a sufficient number of samples to enable analysis across different clustering granularities. For example, the energy consumption profiles used in [4] consist of household data collected over relatively short periods, such as one year or even just one month, limiting their suitability for our purposes.

The consumption data were recorded at 15-minute intervals for REC1 and REC2 profiles, and at 30-minute intervals for the UK profiles. To ensure a consistent resolution across all profiles, we adjusted the UK profiles by halving each recorded value. Since our analysis does not rely on detecting fine-grained temporal variations, any impact of smoothing within these intervals is considered negligible.

Each dataset (i.e., each energy consumption profile) within the UK, REC1, and REC2 sets has an identifier. In the UK and REC1 sets, these identifiers are numerical but do not follow a sequential order, as they correspond to specific profiles selected from the respective sets. In the REC2 set, the identifiers are represented by letters (A to E).

Table 2a shows the cluster sizes for each one of the UK datasets under the three clustering granularities considered in this study (DH, SDH, and MDH). Its column names are as follow: "D." stands for *dataset* and contains the identifier of the dataset (energy consumption profile), "Min." stands for *minimum value*, "Max." stands for *maximum value*, "Q1" stands for *first quartile*, and "Q3" stands for *third quartile*. For instance, considering the dataset *0* and the *DH* category, the cluster with the fewest data points has 568, while the cluster with the most data points has 588. Similarly, Table 2b shows the cluster sizes for the REC1 and REC2 datasets. For these two sets, under a same cluster category, the clusters have the same size across their datasets.

Our experiments did not consider the MDH category for REC1 due to its limited sample size. Fitting models to datasets with such a small number of instances produces results that are effectively meaningless.

5.2. Implementation details and parameter settings

We implemented our own version of the Mixture Models (MMs) and used the Expectation-Maximization (EM) algorithm to estimate their parameters. For GMMs, we relied on the scikit-learn library's EM implementation[4]. In the case of WMMs, we employed our custom EM implementation. We evaluated MMs with configurations of 2 and 3 components.

Regarding the fourteen canonical probability distributions, we used the *Scipy* Python library[5], and their parameters were estimated using maximum likelihood estimation (MLE). We also used the *Scipy* Python library for KDE. We employed Scott's and Silverman's rules of thumb for setting the KDE's bandwidth.

[3] `https://www.wesmart.com/`

[4] `https://scikit-learn.org/1.5/modules/generated/sklearn.mixture.GaussianMixture.html`

[5] `https://docs.scipy.org/doc/scipy/reference/stats.html)`

Table 2: Cluster sizes for the datasets.

(a) UK

Cat.	D.	Min.	Max.	Median	Q1	Q3
DH	0	568	588	576	576	576
	1	836	864	847	840	848
	12	928	952	928	928	936
	13	928	952	928	928	928
	14	924	952	928	928	936
	15	925	952	928	928	935
	16	919	952	928	928	936
	17	924	949	928	928	931
	18	920	952	928	928	932
	20	924	952	928	928	928
SDH	0	104	189	142	104	184
	1	204	224	208	208	216
	12	208	296	216	208	244
	13	208	296	216	208	244
	14	208	296	216	208	243
	15	208	296	216	208	244
	16	208	296	216	208	242
	17	205	296	216	208	240
	18	200	296	216	208	244
	20	200	296	216	208	244
MDH	0	32	80	40	32	64
	1	61	84	72	64	72
	12	64	116	72	64	85
	13	64	116	72	64	85
	14	63	116	72	64	85
	15	60	116	72	64	85
	16	61	116	72	64	84
	17	64	116	72	64	85
	18	56	116	72	64	85
	20	56	116	72	64	85

(b) REC1 and REC2

Cat.	Set	Min.	Max.	Median	Q1	Q3
DH	**REC1**	192	192	192	192	192
	REC2	728	732	732	728	732
SDH	**REC1**	36	56	52	45	52
	REC2	152	212	184	159	205
MDH	**REC1**	4	20	16	16	20
	REC2	48	76	62	52	68

We used the Scipy Python library implementation of the KS test[6] when evaluating the canonical continuous probability distributions described in Section 4.2. We used our own implementation of the KS test to evaluate the MMs and the KDE. The number of bootstrap samples drawn from the null hypothesized distribution to form the null distribution of the KS statistic was set to the default value in the Scipy library (i.e., $n\_mc\_samples = 9999$).

For both the LLR and KS tests, we used a significance level equal to 0.05 in our experiments.

Our codes, datasets and figures (for clearer viewing) are available in our GitHub repository at `https://github.com/15154/Modeling-Electricity-Consumption-Patterns-in-Buildings-Using-Probability-Distribution-Functions`.

5.3. Hardware and computing environment

The experiments were conducted on machines of the *Consortium des Équipements de Calcul Intensif (CECI)*. Table 3 describes the machines used for the experiments. More details can be found at `https://www.ceci-hpc.be/clusters.html`

Table 3: CECI machines used for the experiments.

Name	**Specifications**	**Opertating System**
Dragon1	28 computing nodes, 26 computing nodes with two Intel Sandy Bridge (2 x 8-cores E5-2670 processors at 2.6 GHz) and 2 computing nodes with Intel Sandy Bridge (2 x 8-cores E5-2650 processors at 2.00GHz), 128 GB of RAM	NAME="CentOS Linux" VERSION="7 (Core)"
Dragon2	17 computing nodes, each with two Intel Skylake 16-cores Xeon 6142 processors at 2.6 GHz, with 15 nodes having 192GB of RAM and 2 with 384GB	NAME="Rocky Linux" VERSION="8.10 (Green Obsidian)"
Hercules2	1536 cores spread across 30 AMD Epyc Naples and 32 Intel Sandy Bridge compute nodes. The group of AMD nodes are composed of 24 ones with a single 32-core AMD Epyc 7551P CPU at 2.0 GHz and 256 GB of RAM, 4 nodes with the same CPUs and 512 GB of RAM and 2 nodes with dual 32-core AMD Epyc 7501 CPU at 2.0 GHz and 2 TB of RAM. The Intel nodes have dual 8-core Xeon E5-2660 CPU at 2.2 GHz and 64 or 128 GB of RAM (8 nodes)	NAME="CentOS Linux" VERSION="7 (Core)"
Nic5	4672 cores spread across 73 compute nodes with two 32 cores AMD Epyc Rome 7542 CPUs at 2.9 GHz. The default partition holds 70 nodes with 256GB of RAM, and a second "hmem" partition with 3 nodes with 1TB of RAM is also available	NAME="Rocky Linux" VERSION="8.10 (Green Obsidian)"

6. Experimental results

This section is organized into five subsections, each addressing a specific research question (RQ) that guides our analysis and evaluation of the models.

In the first part, we address **RQ1: What percentage of model fittings are successful based on the criteria of (1) error-free completion, (2) validity of estimated parameters, and (3) validity of log-likelihood values?** This step is fundamental because a model that fails to meet these criteria cannot provide reliable metrics or meaningful interpretations. Ensuring the fitting process is successful establishes a foundation for all subsequent analyses.

The second part focuses on **RQ2: What percentage of cases indicate the reference model as a good candidate to represent the data (based on achieving a p-value greater than the significance level in the KS test)?** Evaluating the GoF through the KS test is critical to determine if the hypothesized distributions are plausible representations of the observed data.

[6] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.goodness_of_fit.html`

In the third part, we delve deeper into comparing the reference models using their log-likelihood and AIC values. This analysis provides insights into the relative performance of the models, balancing model complexity with GoF. It addresses **RQ3: How do models compare in terms of log-likelihood and AIC values when fitting the data?**

The fourth part examines models within the same family using the LLR test, addressing **RQ4: Are there statistically significant differences between models of the same family in their ability to represent the data?** This step is essential for refining model selection by identifying the best-performing variants within a family of similar distributions.

Finally, the fifth part presents a visual assessment to illustrate and evaluate the results, answering **RQ5: How well do the models visually capture the patterns and characteristics of the observed data?** Visual assessments provide an intuitive way to complement statistical metrics, helping to identify patterns, anomalies, or areas where models perform well or poorly.

This section addresses these RQs and comprehensively evaluates the tested models, offering a robust framework for understanding their performance and suitability for modeling energy consumption in different types of buildings (houses, apartments, and workshops).

6.1. Successful model fitting

Table 4: Percentage of successful model fitting for the **UK** datasets.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	0	100.0	92.9	100.0	100.0	92.9	100.0	100.0	92.9	100.0	100.0	92.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	12	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	13	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	14	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	15	100.0	98.8	100.0	100.0	98.8	100.0	98.8	98.8	100.0	100.0	98.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	16	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	100.0	100.0	100.0	99.4	100.0	100.0	100.0
	17	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	18	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	98.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	98.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Mean	100.0	98.2	100.0	100.0	98.2	100.0	99.9	98.2	100.0	100.0	98.2	100.0	99.9	100.0	100.0	100.0	99.9	100.0	100.0	100.0
	Std	0.0	1.9	0.0	0.0	1.9	0.0	0.4	1.9	0.0	0.0	1.9	0.0	0.4	0.0	0.0	0.0	0.2	0.0	0.0	0.0
	Rank	1.0	16.8	1.0	1.0	16.8	1.0	2.5	16.8	1.0	1.0	16.8	1.0	2.5	1.0	1.0	1.0	2.5	1.0	1.0	1.0
SDH	0	100.0	98.1	98.2	100.0	98.1	100.0	100.0	98.1	100.0	100.0	98.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1	100.0	99.7	99.9	100.0	99.7	100.0	99.9	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0
	12	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0
	13	100.0	99.7	100.0	100.0	99.7	100.0	99.9	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	14	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0
	15	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	16	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	17	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	18	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Mean	100.0	99.5	99.8	100.0	99.5	100.0	100.0	99.5	100.0	100.0	99.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Std	0.0	0.5	0.6	0.0	0.5	0.0	0.1	0.5	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
	Rank	1.0	17.0	3.8	1.0	17.0	1.0	3.8	17.0	1.0	1.0	17.0	1.0	2.5	1.0	1.0	1.0	2.3	4.0	1.0	1.0
MDH	0	100.0	99.4	99.4	100.0	99.4	100.0	100.0	99.4	100.0	100.0	99.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	1	100.0	99.9	100.0	100.0	99.9	100.0	99.7	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	12	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	13	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	14	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	15	100.0	99.9	100.0	100.0	99.9	100.0	99.6	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	16	100.0	99.9	100.0	100.0	99.9	100.0	99.2	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0
	17	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	18	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Mean	100.0	99.8	99.9	100.0	99.8	100.0	99.8	99.8	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Std	0.0	0.2	0.2	0.0	0.2	0.0	0.3	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Rank	1.0	16.6	14.9	1.0	16.6	1.0	10.8	16.6	1.0	1.0	16.6	1.0	2.3	1.0	1.0	1.0	3.7	5.1	1.0	1.0

Table 5: Percentage of successful model fitting for the **REC1** datasets.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	0	100.0	28.0	28.6	100.0	28.0	100.0	100.0	28.0	100.0	100.0	28.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	10	100.0	28.0	28.6	100.0	28.0	100.0	99.4	28.0	100.0	100.0	28.0	100.0	99.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	12	100.0	28.0	28.6	100.0	28.0	100.0	99.4	28.0	100.0	100.0	28.0	100.0	89.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	13	100.0	28.0	28.6	100.0	28.0	100.0	100.0	28.0	100.0	100.0	28.0	100.0	83.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	18	100.0	28.0	28.6	100.0	28.0	100.0	98.2	28.0	100.0	100.0	28.0	100.0	90.5	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	100.0	28.0	28.6	100.0	28.0	100.0	98.8	28.0	100.0	100.0	28.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	21	100.0	28.0	28.6	100.0	28.0	100.0	100.0	28.0	100.0	100.0	28.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	22	100.0	28.0	28.6	100.0	28.0	100.0	100.0	28.0	100.0	100.0	28.0	100.0	85.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	8	100.0	28.0	28.6	100.0	28.0	100.0	100.0	28.0	100.0	100.0	28.0	100.0	76.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	9	100.0	28.0	28.6	100.0	28.0	100.0	99.4	28.0	100.0	100.0	28.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Mean	100.0	28.0	28.6	100.0	28.0	100.0	99.5	28.0	100.0	100.0	28.0	100.0	92.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Std	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	8.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Rank	1.0	17.0	16.0	1.0	17.0	1.0	7.7	17.0	1.0	1.0	17.0	1.0	9.3	1.0	1.0	1.0	1.0	1.0	1.0	1.0
SDH	0	100.0	78.4	78.4	100.0	78.4	100.0	99.6	78.4	100.0	100.0	78.4	100.0	99.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	10	100.0	78.4	78.4	100.0	78.4	100.0	99.9	78.4	100.0	100.0	78.4	100.0	99.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0
	12	100.0	78.4	78.4	100.0	78.4	100.0	99.4	78.4	100.0	100.0	78.4	100.0	97.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	13	100.0	78.4	78.4	100.0	78.4	100.0	100.0	78.4	100.0	100.0	78.4	100.0	95.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	18	100.0	78.4	78.4	100.0	78.4	100.0	99.3	78.4	100.0	100.0	78.4	100.0	97.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	20	100.0	78.4	78.4	100.0	78.4	100.0	100.0	78.4	100.0	100.0	78.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	21	100.0	78.4	78.4	100.0	78.4	100.0	99.9	78.4	100.0	100.0	78.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	22	100.0	78.4	78.4	100.0	78.4	100.0	100.0	78.4	100.0	100.0	78.4	100.0	94.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	8	100.0	77.4	77.4	100.0	77.4	100.0	100.0	77.4	100.0	100.0	77.4	100.0	95.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	9	100.0	75.6	75.6	100.0	75.6	100.0	99.9	75.6	100.0	100.0	75.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Mean	100.0	78.0	78.0	100.0	78.0	100.0	99.8	78.0	100.0	100.0	78.0	100.0	97.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Std	0.0	0.9	0.9	0.0	0.9	0.0	0.3	0.9	0.0	0.0	0.9	0.0	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Rank	1.0	16.0	16.0	1.0	16.0	1.0	8.9	16.0	1.0	1.0	16.0	1.0	10.8	1.0	1.0	1.0	2.2	1.0	1.0	1.0

Table 6: Percentage of successful model fitting for the **REC2** datasets.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	A	100.0	99.4	100.0	100.0	99.4	100.0	100.0	99.4	100.0	100.0	99.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	B	100.0	99.4	100.0	100.0	99.4	100.0	100.0	99.4	100.0	100.0	99.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	C	99.4	0.0	2.4	99.4	0.0	99.4	99.4	0.0	99.4	99.4	0.0	99.4	3.0	99.4	99.4	99.4	99.4	99.4	99.4	99.4
	D	100.0	0.0	14.3	100.0	0.0	100.0	100.0	0.0	100.0	100.0	0.0	100.0	10.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	E	98.8	0.0	39.9	98.8	0.0	98.8	98.8	0.0	98.8	98.8	0.0	98.8	20.2	98.8	98.8	98.8	98.8	98.8	98.8	98.8
	Mean	99.6	39.8	51.3	99.6	39.8	99.6	99.6	39.8	99.6	99.6	39.8	99.6	46.8	99.6	99.6	99.6	99.6	99.6	99.6	99.6
	Std	0.5	54.4	46.5	0.5	54.4	0.5	0.5	54.4	0.5	0.5	54.4	0.5	49.0	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	Rank	1.0	17.0	9.6	1.0	17.0	1.0	1.0	17.0	1.0	1.0	17.0	1.0	9.8	1.0	1.0	1.0	1.0	1.0	1.0	1.0
SDH	A	100.0	99.7	99.9	100.0	99.7	100.0	99.9	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	B	100.0	99.7	99.9	100.0	99.7	100.0	100.0	99.7	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	C	99.9	2.2	5.7	99.9	2.2	99.9	99.9	2.2	99.9	99.9	2.2	99.9	9.2	99.9	99.9	99.9	99.9	99.9	99.9	99.9
	D	100.0	7.7	14.0	100.0	7.7	100.0	100.0	7.7	100.0	100.0	7.7	100.0	17.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	E	99.7	17.6	47.0	99.7	17.6	99.7	99.7	17.6	99.7	99.7	17.6	99.7	33.9	99.7	99.7	99.7	99.7	99.7	99.7	99.7
	Mean	99.9	45.4	53.3	99.9	45.4	99.9	99.9	45.4	99.9	99.9	45.4	99.9	52.1	99.9	99.9	99.9	99.9	99.9	99.9	99.9
	Std	0.1	49.9	45.2	0.1	49.9	0.1	0.1	49.9	0.1	0.1	49.9	0.1	44.6	0.1	0.1	0.1	0.1	0.1	0.1	0.1
	Rank	1.0	17.0	15.6	1.0	17.0	1.0	3.8	17.0	1.0	1.0	17.0	1.0	9.6	1.0	1.0	1.0	1.0	1.0	1.0	1.0
MDH	A	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	B	100.0	99.9	100.0	100.0	99.9	100.0	99.9	99.9	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	C	100.0	8.7	10.7	100.0	8.7	100.0	100.0	8.7	100.0	100.0	8.7	100.0	19.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	D	98.3	19.0	22.0	100.0	19.0	100.0	100.0	19.0	98.3	98.3	19.0	98.3	24.8	100.0	100.0	100.0	100.0	100.0	98.3	98.3
	E	99.9	44.5	54.0	99.9	44.5	99.9	99.4	44.5	99.9	99.9	44.5	99.9	54.0	99.9	99.9	99.9	99.9	99.9	99.9	99.9
	Mean	99.6	54.4	57.3	100.0	54.4	100.0	99.8	54.4	99.6	99.6	54.4	99.6	59.6	100.0	100.0	100.0	100.0	100.0	99.6	99.6
	Std	0.8	43.5	42.0	0.0	43.5	0.0	0.3	43.5	0.8	0.8	43.5	0.8	39.2	0.0	0.0	0.0	0.0	0.0	0.8	0.8
	Rank	2.6	16.8	15.6	1.0	16.8	1.0	6.6	16.8	2.6	2.6	16.8	2.6	9.4	1.0	1.0	1.0	1.0	1.0	2.6	2.6

Table 4 presents the percentage of successful model fitting for the UK datasets. Overall, the results were very good for this group of datasets. The worst results are for the dataset 0 under the category DH for the distributions lognorm2, invgauss2, gamma2, and Rayleigh1, but their fitting success was around 93%.

Table 5 summarizes the percentage of successful model fitting for the REC1 datasets. The models lognorm2, lognorm, invgauss2, gamma2, and Rayleigh1 demonstrated the worst results, with around 28% success for DH and 78% for SDH. This represented approximately a 2.8× improvement in the SDH category over the DH category. While Weibull2 did not achieve around 100% success across all cases, its performance was expressively better than the aforementioned models. The fitting process for the other models was consistently successful. Results were largely uniform across datasets, with Weibull2 showing the greatest variation: success rates ranged from 76.2% to 100% for DH and from 94.8% to 100% for SDH.

Table 6 presents the percentage of successful model fitting for the REC2 datasets. The results varied significantly across datasets, with A and B being the easiest to model and C, D, and E posing greater challenges. Even among the more difficult datasets, there were noticeable differences, with performance ranking as C $\leq$ D $\leq$ E. For datasets A and B, the fitting process was successful for all models, achieving over 99% success. In contrast, datasets C, D, and E showed poor performance for certain models, which are lognorm2, lognorm, invgauss2, gamma2, Rayleigh1, and Weibull2. Across these more challenging datasets, performance generally improved with increasing granularity in the categories, following the trend DH $<$ SDH $<$ MDH. This highlights the influence of dataset characteristics and modeling granularity on the success of the fitting process.

For clarity, the row labeled "Rank" in these tables represents the average ranking of each model, computed across the datasets. The ranking ranges from 1 to 20, as our experiments included 20 models, with a rank of 1 assigned to the best-performing model. In cases where multiple models achieved the same performance, the best rank among them was assigned to each tied models.

6.2. Kolmogorov-Smirnov test

Tables 7 to 9 present the percentage of model fittings that achieved a p-value greater than the significance level in the KS test. Success in the KS test implicitly requires prior success in the model fitting process. For readers who prefer an alternative perspective, Tables A.13 to A.15 in Appendix A provide equivalent results, but the percentages are computed considering only cases where the model fitting was successful. Notably, Table 7 can be interpreted as the product of Table 4 (percentage of successful model fittings) and Table A.13. Equivalently, for Tables 8 and 9.

Table 7: Percentage of Kolmogorov-Smirnov (KS) tests with p-values exceeding the significance level for the **UK** datasets.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	0	0.0	1.8	2.4	1.8	1.8	2.4	4.2	3.6	4.8	0.0	0.0	0.0	1.8	2.4	13.1	32.7	48.8	57.7	10.1	9.5
	1	0.0	3.6	3.6	6.0	2.4	2.4	1.8	1.8	2.4	1.2	0.0	0.0	0.0	0.6	11.3	40.5	53.6	61.3	22.6	18.5
	12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.6	16.1	61.3	67.9	0.0	0.0
	13	0.0	1.2	1.2	0.6	1.2	1.2	1.2	1.2	1.2	0.0	0.6	0.6	1.2	1.2	8.3	37.5	66.7	77.4	32.7	28.0
	14	1.8	6.5	6.5	8.3	7.7	8.3	1.8	1.8	1.8	0.0	0.0	0.0	0.6	1.2	8.3	33.3	58.3	76.2	13.7	12.5
	15	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.0	26.8	56.5	64.3	13.1	12.5
	16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	34.5	4.2	11.9	0.0	0.0
	17	0.0	0.0	0.0	1.2	0.0	0.0	1.2	0.6	1.2	0.0	0.0	0.0	1.2	1.2	0.0	9.5	38.1	51.8	0.6	0.6
	18	0.0	0.6	0.6	0.0	2.4	3.6	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	13.1	53.0	60.7	0.0	0.0
	20	0.0	3.0	3.0	4.8	3.0	3.0	2.4	3.0	4.2	2.4	0.0	0.0	0.6	0.6	7.7	20.8	58.9	69.6	8.3	8.3
	Mean	0.2	1.7	1.7	2.3	1.8	2.1	1.4	1.2	1.5	0.4	0.1	0.1	0.5	0.7	7.4	26.5	49.9	59.9	10.1	9.0
	Std	0.6	2.2	2.2	3.0	2.4	2.6	1.2	1.3	1.8	0.8	0.2	0.2	0.7	0.8	7.0	10.9	17.8	18.6	11.0	9.4
	Rank	12.1	8.2	7.9	7.9	8.2	7.5	8.0	8.5	7.4	12.1	12.4	12.4	10.0	9.5	6.1	2.8	2.1	1.1	5.3	5.8
SDH	0	7.0	19.3	20.1	25.4	18.8	19.0	25.7	17.0	25.3	8.8	4.3	4.5	14.4	20.5	48.7	70.7	49.0	58.9	23.7	22.0
	1	1.9	14.9	15.2	17.4	13.5	14.4	14.1	10.6	12.6	9.2	5.2	5.2	9.1	11.5	32.4	60.6	55.4	65.3	27.2	25.1
	12	0.1	5.8	6.8	4.5	4.5	6.0	3.7	1.6	4.2	1.0	0.7	0.7	1.8	3.1	19.2	44.3	36.2	57.0	2.5	2.2
	13	5.2	11.5	11.8	15.0	10.0	10.0	18.5	8.3	10.6	5.1	3.7	3.9	10.3	11.5	45.1	67.4	60.1	76.3	43.6	42.6
	14	7.1	28.7	30.1	26.9	26.9	28.4	22.9	19.2	23.1	10.1	4.3	4.5	14.4	19.8	40.2	67.6	52.5	75.7	28.7	27.1
	15	4.5	13.1	13.8	18.2	11.6	11.8	12.6	8.9	11.6	7.1	3.0	3.0	7.1	8.2	29.2	59.2	44.3	58.2	25.3	23.8
	16	0.1	2.1	2.1	2.7	2.2	2.4	4.6	0.9	2.2	0.9	0.4	0.4	1.6	2.4	10.9	39.6	23.1	35.7	5.4	4.9
	17	2.2	6.8	7.4	9.5	7.0	7.3	14.4	7.4	11.2	1.0	3.0	3.0	6.5	9.7	12.8	43.5	39.4	55.1	5.7	5.2
	18	0.0	18.9	19.5	9.1	20.1	22.0	17.3	11.0	15.8	0.3	0.0	0.0	8.5	13.4	9.1	33.2	50.6	59.2	0.9	0.7
	20	5.8	21.7	23.2	21.1	19.3	20.1	19.0	15.6	16.8	12.6	4.8	4.8	12.2	14.3	38.2	58.6	49.3	67.4	24.1	23.4
	Mean	3.4	14.3	15.0	15.0	13.4	14.1	15.3	10.1	13.3	5.6	2.9	3.0	8.6	11.4	28.6	54.5	46.0	60.9	18.7	17.7
	Std	2.9	8.1	8.5	8.4	7.8	8.2	7.1	6.0	7.3	4.6	1.9	1.9	4.5	6.0	14.7	13.2	10.7	11.6	14.2	13.7
	Rank	18.6	9.4	8.1	8.3	10.9	9.6	8.8	14.1	10.5	17.0	18.7	18.4	15.3	11.6	4.9	1.8	2.9	1.3	8.2	9.4
MDH	0	28.8	47.7	51.3	53.6	47.1	49.7	44.8	47.9	55.6	34.4	27.6	27.9	42.9	59.0	72.3	84.0	48.9	60.4	41.9	40.7
	1	16.0	37.7	38.9	40.7	36.6	37.3	45.0	36.6	47.0	26.3	20.1	20.0	33.0	42.3	60.7	80.7	58.3	67.4	39.1	36.9
	12	3.3	18.2	21.1	19.8	17.4	21.1	21.1	12.9	31.7	6.8	5.0	5.0	10.4	27.7	44.0	64.3	26.2	44.9	10.1	9.4
	13	20.9	32.2	34.2	39.1	30.0	31.3	48.0	31.4	42.6	20.0	20.1	20.1	37.3	45.3	68.7	84.2	54.0	75.8	52.7	50.4
	14	19.3	51.1	53.4	50.4	48.7	51.7	48.8	42.6	52.9	29.4	20.3	20.3	38.9	52.8	70.3	84.9	49.9	72.2	41.6	39.8
	15	15.8	36.5	38.8	43.1	36.0	37.1	36.3	31.2	43.8	23.9	17.9	17.7	29.7	41.9	60.3	80.8	41.6	55.4	39.4	37.1
	16	6.3	14.5	15.9	19.8	13.7	15.4	22.0	12.3	24.8	8.3	6.8	6.9	11.9	22.2	35.0	65.5	24.6	44.7	16.3	15.5
	17	12.6	29.0	31.1	37.0	27.5	29.3	39.4	24.6	44.0	18.0	15.5	15.5	25.7	39.1	50.5	71.4	41.5	62.5	23.7	22.4
	18	3.2	42.0	43.8	29.4	40.2	41.9	42.5	48.7	47.1	7.9	4.4	4.4	36.4	47.7	49.2	66.7	50.8	61.2	7.8	7.5
	20	21.4	44.7	46.8	46.7	44.3	46.0	43.3	37.1	48.1	31.0	24.1	24.2	35.1	47.4	68.4	83.1	45.7	66.7	40.4	39.2
	Mean	14.8	35.4	37.5	37.9	34.1	36.1	39.1	32.5	43.8	20.6	16.2	16.2	30.1	42.5	57.9	76.6	44.1	61.1	31.3	29.9
	Std	8.4	12.1	12.3	11.8	12.0	12.0	10.0	12.9	9.3	10.2	8.2	8.2	11.1	10.9	12.7	8.5	11.1	10.4	15.5	14.9
	Rank	19.5	11.0	8.2	8.5	12.9	10.0	9.3	12.8	5.1	17.1	18.6	18.3	14.5	6.0	2.8	1.0	6.2	2.3	11.9	13.5

Table 8: Percentage of Kolmogorov-Smirnov (KS) tests with p-values exceeding the significance level for the **REC1** datasets.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	**0**	0.0	0.6	1.2	0.0	1.2	1.2	1.2	0.6	1.2	0.6	0.0	0.0	0.0	0.6	22.6	49.4	8.3	19.0	3.0	2.4
	10	3.0	3.6	3.6	7.1	1.8	4.2	5.4	1.2	3.6	3.0	0.0	0.6	7.1	8.9	30.4	55.4	19.0	25.0	26.2	25.0
	12	0.0	5.4	5.4	2.4	3.0	4.8	4.8	1.8	3.0	1.8	0.6	1.2	3.6	4.2	25.0	42.3	19.0	21.4	2.4	2.4
	13	0.0	0.6	1.8	0.0	1.2	1.8	0.6	0.0	0.6	0.0	0.0	0.0	0.0	0.6	5.4	23.2	16.1	16.7	0.0	0.0
	18	0.0	0.6	0.6	0.6	0.6	0.6	1.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	38.1	20.2	22.0	0.6	0.6
	20	0.6	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.6	0.0	1.2	1.2	0.0	0.0	28.6	47.6	15.5	25.0	13.7	12.5
	21	8.3	4.8	4.8	22.6	5.4	16.1	22.6	6.0	20.8	14.9	5.4	10.7	13.1	21.4	47.6	50.0	19.6	21.4	29.2	27.4
	22	0.0	1.8	1.8	7.1	2.4	6.0	2.4	0.6	1.8	2.4	0.0	0.6	3.0	1.8	18.5	54.2	16.1	22.6	4.2	3.0
	8	0.0	4.2	4.2	1.8	4.2	4.8	0.0	0.0	0.0	0.6	0.0	0.0	0.6	3.0	14.3	36.9	22.6	22.6	0.6	0.6
	9	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.3	31.5	1.8	1.2	7.1	6.0
Mean		1.2	2.1	2.3	4.2	2.0	3.9	4.2	1.0	3.2	2.3	0.7	1.4	2.7	4.0	23.2	42.9	15.8	19.7	8.7	8.0
Std		2.7	2.1	2.0	7.1	1.8	4.8	6.7	1.8	6.3	4.5	1.7	3.3	4.3	6.7	11.9	10.4	6.3	7.0	10.8	10.3
Rank		14.3	10.1	9.3	9.5	10.1	7.8	8.1	13.9	10.6	12.1	14.7	14.0	11.1	9.8	2.5	1.0	4.6	3.5	6.8	7.4
SDH	**0**	8.5	16.5	19.0	20.7	18.0	22.2	20.2	12.9	22.9	10.1	7.1	7.3	11.3	25.1	40.5	54.8	26.3	43.9	10.9	10.0
	10	13.7	35.4	38.7	39.4	35.9	41.5	38.5	34.5	40.6	18.3	12.6	14.1	32.6	49.1	54.5	74.6	42.0	54.3	30.2	28.1
	12	10.1	33.8	36.9	32.6	31.5	39.0	39.7	40.6	39.1	23.1	14.6	16.8	34.1	45.1	56.8	75.0	41.7	56.0	22.9	21.9
	13	4.5	12.8	14.1	19.5	14.9	17.9	19.8	22.2	20.8	4.8	3.7	3.6	18.8	24.6	38.1	54.6	38.5	49.7	10.7	10.1
	18	8.9	13.1	15.6	15.9	14.3	18.0	20.1	16.8	15.2	6.8	4.2	4.3	14.0	19.3	38.8	49.3	41.8	58.2	12.6	12.4
	20	18.9	29.2	29.2	30.8	20.2	27.7	28.1	24.1	33.0	18.8	17.0	17.3	30.1	35.6	64.7	83.8	24.1	28.9	34.8	34.5
	21	29.6	35.6	36.9	50.7	33.9	41.5	40.8	31.4	49.4	34.1	26.0	30.2	33.0	52.4	62.4	75.3	45.2	51.3	45.4	44.0
	22	8.8	26.9	28.6	25.4	25.6	32.4	26.5	29.9	30.4	15.5	7.6	9.8	24.6	29.2	55.7	61.9	41.7	54.0	18.3	16.5
	8	8.3	28.1	31.8	28.4	31.5	41.2	27.5	34.2	29.2	14.3	7.6	11.3	25.1	35.9	53.7	68.3	43.2	58.2	17.3	16.2
	9	16.4	28.9	29.0	35.6	21.1	28.4	35.6	23.2	37.6	20.5	18.8	19.2	22.9	36.6	50.7	79.5	13.4	13.4	36.5	35.0
Mean		12.8	26.0	28.0	29.9	24.7	31.0	29.7	27.0	31.8	16.6	11.9	13.4	24.6	35.3	51.6	67.7	35.8	46.8	24.0	22.9
Std		7.3	8.8	8.9	10.4	8.1	9.6	8.4	8.6	10.3	8.5	7.2	8.0	8.0	11.0	9.5	11.9	10.7	14.6	12.1	11.9
Rank		18.4	11.8	9.9	8.9	12.6	8.4	9.2	10.3	7.0	16.4	19.5	18.3	12.7	4.8	2.6	1.1	6.9	5.0	12.2	13.6

Table 9: Percentage of Kolmogorov-Smirnov (KS) tests with p-values exceeding the significance level for the **REC2** datasets.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	5.4	48.2	82.7	0.0	0.0
	B	0.0	3.6	3.6	3.6	3.0	3.0	1.2	2.4	3.0	6.0	0.0	0.0	0.0	0.0	17.9	23.8	55.4	56.0	20.8	19.0
	C	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	19.0	50.0	0.0	0.0	0.0	0.0
	D	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.2	0.0	0.0	0.0	0.0
	E	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.2	4.8	1.8	0.6	0.0	0.0
Mean		0.0	0.7	0.7	0.7	0.6	0.6	0.2	0.5	0.7	1.2	0.0	0.0	0.0	0.0	7.6	17.6	21.1	27.9	4.2	3.8
Std		0.0	1.6	1.6	1.6	1.3	1.3	0.5	1.1	1.3	2.7	0.0	0.0	0.0	0.0	9.9	19.9	28.2	39.0	9.3	8.5
Rank		6.2	4.6	4.6	4.6	5.2	5.2	6	5.8	5	4.4	6.2	6.2	6.2	6.2	3.6	1.8	2.2	2.2	3.8	4
SDH	A	1.9	2.1	2.2	2.5	0.9	1.0	2.2	2.1	2.1	0.0	0.1	0.1	1.8	1.2	13.5	23.8	43.6	42.7	5.2	4.9
	B	14.1	19.0	19.8	24.1	17.4	17.4	24.1	16.7	19.8	11.0	9.4	9.5	18.2	19.3	42.1	60.9	46.3	53.9	36.6	35.7
	C	0.0	0.1	0.1	1.3	0.3	0.9	0.9	0.3	0.9	0.3	0.0	0.1	0.9	0.7	11.8	25.7	1.3	0.9	0.1	0.1
	D	0.0	2.2	2.8	8.9	2.7	5.8	2.8	2.2	2.2	3.1	0.3	1.6	4.6	3.3	14.7	19.2	4.2	4.3	0.6	0.6
	E	0.0	0.0	0.0	1.0	0.1	0.3	1.8	1.0	0.9	0.0	0.1	0.1	0.9	0.9	16.1	22.6	8.6	9.5	0.3	0.3
Mean		3.2	4.7	5.0	7.6	4.3	5.1	6.4	4.5	5.2	2.9	2.0	2.3	5.3	5.1	19.6	30.4	20.8	22.3	8.6	8.3
Std		6.2	8.1	8.4	9.8	7.4	7.2	9.9	6.9	8.2	4.7	4.1	4.1	7.4	8.0	12.7	17.2	22.2	24.3	15.8	15.4
Rank		17.2	13.2	11.6	5.2	13.6	10.0	7.0	11.2	9.0	15.0	18.0	16.2	9.0	10.4	2.8	1.4	3.6	3.6	10.4	10.8
MDH	A	10.6	12.3	13.2	15.8	12.3	12.8	20.9	14.7	17.3	7.5	7.4	7.4	14.0	17.3	28.9	37.6	39.1	42.1	15.2	15.0
	B	33.5	41.6	43.3	48.1	40.2	40.9	48.7	40.5	49.8	31.5	29.2	29.3	40.3	47.6	66.9	83.2	50.4	52.3	53.2	52.1
	C	0.9	0.8	0.7	5.2	1.4	3.9	2.9	3.3	3.6	2.8	0.5	2.1	3.8	2.6	16.4	26.8	5.3	5.3	1.9	1.8
	D	3.7	9.5	9.2	15.5	11.3	13.3	10.6	13.2	10.0	9.2	4.5	6.2	14.5	11.7	30.2	36.8	14.5	14.8	5.1	4.8
	E	5.0	5.9	5.1	10.5	7.7	6.6	12.6	14.2	10.0	5.3	4.5	5.7	9.1	8.5	28.8	43.8	28.6	32.0	6.7	6.5
Mean		10.7	14.0	14.3	19.0	14.6	15.5	19.1	17.2	18.1	11.3	9.2	10.1	16.3	17.5	34.2	45.7	27.6	29.3	16.4	16.1
Std		13.2	16.0	16.8	16.8	15.0	14.8	17.7	13.9	18.4	11.6	11.4	10.9	14.1	17.6	19.1	21.9	18.2	19.3	21.2	20.7
Rank		18.0	14.6	15.0	6.4	13.6	10.6	8.0	9.4	8.2	15.6	19.6	16.8	9.6	9.4	2.6	1.4	4.0	2.8	11.0	12.4

For the UK datasets, the percentage of successful model fittings achieving a p-value greater than the significance level in the KS test followed the trend MDH > SDH > DH. For DH, the best-performing models were (in terms of mean percentage and standard deviation): WMM3 (59.9 ± 18.6), WMM2 (49.9 ± 17.8), and GMM3 (26.5 ± 10.9), KDE_scott (10.1 ± 11.0), KDE_silverman (9.0 ± 9.4), and GMM2 (7.4 ± 7.0). For SDH, the best models were: WMM3 (60.9 ± 11.6), GMM3 (54.5 ± 13.2), WMM2 (46.0 ± 10.7), GMM2 (28.6 ± 14.7), and KDE_scott (18.7 ± 14.2). For MDH, the best models were: GMM3 (76.6 ± 8.5), WMM3 (61.1 ± 10.4), GMM2 (57.9 ± 12.7), WMM2 (44.1 ± 11.1), gamma (43.8 ± 9.3), and Weibull (42.5 ± 10.9). Performance varied among datasets and models, reflecting the influence of dataset-specific characteristics on modeling success.

For the REC1 datasets, the percentage of successful model fittings achieving a p-value greater than the significance level in the KS test was consistently higher for SDH compared to DH. The best-performing models for DH were GMM3 (42.9 ± 10.4), GMM2 (23.2 ± 11.9), WMM3 (19.7 ± 7.0), WMM2 (15.8 ± 6.3), and KDEscott (8.7 ± 10.8). For SDH, the best models were GMM3 (67.7 ± 11.9), GMM2 (51.6 ± 9.5), WMM3 (46.8 ± 14.6), Weibull (35.3 ± 11.0), and WMM2 (35.8 ± 10.7). The improvement of SDH over DH varied significantly across models and datasets, ranging from $1.1\times$ to $59.8\times$ better. Additionally, performance varied substantially among datasets; for example, dataset 21 exhibited much better results than dataset 13, highlighting variability in model effectiveness across different data scenarios.

For the REC2 datasets, MMs once again emerged as top performers. For DH, the best models were WMM3 (27.9 ± 39.0), WMM2 (21.1 ± 28.2), GMM3 (17.6 ± 19.9), and GMM2 (7.6 ± 9.9). For SDH, the leading models were GMM3 (30.4 ± 17.2), WMM3 (22.3 ± 24.3), WMM2 (20.8 ± 22.2), and GMM2 (19.6 ± 12.7). For MDH, the best results were achieved by GMM3 (45.7 ± 21.9), GMM2 (34.2 ± 19.1), WMM3 (29.3 ± 19.3), and WMM2 (27.6 ± 18.2). While the trend MDH > SDH > DH was observed overall, it did not hold consistently across all cases (e.g., for datasets A and B, we had MDH < SDH < DH for WMM3). In terms of datasets, the results for A and B outperformed C, D, and E, with B having better results than A.

6.3. Log-Likelihood and AIC

Figs. 2 to 4 and 5 to 7 display the log-likelihood and AIC values, respectively, for the UK, REC1, and REC2 datasets. Note that a box is absent if the percentage of success in the KS test equals 0%. We fixed the y-limits to prevent scaling issues caused by some distributions yielding very poor results.

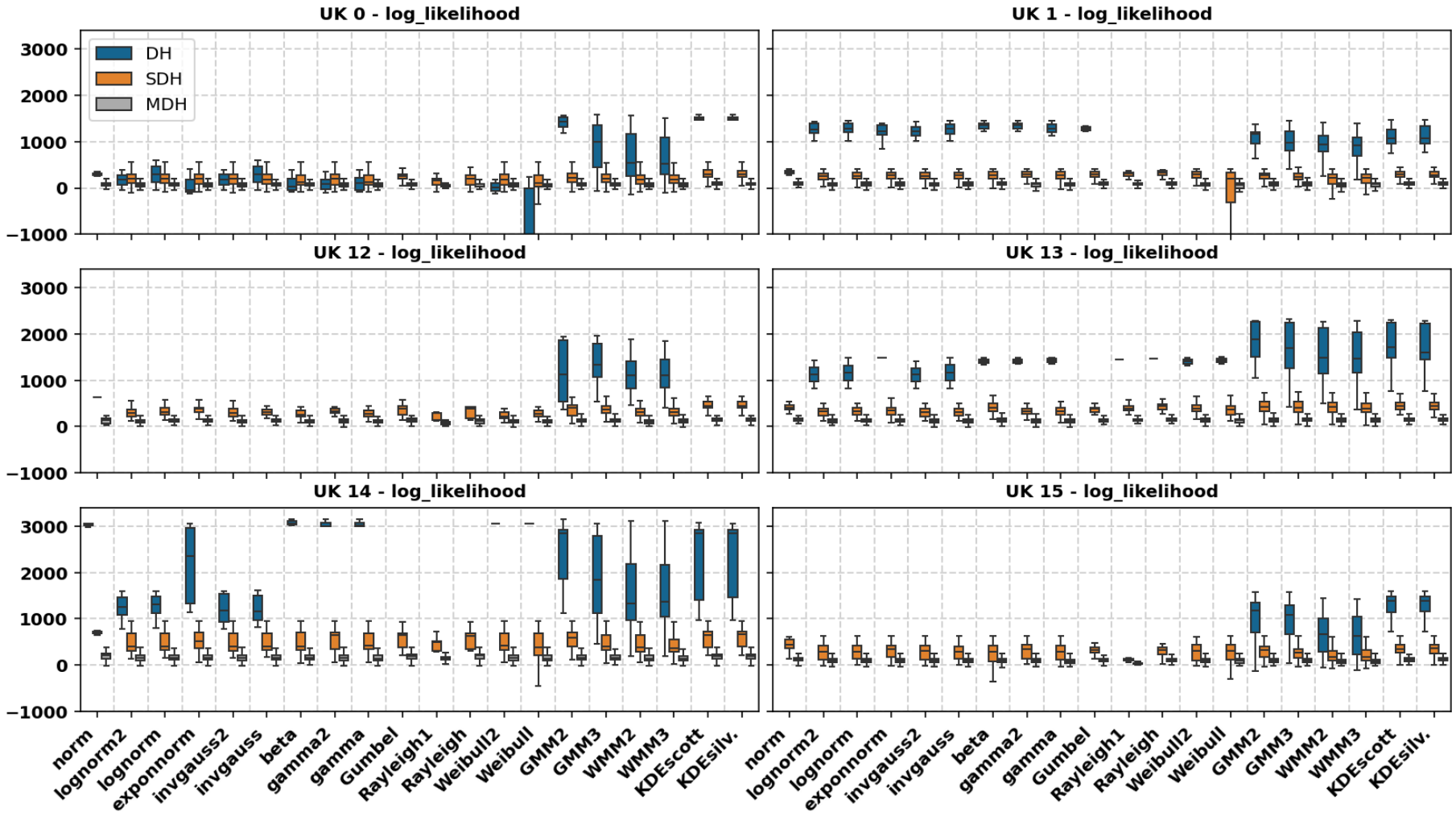


Figure 2: Log-likelihood for the **UK** datasets (part 1 of 2).

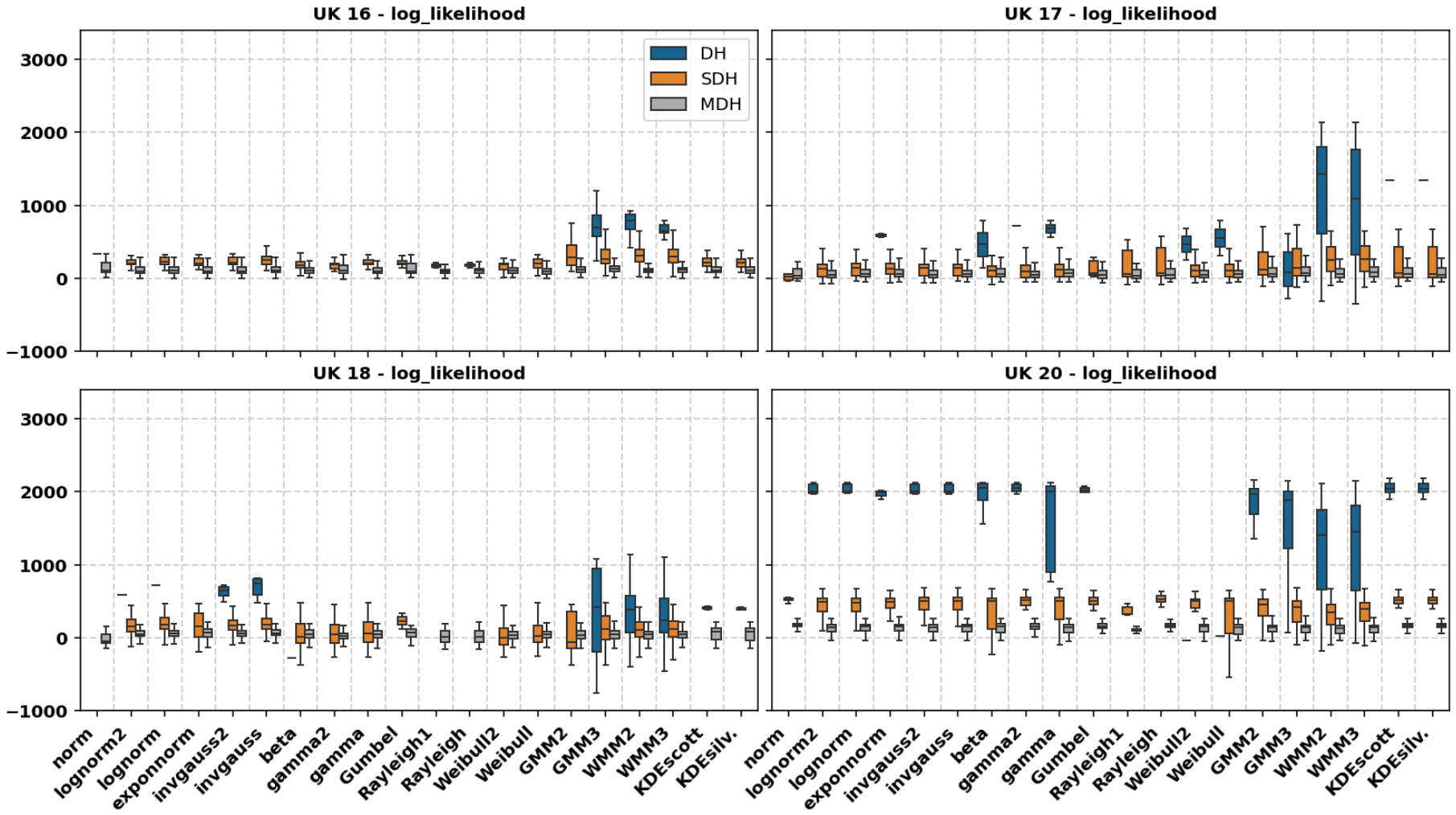


Figure 2: Log-likelihood for the **UK** datasets (part 2 of 2).

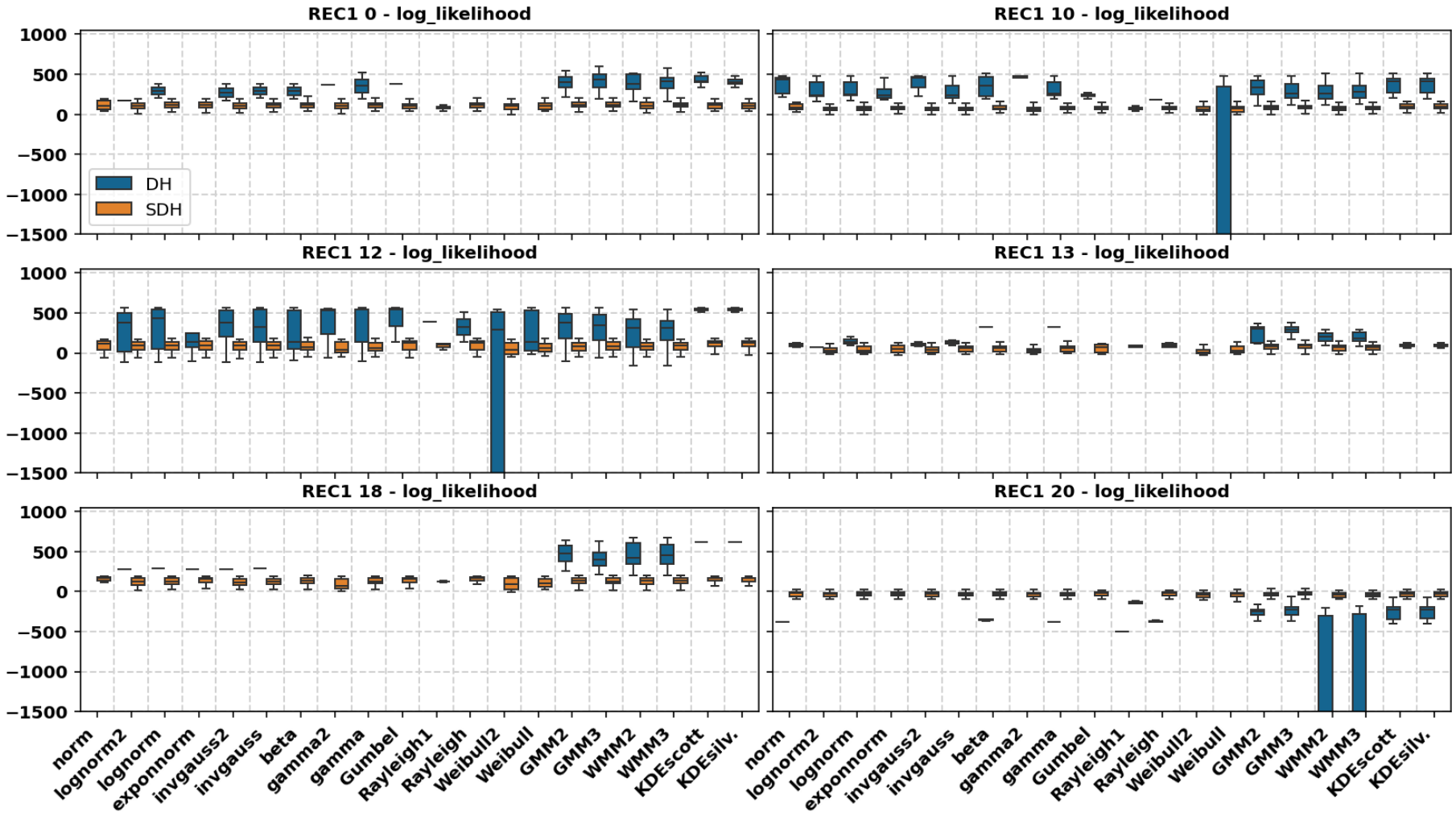


Figure 3: Log-likelihood for the **REC1** datasets (part 1 of 2).

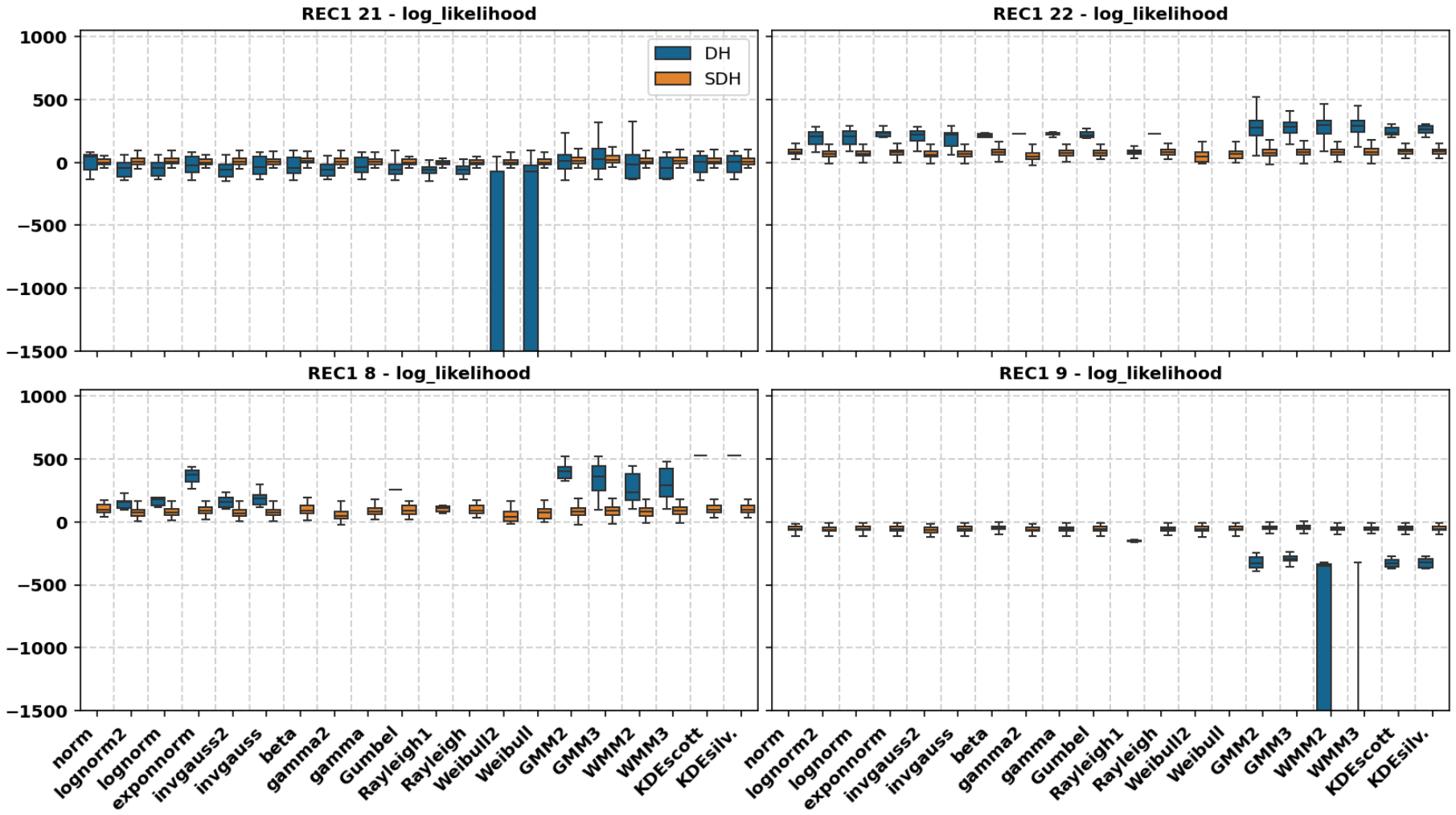


Figure 3: Log-likelihood for the **REC1** datasets (part 2 of 2).

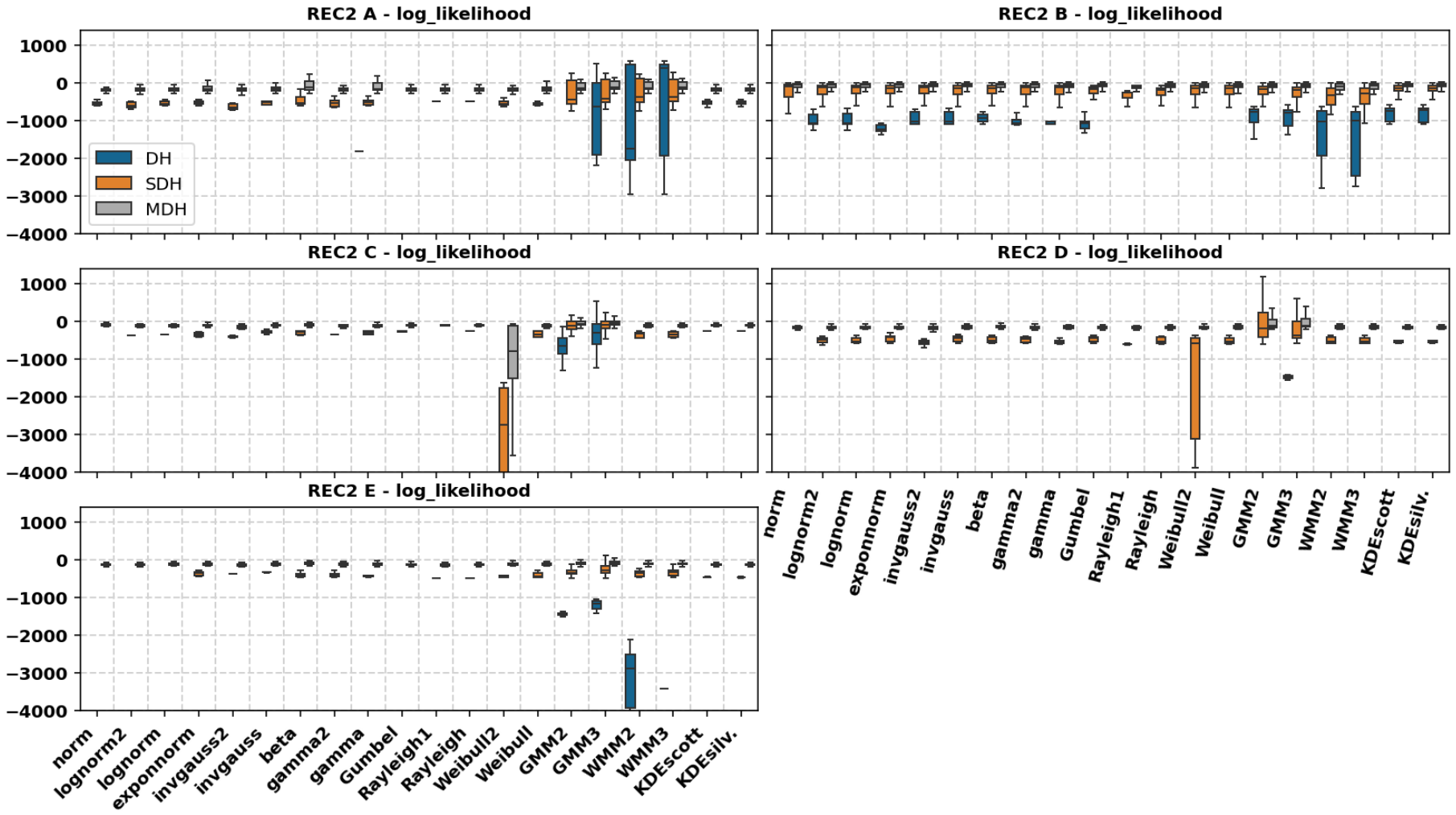


Figure 4: Log-likelihood for the **REC2** datasets.

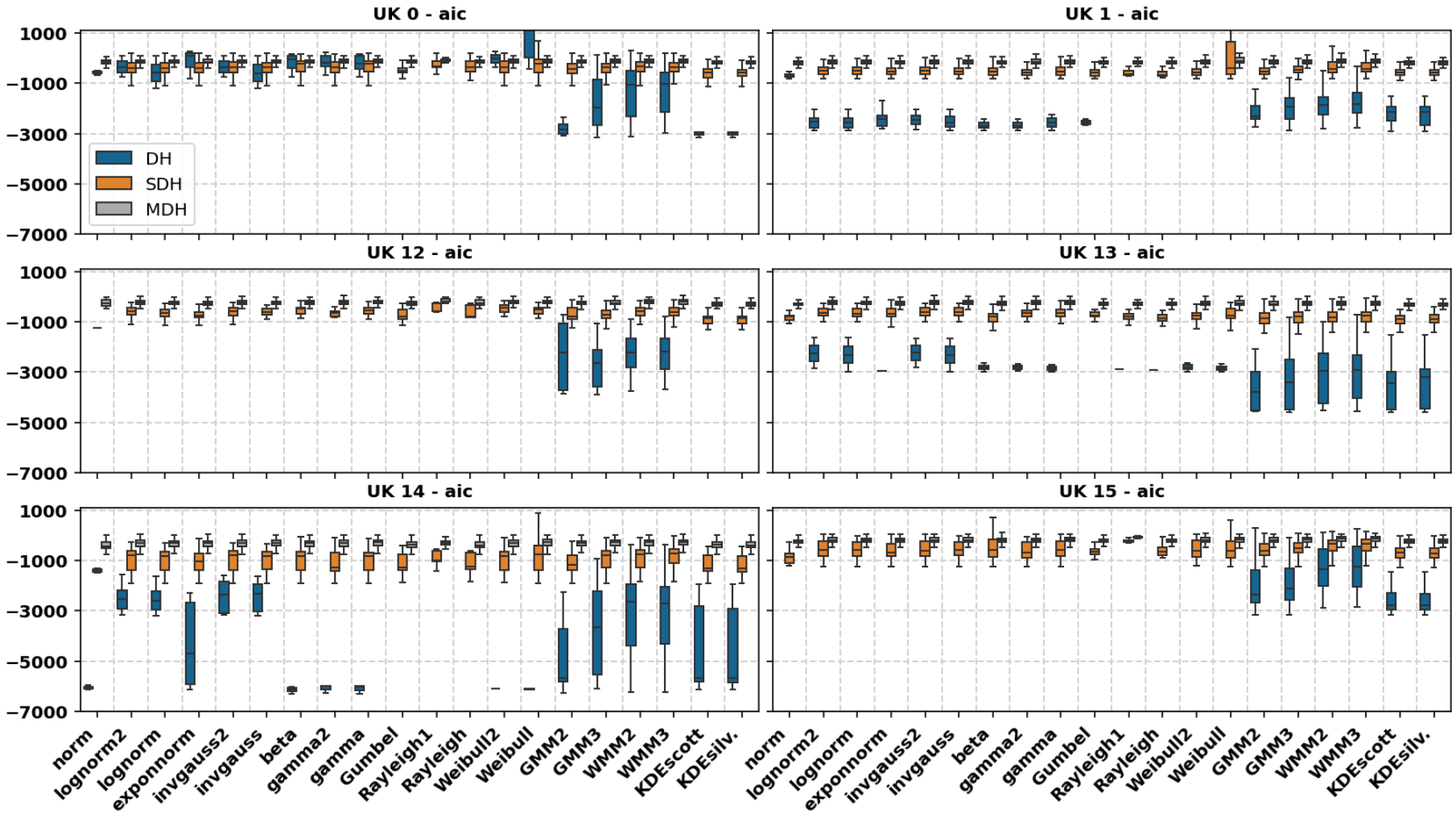


Figure 5: AIC for the **UK** datasets (part 1 of 2).

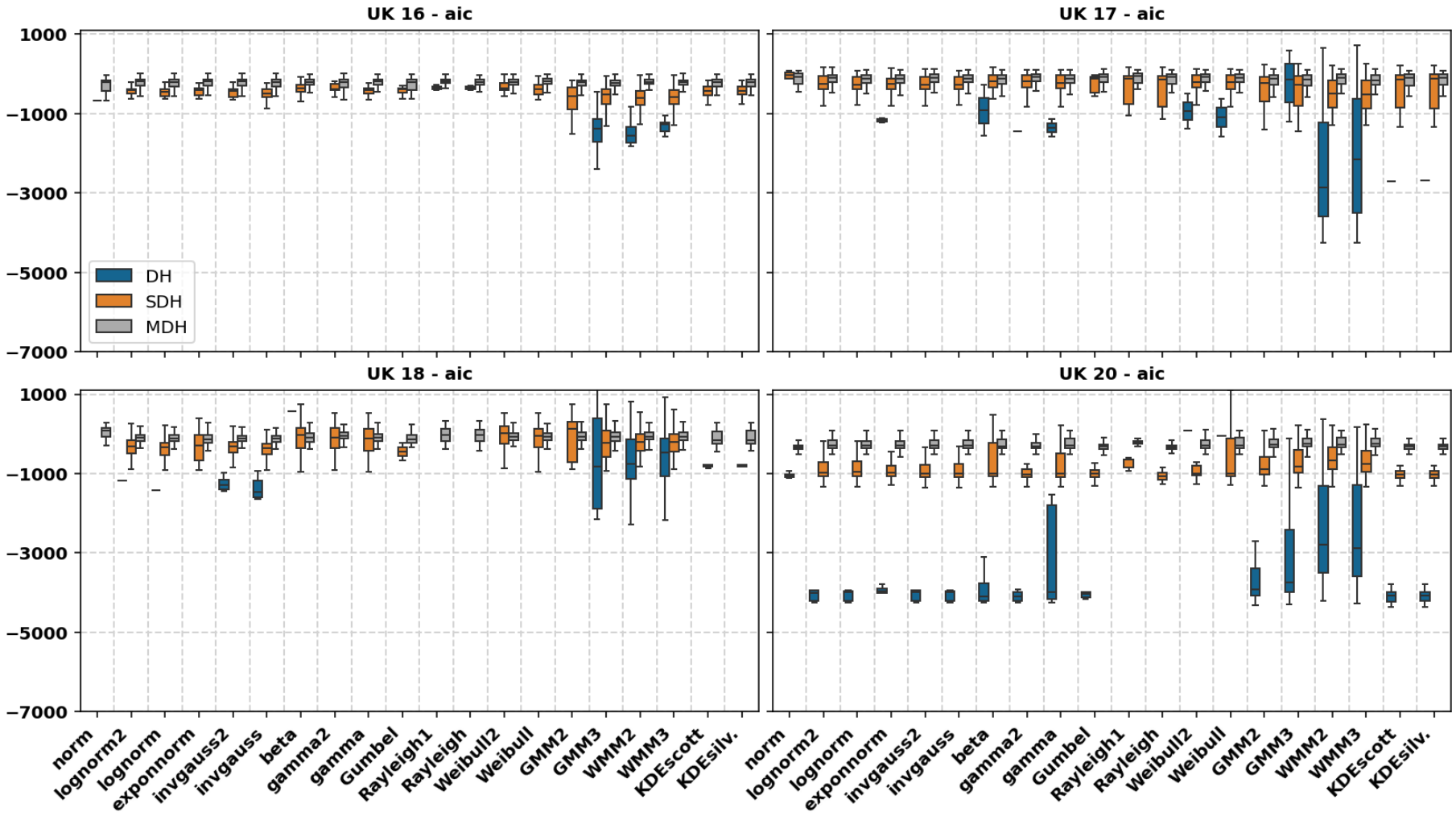


Figure 5: AIC for the **UK** datasets (part 2 of 2).

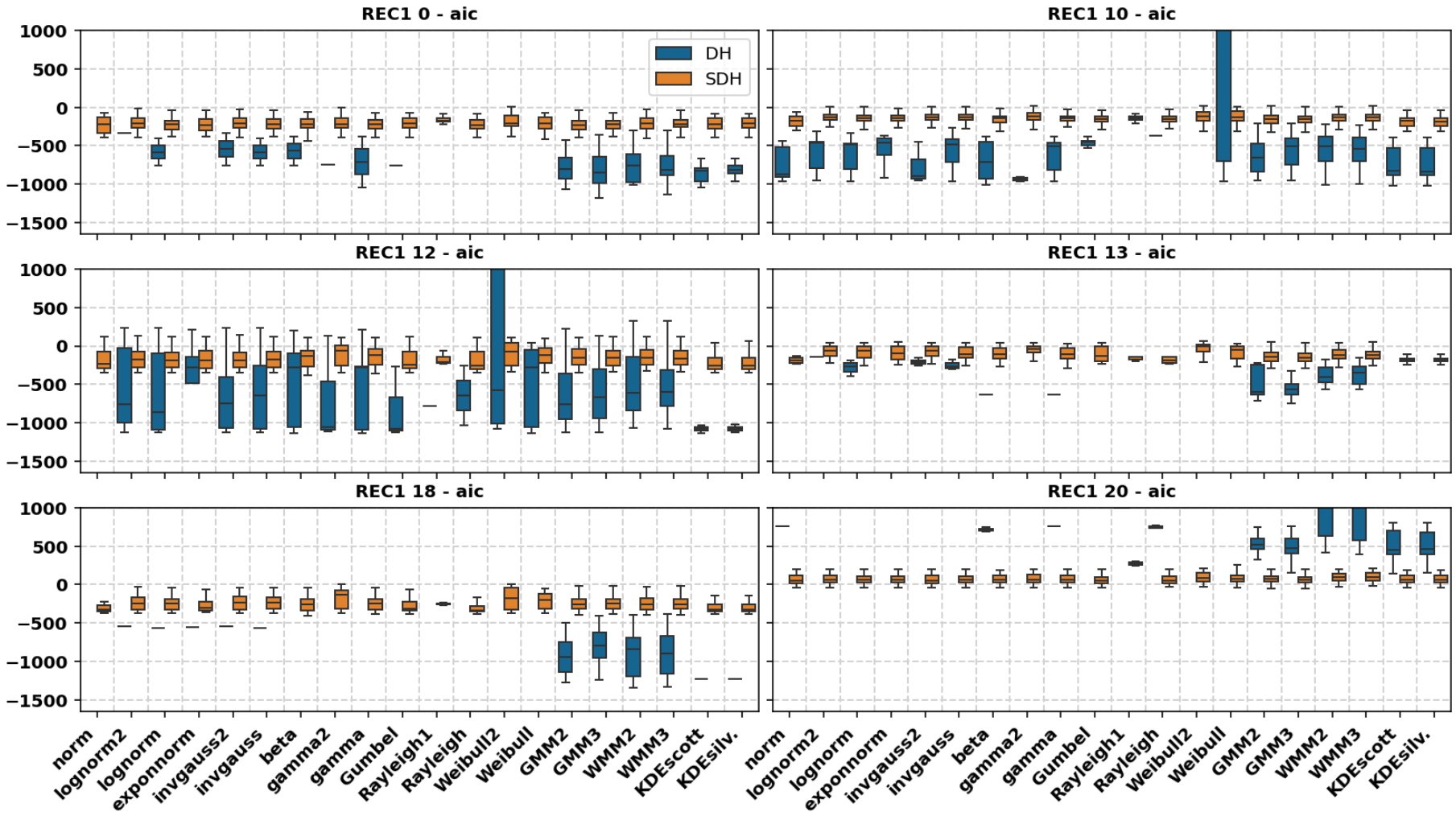


Figure 6: AIC for the **REC1** datasets (part 1 of 2).

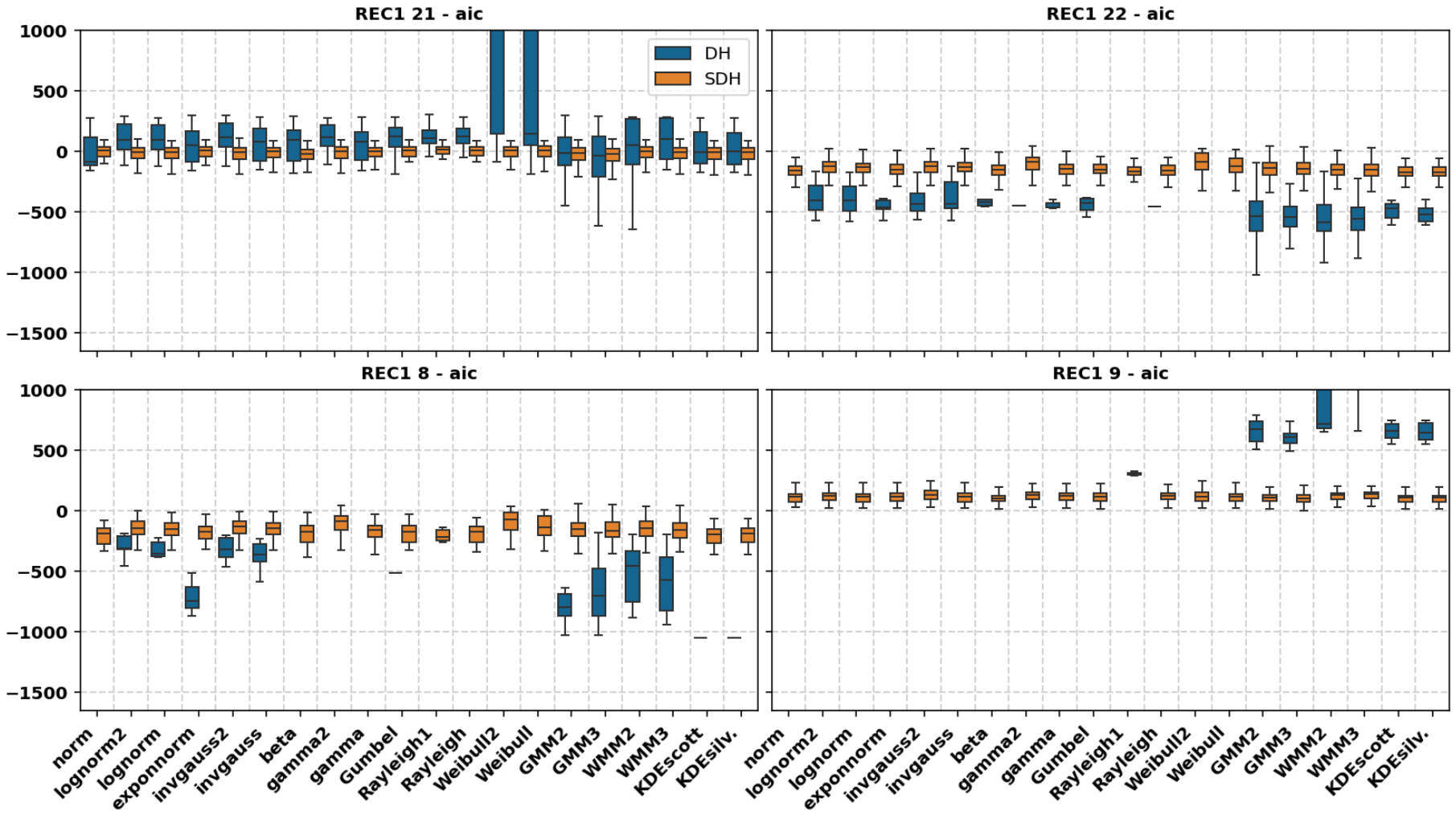


Figure 6: AIC for the **REC1** datasets (part 2 of 2).

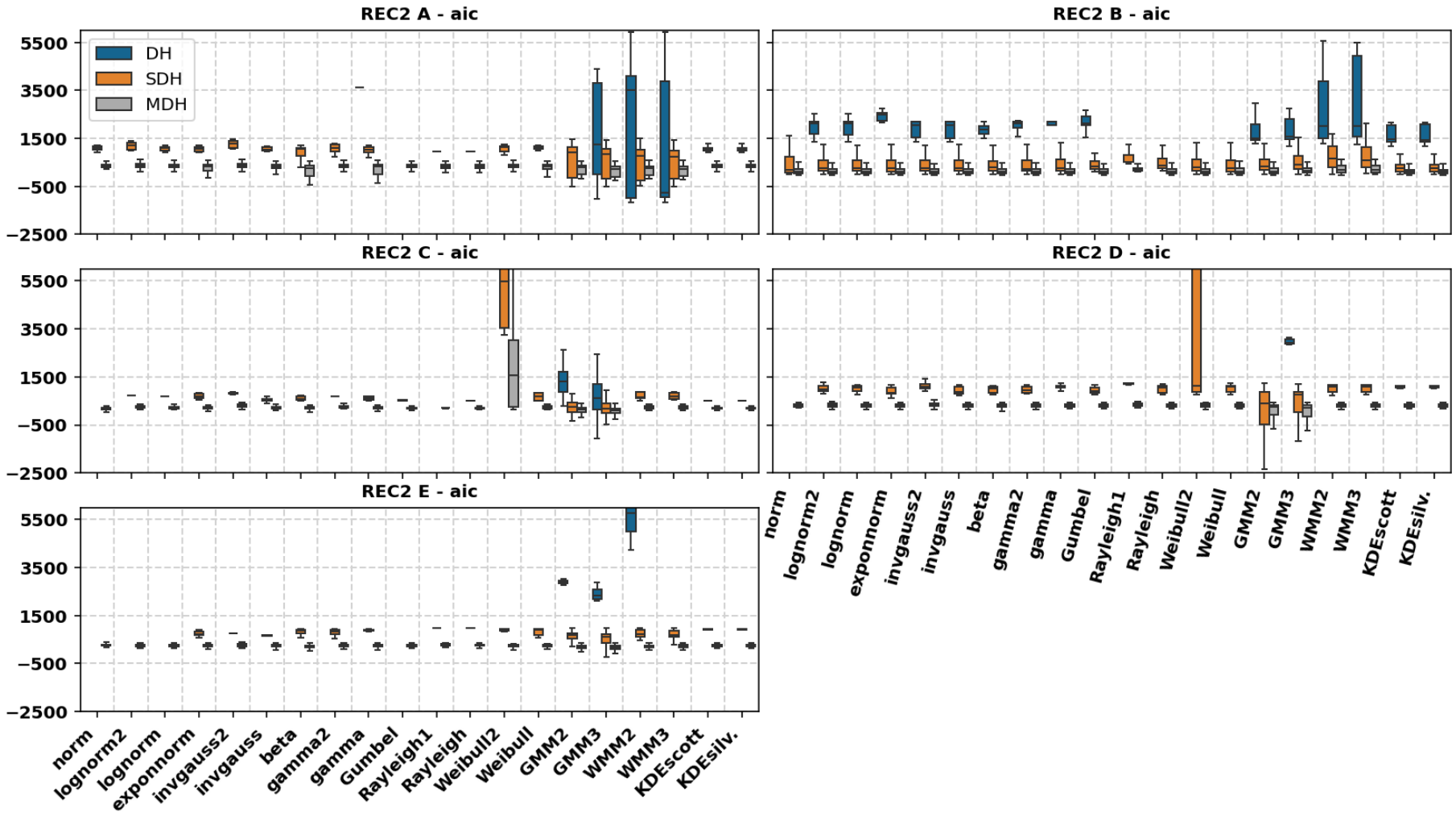


Figure 7: AIC for the **REC2** datasets.

The log-likelihood results for the UK datasets reveal notable patterns across different levels of clustering granularity (DH, SDH, and MDH). Overall, the performance can be summarized as DH > SDH > MDH, with some exceptions, such as the case where Weibull in dataset 0 performed better under SDH than DH. Additionally, increasing the clustering granularity (i.e., transitioning from DH to MDH) led to more stable log-likelihood results, as evidenced by decreasing variance.

Identifying the more competitive models for the UK datasets under the DH category was easier. For datasets 0 and 13, GMMs, WMMs, and KDE achieved the highest log-likelihood values, with WMMs standing out due to their superior KS test success rates. For datasets 1 and 20, most models (excluding norm, Rayleigh1, Rayleigh, Weibull12, and Weibull1) demonstrated competitive log-likelihood results (when considering both log-likelihood values and KS test success rates, WMMs consistently outperformed the other models). For dataset 12, GMMs outperformed WMMs in log-likelihood, but WMMs achieved higher KS test success rates, while other models failed to perform adequately. For dataset 15, GMMs and KDE had better log-likelihood values than WMMs, but the latter excelled in KS test success rates, while other models also failed to perform adequately. For dataset 14, several models, including norm, exponnorm, beta, gamma variants, Weibull variants, GMMs, and KDE, delivered high log-likelihood values; however, only exponnorm, GMMs, and KDE had KS test success rates above 2%, with GMM3 demonstrating more than double the rate of the others. For dataset 16, GMM3 achieved the best log-likelihood results and KS test success rates, while WMMs were the only other models with a KS test success rate greater than 0%. For dataset 17, WMMs outperformed all other models in both log-likelihood and KS test success rates. For dataset 18, invgauss, GMM3, and WMMs excelled in log-likelihood values, but only WMMs surpassed a 50% KS test success rate.

For the UK datasets under the SDH category, the log-likelihood discrepancy among models decreased, making it harder to identify clear winners. Despite this, some observations stand out. For dataset 0, KDE emerged as the best model, though its KS test success rate remained below 25%. Norm also performed competitively for dataset 0 but with only a 7% KS test success rate. For dataset 1, norm achieved the highest log-likelihood value but with a KS test success rate of only 1.9%, followed by KDE. For dataset 12, KDE was the top performer, though its KS test success rate was below 2.5%. For dataset 13, MMs and KDE achieved the best log-likelihood results. For dataset 14, gamma2, Gumbel, Rayleigh, and KDE stood out as the top-performing models. For dataset 15, norm was the best, albeit with a KS test success rate of only 4.5%. For dataset 16, GMM2 and WMMs were among the best performers. For dataset 17, WMMs stood out among the top models. For dataset 18, KDE achieved the best log-likelihood values but with a KS test success rate below 1%. For dataset 20, KDE models were among the top performers.

For the UK datasets under the MDH category, the log-likelihood results showed even greater stability across models, with lower variability than SDH. The differences between models were subtler. For instance, for dataset 0, norm and KDE showed slight superiority over other models. For dataset 12, Rayleigh1 was identified as the worst-performing model. For dataset 14, norm emerged as the best-performing model. For dataset 18, KDE was among the top performers. For dataset 20, norm and KDE were among the best results.

The log-likelihood results for the REC1 datasets under the DH category demonstrated significant variability across datasets and models. GMMs and WMMs consistently achieved the highest log-likelihood values, but their performance varied depending on the dataset. For dataset 0, the MMs were the standout performers, achieving the best KS test success rates and good likelihood values. While KDE and gamma models also showed good log-likelihood values, their low KS test success rates (e.g., gamma: 1.2%; KDE_scott: 3.0%; KDE_silverman: 2.4%) can limit their practical applicability. Dataset 8 followed a similar pattern, with the MMs again dominating the

log-likelihood rankings. However, while KDE and exponnorm achieved competitive log-likelihoods, they were hampered by extremely low success rates in the KS test (0.6% and 1.8%, respectively). The more difficult datasets, such as 9, 13, 18, and 20, continued to highlight the dominance of MMs in log-likelihood performance. For dataset 9, 13 out of 20 models failed completely in the KS test (not having subsequent analysis). GMM2, GMM3, and KDE provided the best log-likelihood results. The WMM models displayed significant variability in the log-likelihood results, with also lower success rates in the KS test. In dataset 13, GMM2's log-likelihood results were among the best ones but it had only a 5.4% KS test success rate, while other models like KDE completely failed. Dataset 18 showed similar trends, with the MMs outperforming other candidates, though many models achieved valid log-likelihood values with only a 0.6% KS test success rate. In dataset 20, while GMMs and KDE performed best in log-likelihood, the WMM models suffered substantial variability, undermining their overall applicability. The modeling task appeared easier for datasets 10, 12, and 21. All models except Rayleigh1 achieved valid log-likelihood values for dataset 10, with the MMs and KDE leading in performance. While the other models showed good log-likelihood values, their KS test success rates remained modest (1.8% to 7.1%). Similarly, dataset 12 saw strong performance from MMs, while most other models also produced valid log-likelihoods, albeit with low KS test success rates (0.6% to 5.4%). All tested models achieved valid log-likelihood values for dataset 21 (with success rates ranging from 4.8% to 50.0%); and log-likelihood values were generally stable across models (except for Weibull distributions). For dataset 22, the MMs again achieved the best log-likelihood values, while other models provided valid results but struggled with low KS test success rates.

The log-likelihood results for the REC1 datasets under the SDH category were generally lower than for DH, except for datasets 9, 20, and 21, where SDH outperformed DH. Notably, SDH results were more stable, with lower variability in log-likelihood values and all models achieving KS test success rates above 0%. This stability made it easier to identify the worst-performing models in certain cases. For instance, Rayleigh1 consistently performed poorly for datasets 0, 9, and 20. However, identifying clear distinctions between the best and worst models was more challenging for most datasets, as log-likelihood values were more uniform across models. Overall, the DH category showcased a broad range of performance, while the SDH category offered a more uniform performance.

The log-likelihood results for the REC2 datasets reveal distinct patterns that highlight the influence of data specificity and dataset complexity on model performance. A general trend observed across all datasets is that log-likelihood improves with increased specificity, where MDH $>$ SDH $>$ DH. This pattern aligns with the percentage of success in the KS test, which also increases with specificity. Notably, this trend differs from the UK and REC1 datasets, where SDH generally outperformed DH in terms of the KS test success rate, but log-likelihood values tended to be higher for DH (the same for SDH compared to MDH in the UK datasets). Additionally, the variability in log-likelihood results decreased with greater specificity.

The log-likelihood results for the REC2 datasets under the DH category are limited for some datasets, such as A, C, D, and E, where very few models achieved a percentage of success in the KS test above zero. Consequently, no meaningful log-likelihood values are available for these datasets. For dataset A, the WMMs and GMM3 showed the best log-likelihood performance. However, GMM3 had a success rate in the KS test of only 5.4%, and the log-likelihood values exhibited significant variability. For dataset B, considering both log-likelihood values and the percentage of success in the KS test, the MMs and KDE models emerged as the best performers. In dataset C, GMM2 achieved the best log-likelihood results, although its success rate in the KS test was less than half that of GMM3 (GMM2 and GMM3 were the only models with any KS test success).

GMM3 was the only model with a KS test success rate above zero for dataset D. For dataset E, GMM2 and GMM3 achieved the best log-likelihood results, albeit with small success rates in the KS test.

For the REC2 datasets under the SDH category, the log-likelihood results were more balanced across models than DH. In dataset A, the MMs again delivered the best results, although their log-likelihood values exhibited significant variability. For datasets C and D, GMMs consistently outperformed other models. For dataset E, MMs, particularly GMM3, achieved the best log-likelihood values.

For the REC2 datasets under the MDH category, the log-likelihood results were more balanced across models than SDH. Beta, gamma, and the MMs were among the best-performing models for dataset A. The log-likelihood results were more similar across the models for datasets B and E. For dataset C, the GMMs once again demonstrated superior performance, while Weibull2 showed significant variability in its results. For dataset D, GMMs achieved the best log-likelihood values.

As expected, the AIC results were strongly related to the log-likelihood results since AIC is directly computed from the log-likelihood. Consequently, many trends observed in the log-likelihood analysis were reflected in the AIC outcomes. However, AIC offers additional insights into the relative performance of models by penalizing complexity, which can help differentiate models with similar log-likelihoods.

The analysis of AIC values for the UK datasets highlights distinct trends across clustering granularities. Under DH, WMMs consistently demonstrated good AIC performance, though they were not always the top-performing models – for instance, GMM2 and KDE outperformed WMMs for dataset 0. For challenging datasets with limited KS test success rates (e.g., datasets 12, 15, and 16), WMMs maintained strong AIC results, though GMM2 achieved better outcomes for dataset 12 and KDE for dataset 15. Under SDH, KDE exhibited consistently good AIC values, but MMs were clearly superior for dataset 16. The norm model also performed well, especially for datasets 14 and 15. Similarly, under MDH, KDE continued to provide reliable AIC results, with the norm model also excelling, particularly in datasets 14, 15, 16, and 20.

Regarding the REC1 datasets under the DH category, for the easier datasets (10, 12, and 21), AIC values were generally more competitive across models. In dataset 10, for example, AIC results were relatively close among models, though Gumbel and Rayleigh models were clearly overcome. Similarly, for dataset 21, GMMs and WMM2 occasionally surpassed other models, with Rayleigh distributions being consistently overcome again. For the more challenging datasets (9, 13, 18, and 20), GMMs consistently emerged among the best performers. In datasets 9 and 13, GMMs achieved the lowest AIC values, underscoring their effectiveness in modeling these complex distributions. Dataset 18 also favored MMs overall, while dataset 20 highlighted GMMs and KDE as the top-performing models. For the remaining datasets, AIC results largely mirrored the trends observed for log-likelihood. For example, GMMs and WMMs were among the best models for dataset 0. Dataset 8 highlighted GMMs as the best models, while MMs dominated the AIC ranking for dataset 22.

For REC1 datasets under the SDH category, AIC results again followed similar patterns to those observed for log-likelihood. Typically, AIC values for DH were better than those for SDH, except in datasets 9, 20, and 21. Also, AIC values under the SDH category presented less variation and more similarity across models. For datasets 0, 9, and 20, Rayleigh1 was consistently overcome by other models, indicating its limitations in these cases.

For REC2 datasets under the DH category, WMM3 was the best model in terms of AIC for dataset A. GMMs, and KDE emerged as top contenders for dataset B. For datasets C and E, GMM3 stood out as the best model. Regarding the SDH category, MMs were the best performers

for dataset A, but they exhibited significant variability in AIC values compared to other models. GMMs were the best models for dataset C. GMMs were also the best models for dataset D, besides showing substantial variability in their AIC results. For dataset E, GMM3 was identified as the best model. Under the MDH category, GMMs were consistently identified as the best models for datasets C and D. However, determining the best models was more challenging for datasets A, B, and E, suggesting that no single model consistently outperformed others in these cases.

6.4. Likelihood-ratio test

Tables 10 to 12 present the percentage of cases where the likelihood-ratio (LLR) test identified a statistically significant improvement (p-value ≤ 0.05) in the fit of more complex models over their simpler counterparts. The percentages are calculated based solely on cases where the fitting process was successful, and the KS test indicated a valid model (empty cells denote instances where these conditions were not met for the models being compared). For example, the table shows the proportion of cases where the 3-parameter gamma distribution (gamma) significantly outperformed the 2-parameter gamma distribution (gamma2). This analysis highlights the situations where increased model complexity provides a measurable benefit in representing the data.

Table 10: Percentage of cases where the likelihood-ratio test favored the more complex model (p-value ≤ 0.05) for the **UK** datasets.

Cat.	D.	lognorm	invgauss	gamma	Rayleigh	Weibull	GMM2	GMM3	WMM2	WMM3
DH	0	75.0	75.0	100.0		50.0	100.0	81.8	92.7	68.0
	1	66.7	75.0	25.0		0.0	100.0	91.2	97.8	54.4
	12						100.0	100.0	99.0	65.8
	13	100.0	50.0	100.0	100.0	50.0	100.0	82.5	55.4	42.3
	14	100.0	100.0	66.7		100.0	92.9	89.3	80.6	55.5
	15						100.0	97.8	97.9	57.4
	16							100.0	100.0	90.0
	17			100.0		100.0		100.0	75.0	63.2
	18	100.0	100.0					100.0	100.0	51.0
	20	20.0	20.0	14.3		100.0	100.0	97.1	97.0	59.8
SDH	0	55.6	48.4	61.2	90.0	80.4	86.2	57.5	77.8	46.5
	1	51.0	47.4	68.2	85.7	59.7	95.4	73.0	82.0	45.1
	12	95.7	100.0	82.1	100.0	95.2	99.2	85.9	97.5	65.5
	13	62.0	53.7	67.6	96.2	57.1	94.7	55.4	37.1	33.5
	14	68.3	55.5	61.9	100.0	82.7	85.2	66.1	67.1	49.7
	15	75.3	73.4	62.8	100.0	90.9	90.8	75.4	95.3	51.2
	16	71.4	87.5	86.7	66.7	93.8	100.0	90.2	89.0	63.3
	17	74.0	77.6	98.7	80.0	92.3	94.2	84.6	67.9	45.4
	18	93.9	89.2	98.1		94.4	100.0	91.9	94.7	49.5
	20	59.6	49.6	49.6	93.8	89.6	87.5	65.2	82.8	54.3
MDH	0	38.5	36.0	49.8	81.9	64.8	68.7	34.1	37.2	33.5
	1	46.0	42.4	58.0	72.5	55.4	82.4	51.3	53.3	33.1
	12	86.4	86.2	93.1	97.0	94.6	95.8	65.8	83.1	63.2
	13	52.2	41.5	58.4	83.3	52.2	82.1	39.0	23.7	24.3
	14	56.0	51.0	66.4	98.5	77.3	77.8	41.7	46.3	35.3
	15	58.7	58.7	69.1	99.4	84.4	81.4	47.7	70.6	40.6
	16	60.9	56.5	79.0	77.0	75.2	91.1	66.3	52.4	49.8
	17	57.7	55.3	74.5	72.8	75.5	86.1	60.1	48.0	36.0
	18	72.3	60.2	92.1	51.1	91.4	98.2	65.4	73.3	38.0
	20	49.5	46.4	57.2	95.9	84.6	75.4	41.6	52.2	37.8

Overall, the LLR test results for the UK datasets favored more complex models, with the effect diminishing as clustering granularity increased. Notable exceptions include: lognorm, invgauss, and gamma for dataset 20, where their improvement over simpler counterparts was more pronounced in SDH than DH; and Rayleigh for datasets 16 and 20, where its gains were greater in

Table 11: Percentage of cases where the likelihood-ratio test favored the more complex model (p-value ≤ 0.05) for the **REC1** datasets.

Cat.	D.	lognorm	invgauss	gamma	Rayleigh	Weibull	GMM2	GMM3	WMM2	WMM3
DH	0	100.0	100.0	100.0		0.0	100.0	91.6	71.4	50.0
	10	83.3	100.0	83.3	100.0	80.0	94.1	79.6	84.4	38.1
	12	88.9	100.0	100.0	100.0	100.0	100.0	94.4	84.4	50.0
	13	100.0	100.0	100.0		100.0	100.0	100.0	92.6	50.0
	18	100.0	100.0				100.0	96.9	73.5	56.8
	20			0.0	100.0		100.0	77.5	23.1	73.8
	21	50.0	81.5	88.6	94.4	91.7	85.0	54.8	72.7	30.6
	22	100.0	70.0	100.0	100.0	66.7	100.0	81.3	96.3	55.3
	8	100.0	100.0			0.0	100.0	90.3	97.4	52.6
	9						100.0	79.2	33.3	50.0
SDH	0	75.8	77.2	79.2	100.0	84.6	88.2	62.5	50.3	33.6
	10	64.6	60.6	71.4	88.4	66.1	82.0	62.1	55.0	35.9
	12	70.6	69.1	82.1	88.5	91.1	87.2	58.7	65.0	38.8
	13	63.2	60.8	85.0	83.3	87.3	95.3	76.6	69.5	37.4
	18	64.8	63.6	58.8	100.0	63.8	85.8	63.4	54.1	37.9
	20	47.4	46.8	55.0	100.0	71.5	75.6	43.5	30.9	32.0
	21	38.3	46.2	52.1	73.4	67.0	67.3	44.1	37.8	30.1
	22	66.7	70.6	79.4	80.3	77.0	89.8	57.9	68.9	39.7
	8	78.0	81.6	81.6	100.0	78.0	90.9	68.8	76.2	34.0
	9	63.6	66.5	73.1	100.0	87.0	81.8	63.5	3.3	3.3

Table 12: Percentage of cases where the likelihood-ratio test favored the more complex model (p-value ≤ 0.05) for the **REC2** datasets.

Cat.	D.	lognorm	invgauss	gamma	Rayleigh	Weibull	GMM2	GMM3	WMM2	WMM3
DH	A			0.0				100.0	100.0	50.4
	B	66.7	60.0	60.0			100.0	87.5	91.4	56.4
	C						100.0	100.0		
	D							100.0		
	E						100.0	100.0	100.0	100.0
SDH	A	100.0	100.0	100.0	100.0	100.0	95.6	84.4	98.6	48.8
	B	48.9	39.3	39.1	93.8	72.3	72.4	50.9	63.7	35.4
	C	100.0	100.0	66.7	100.0	80.0	100.0	96.5	66.7	33.3
	D	100.0	100.0	73.3	100.0	68.2	100.0	89.9	3.6	20.7
	E		100.0	0.0	100.0	16.7	100.0	96.1	82.8	48.4
MDH	A	61.3	59.5	69.8	82.6	63.2	74.4	61.9	74.3	33.9
	B	41.3	36.2	47.5	89.7	64.7	63.2	41.6	37.9	20.5
	C	64.3	96.2	68.1	88.1	73.6	98.2	86.9	73.6	38.7
	D	50.5	80.3	56.7	77.6	52.5	93.1	67.4	28.4	16.8
	E	75.7	94.0	53.5	83.5	49.1	95.5	73.8	80.9	37.6

MDH than SDH. Among the models, GMM2 demonstrated the most substantial improvement over its simpler counterpart (norm), with more than 90% of cases showing improvement under DH, over 85% under SDH, and over 68% under MDH. Conversely, WMM3 typically outperformed WMM2 in fewer than 50% of cases under MDH, except for dataset 12, reflecting varying levels of benefit from increased model complexity across different datasets and clustering granularities.

For the REC1 datasets, the LLR test results generally favored the more complex models, particularly for the DH category. However, there were notable exceptions. For example, in datasets 0 and 8 under DH, the Weibull model showed no improvement over Weibull2; and in dataset 20 under DH, the gamma model also showed no improvement over gamma2. Similarly, in dataset 9 under SDH, the improvements were minimal: 3.3% of the cases for WMM2 over Weibull2 and 3.3% for WMM3 over WMM2. The results were not particularly strong in some cases, such as 23.1% of the cases with improvement of WMM2 over Weibull2 for dataset 20 under DH; 33.3% of the cases with improvement of WMM2 over Weibull2 for dataset 9 under DH; 38.1% of the cases with improvement of WMM3 over WMM2 for dataset 10 under DH; and 30.6% of the cases with improvement of WMM3 over WMM2 for dataset 21 under DH. Under SDH, datasets 20 and 21 showed also moderate improvements for lognorm over lognorm2; invgauss over invgauss2; GMM3 over GMM2; and WMM2 over Weibull2. Improvements for WMM3 over WMM2 under SDH were also moderate across all datasets. Overall, while the results favored more complex models, the level of improvement varied considerably by dataset, category, and model.

The LLR test results for the REC2 datasets generally favored more complex models too. However, there were notable instances where simpler models outperformed or matched their more complex counterparts, challenging the assumption that added complexity always yields better fits. In some cases, the simpler models consistently outperformed their complex versions. For example, in dataset A under DH and dataset E under SDH, the gamma model failed to show any improvement over its simpler variant gamma2, with 0% of cases indicating superiority. Similarly, in dataset D under SDH, WMM2 only improved upon Weibull2 in 3.6% of the cases. There were also less extreme cases where simpler models were favored. For instance, WMM3 under the SDH and MDH categories often did not outperform WMM2, with improvement observed in only 16% to 48.8% of cases. A similar trend was seen in dataset B under SDH and MDH, where models like lognorm, invgauss, and gamma failed to be superior to their simpler counterparts in most cases. Another example is dataset E under SDH and MDH, where Weibull did not demonstrate superiority over Weibull2 in the majority of cases.

These findings emphasize that while complexity often leads to better performance, it is not universally advantageous. The success of simpler models in some cases highlights the importance of balancing complexity with parsimony, especially in datasets or conditions where additional parameters do not substantially improve the model fit.

6.5. Visual assessment

Figs. 8 to 15 illustrate data histograms and fitted PDFs for the datasets (UK, REC1, and REC2) across the clustering granularities (DH, SDH, and MDH) analyzed in our experiments. For each dataset (energy consumption profile) and clustering granularity, we identified the cluster with the best fitting for a particular distribution. The best fitting was determined based on a distribution that successfully passed the KS test and ranked in the top in terms of log-likelihood and AIC values. For the selected cluster, we plotted the other distributions that also passed the KS test, with their order in the plot reflecting a rank based on log-likelihood and AIC values. To enhance the clarity of the visualizations, we limited the graphs to the top five distributions for each cluster.

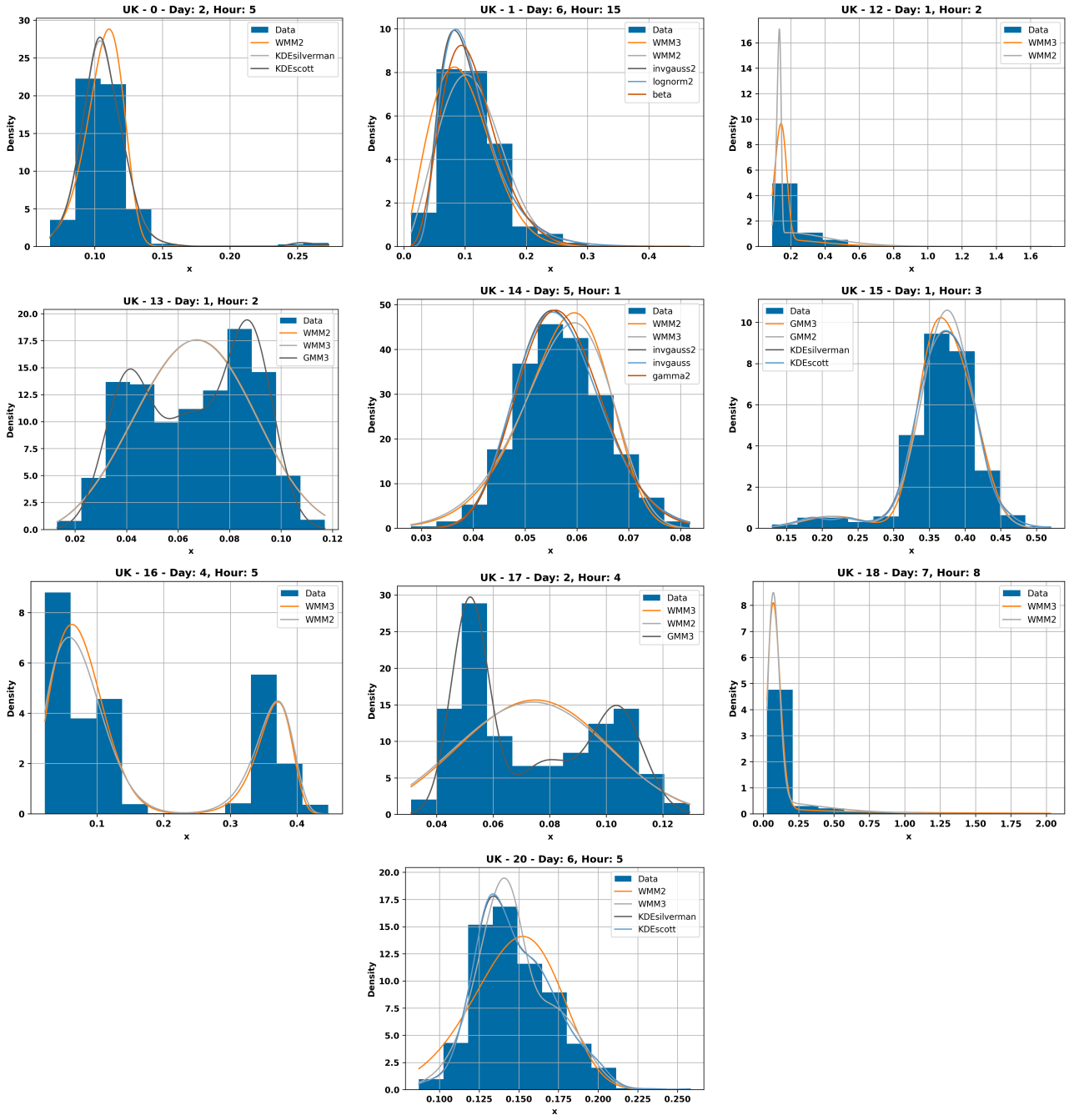


Figure 8: Examples of histograms of electricity use data along with fitted distribution(s) for the **UK** datasets under the category **DH**.

The examples of distribution fittings presented in Figs. 8 to 15 illustrate the diversity of data shapes and the varying performance of models across datasets and clustering granularities. The histograms reveal a wide variety of energy consumption patterns, highlighting the importance of testing multiple probability distribution models to accommodate the varying characteristics of the data. Notably, all 20 tested models appear in these plots, demonstrating that each achieved good fits in at least a few scenarios. However, the WMMs frequently emerged among the selected distributions for the UK datasets, while GMM3 consistently stood out for the REC1 and REC2 datasets.

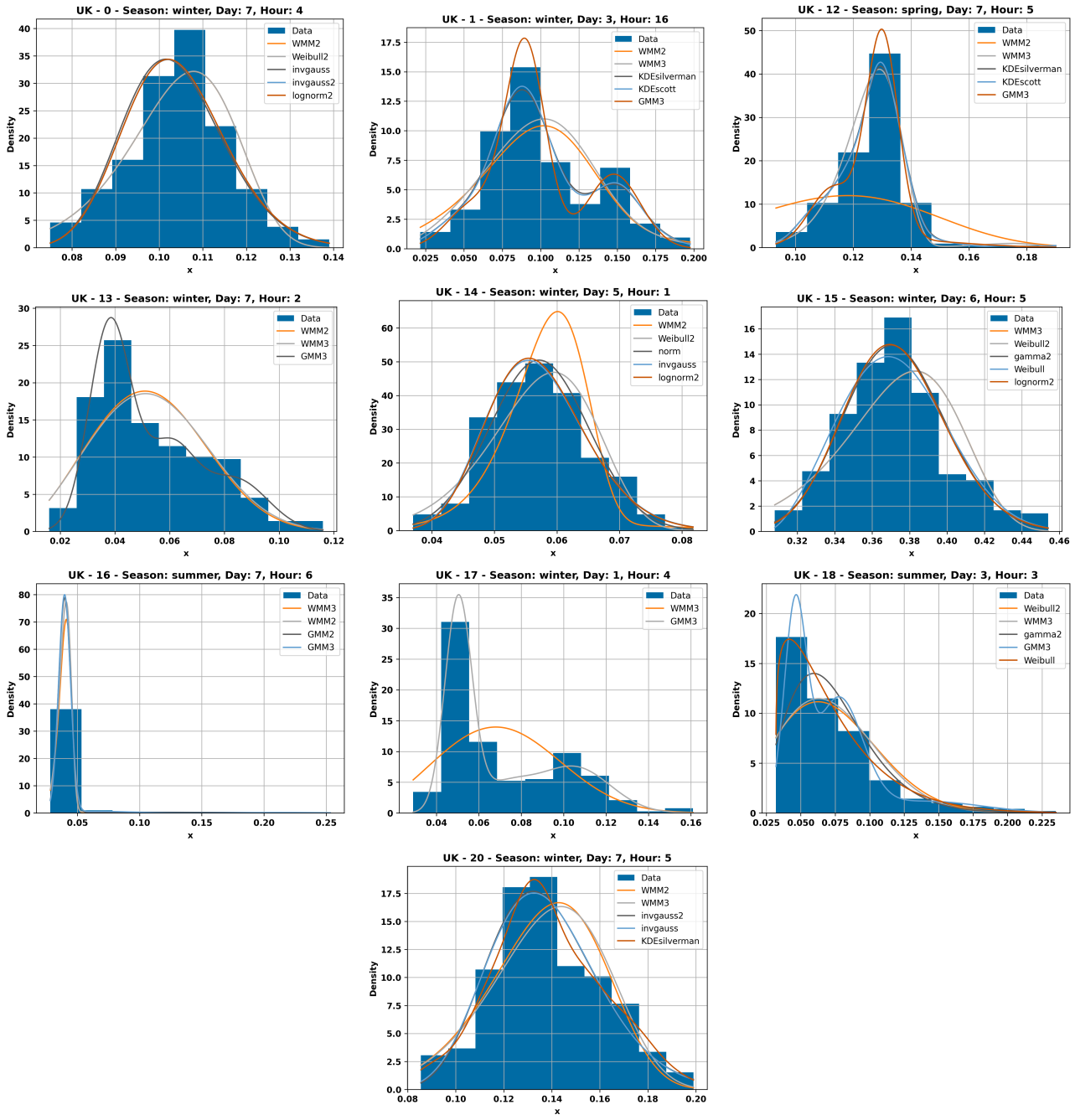


Figure 9: Examples of histograms of electricity use data along with fitted distribution(s) for the **UK** datasets under the category **SDH**.

For the UK and REC1 datasets, which consist entirely of household energy consumption profiles, we observed that the clusters with the best fitting for a specific distribution were most frequently associated with periods when individuals are typically asleep, specifically between midnight and 6 a.m. During these hours, energy consumption patterns tend to be more stable and less variable, likely due to minimal household activity. This reduced variability may explain why these clusters were selected – as the underlying data exhibited smoother and more predictable patterns compared to clusters representing periods of higher activity during the day.

For the REC2 datasets, which consist of energy consumption profiles from workshops (small

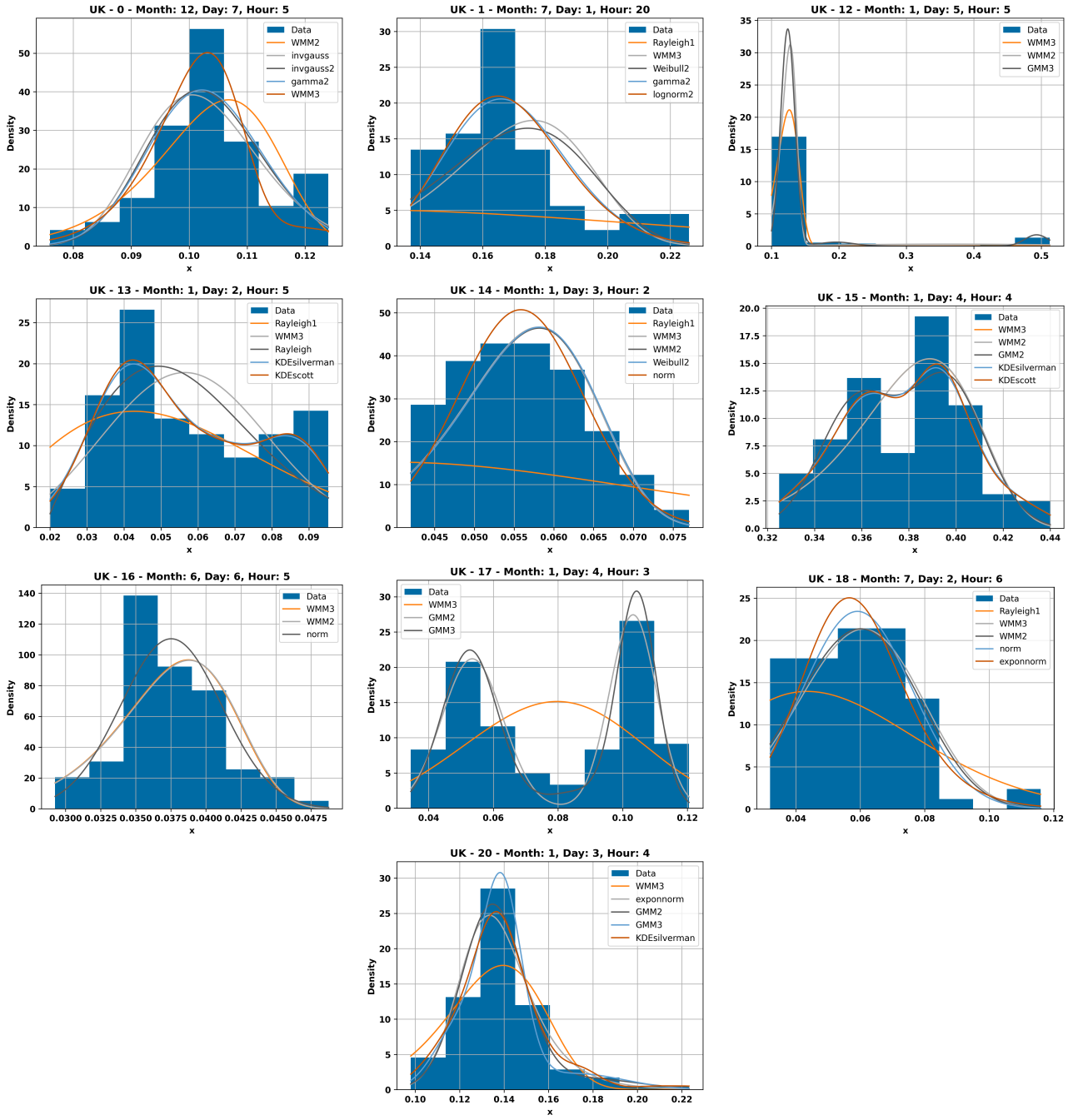


Figure 10: Examples of histograms of electricity use data along with fitted distribution(s) for the **UK** datasets under the category **MDH**.

businesses), we observed that the clusters with the best fitting for a specific distribution predominantly corresponded to periods when the businesses were likely closed – such as Sundays or outside typical business hours on weekdays. Although we lack precise information about the operating schedules of these businesses, this pattern also suggests that periods of minimal activity, characterized by lower and more stable energy consumption, offered the more predictable patterns.

It is important to note that these observations pertain specifically to the clusters identified as the best-fitting for each energy profile. This does not imply that good fittings were absent for clusters corresponding to periods when individuals are not typically asleep or when businesses are

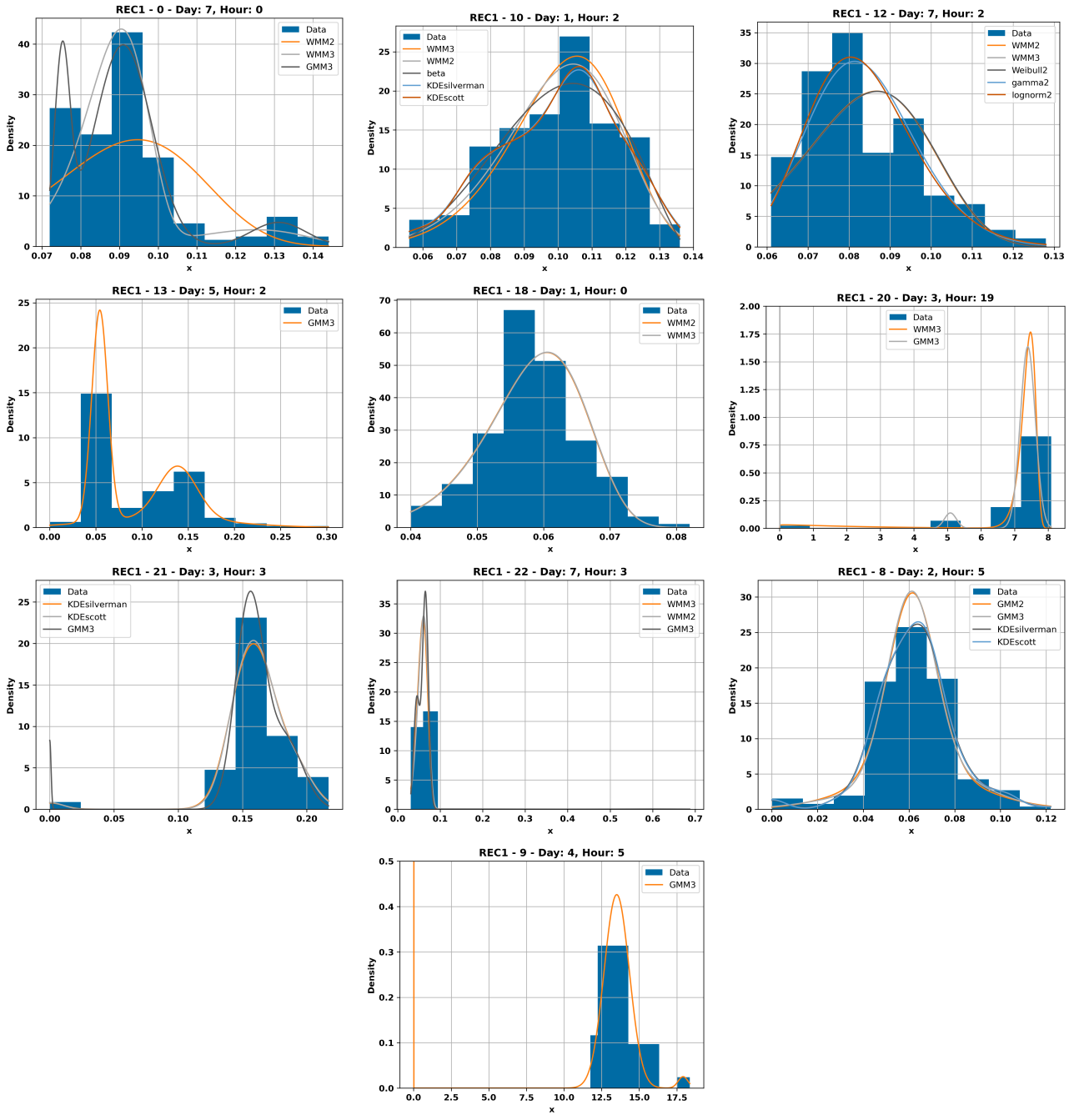


Figure 11: Examples of histograms of electricity use data along with fitted distribution(s) for the **REC1** datasets under the category **DH**.

likely open. In fact, several models demonstrated good performance across a range of clusters representing different times of day and activity levels. For both the UK and REC1 datasets across the three clustering granularities, we observed cases where the majority of the best-fitting clusters did not correspond to periods when individuals are typically asleep. Similarly, for the REC2 datasets, some energy profiles exhibited the majority of their best-fitting clusters during periods when businesses are likely open. Thus, strong model performance is not strictly confined to periods of minimal activity, such as late-night hours. However, the consistency of stable energy consumption patterns during these periods made these clusters particularly favorable for achieving

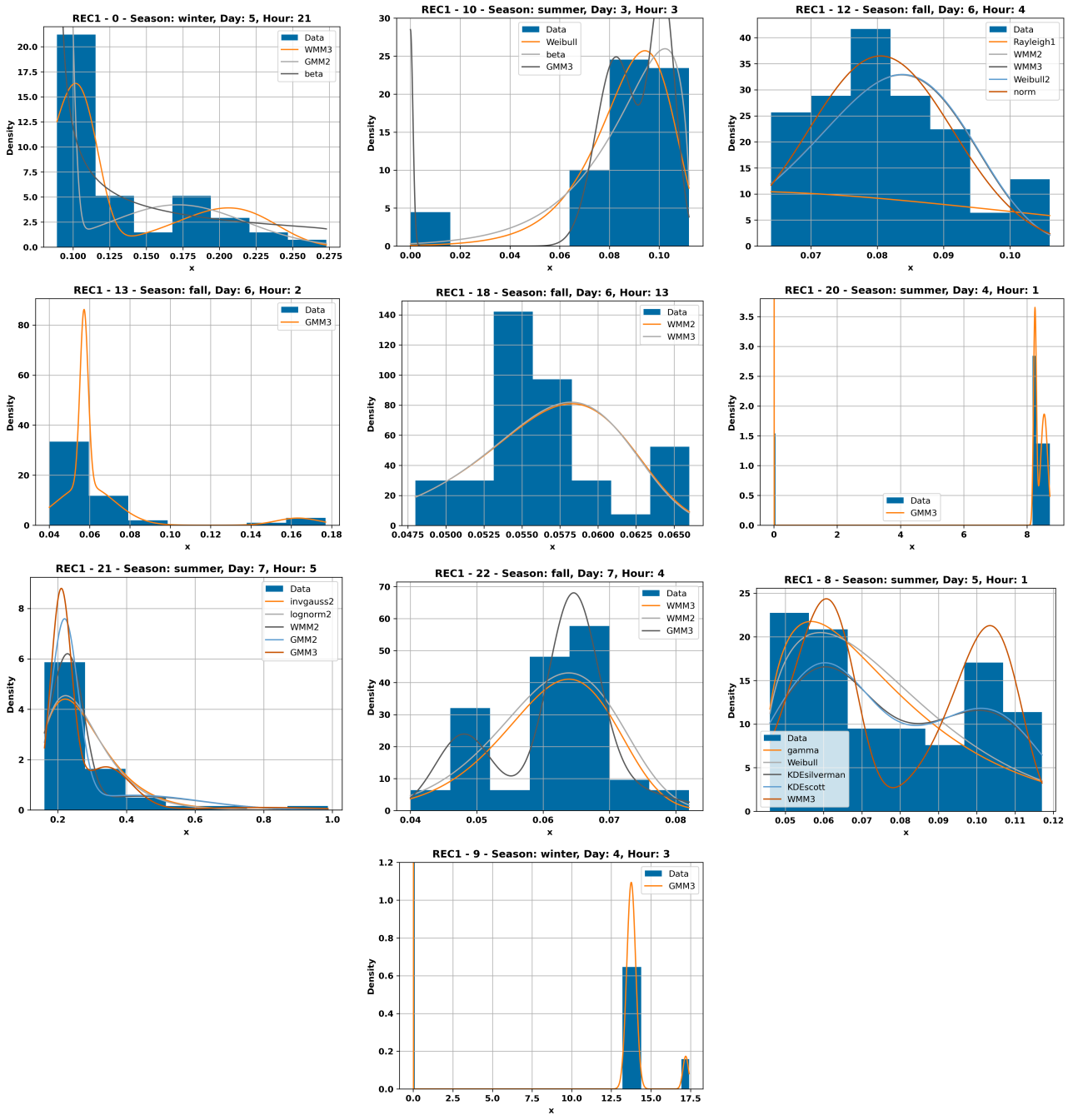


Figure 12: Examples of histograms of electricity use data along with fitted distribution(s) for the **REC1** datasets under the category **SDH**.

the best fits in our analysis.

7. Discussion

7.1. Clustering granularity

We analyzed the data under three levels of temporal resolution: DH, SDH, and MDH, corresponding to different cluster sizes. For instance, for the REC2 datasets, clusters range from 728 to 732 instances under DH, 152 to 212 instances under SDH, and 48 to 76 instances under MDH. So,

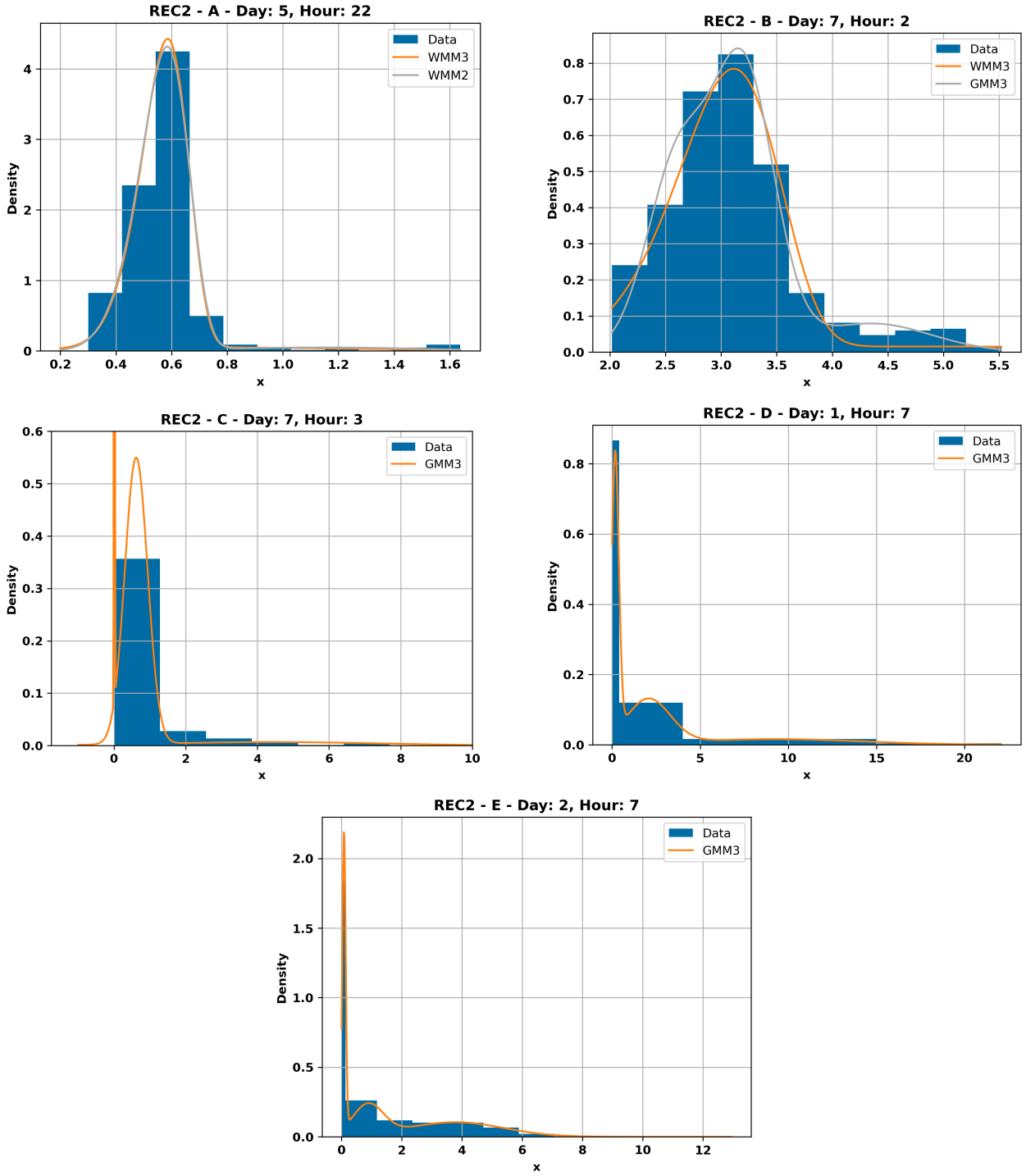


Figure 13: Examples of histograms of electricity use data along with fitted distribution(s) for the **REC2** datasets under the category **DH**.

we need to interpret the results concerning the number of instances used during the distribution fitting. Larger datasets can provide more information for parameter estimation, enabling models to capture underlying patterns better, while smaller datasets may limit the reliability of statistical tests.

The sample size influences the sensitivity of the KS test. The test becomes more sensitive

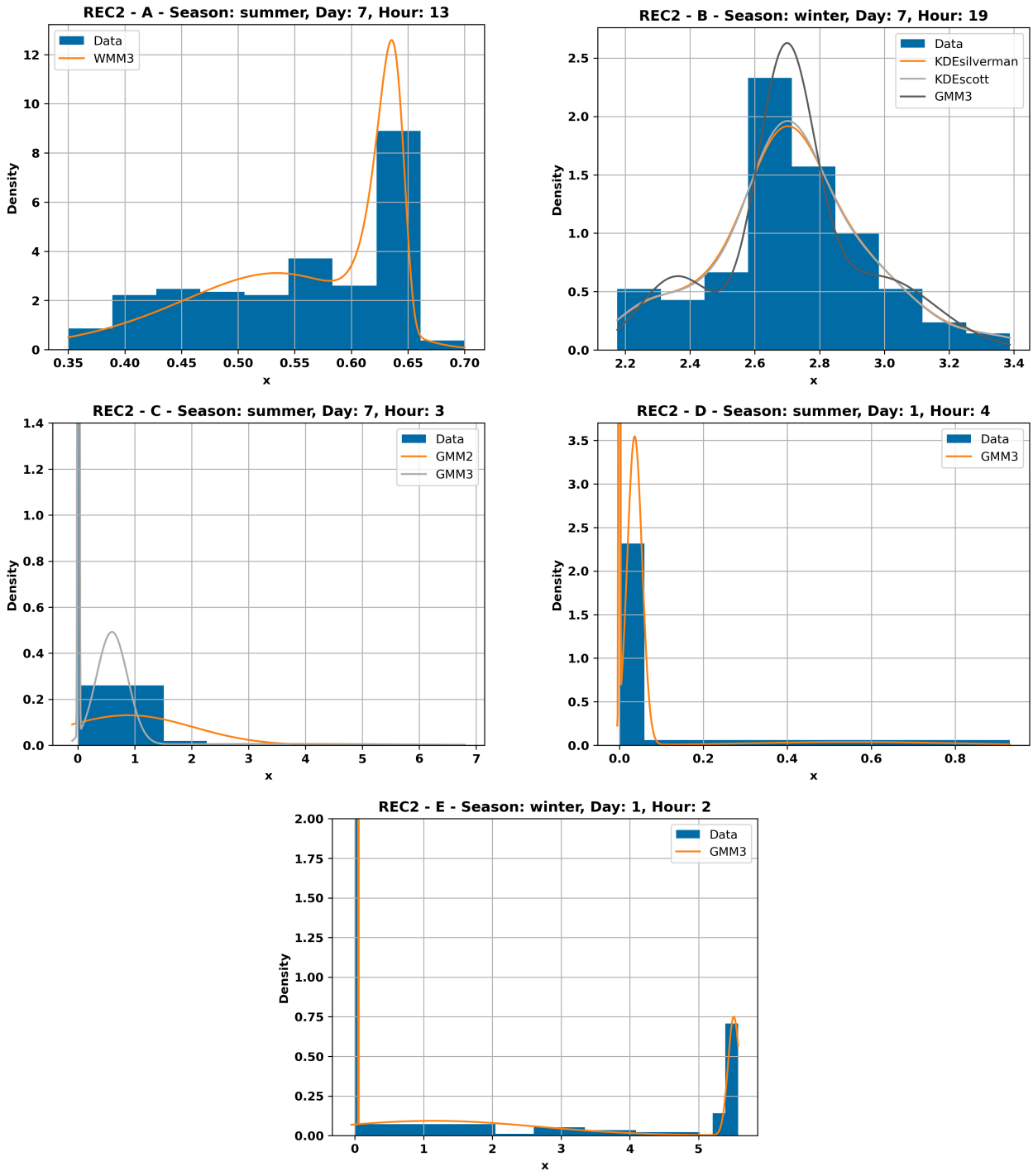


Figure 14: Examples of histograms of electricity use data along with fitted distribution(s) for the **REC2** datasets under the category **SDH**.

with more instances. Thus, for a small number of instances, the lack of sensitivity may allow distributions to pass the test even if they are not ideal.

The bootstrap moderates the tendency of the standard KS test to over-penalize models in large datasets because it considers parameter estimation variability. However, although the bootstrap improves the accuracy of the p-value by adjusting for parameter estimation, it does not entirely eliminate the influence of dataset size. For smaller datasets, the lack of sensitivity may allow distri-

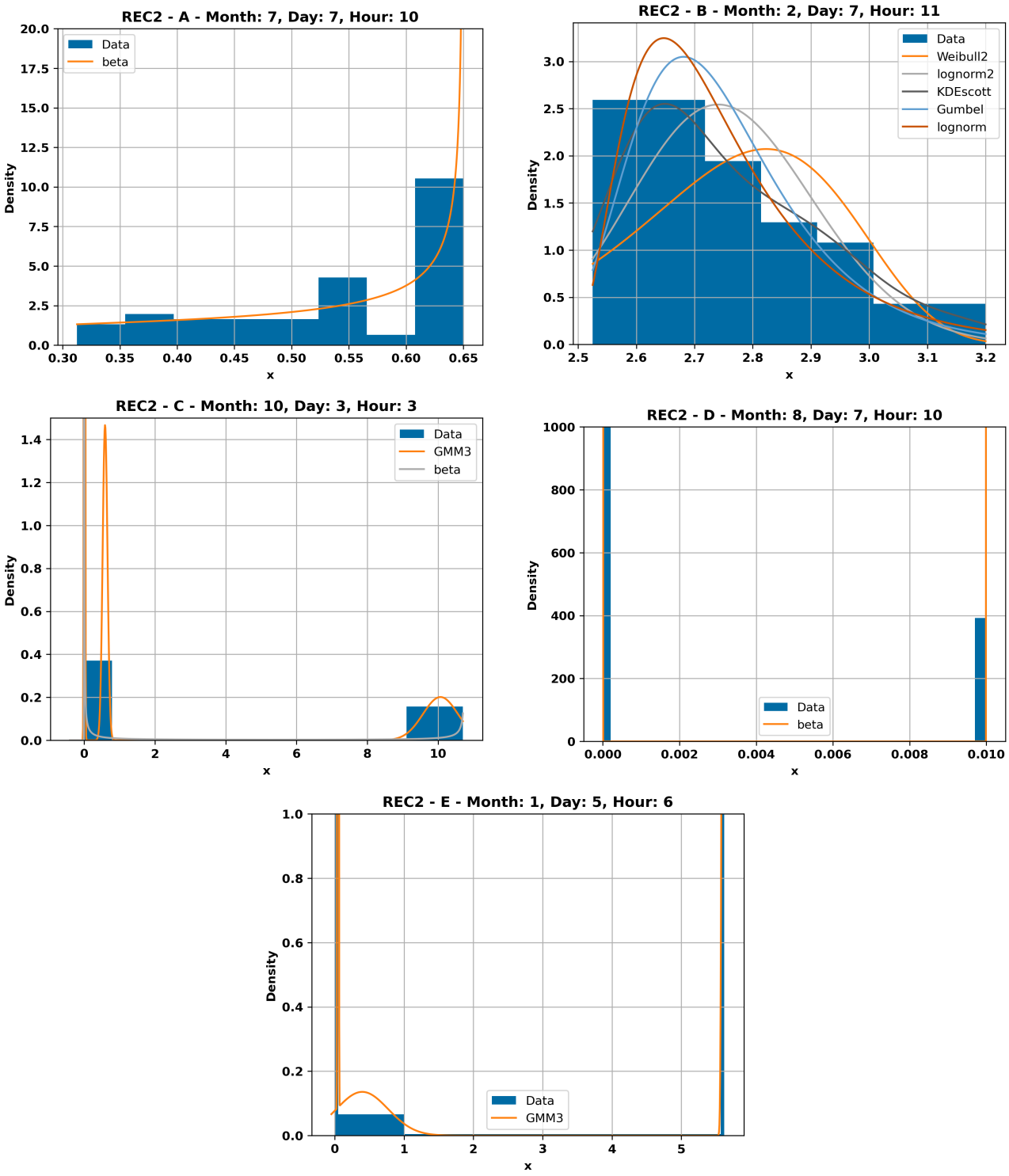


Figure 15: Examples of histograms of electricity use data along with fitted distribution(s) for the **REC2** datasets under the category **MDH**.

butions to pass the test even if they are suboptimal, as the resampling process generates datasets with broader variability around the null hypothesis. This broader variability means that larger deviations between the empirical and theoretical CDFs may still result in ($p > 0.05$), reducing the test's power to reject poorly fitting models. In contrast, larger datasets reduce the variability in re-

sampling, leading to more consistent p-values and enabling the detection of smaller deviations. So, while the bootstrap procedure accounts for parameter estimation variability, it does not eliminate the effect of dataset size, with larger datasets still resulting in stricter evaluations.

In our results, KS test pass rates were higher for MDH compared to SDH and higher for SDH compared to DH. However, this does not necessarily indicate better fits for MDH, as the reduced sensitivity in smaller datasets can mask subtle but meaningful deviations. On the other hand, the stricter evaluations in larger datasets may reject models for minor mismatches that are not practically significant. This highlights the importance of interpreting KS test results alongside other metrics to avoid misleading conclusions.

Metrics like log-likelihood and AIC can also benefit from larger sample sizes. Larger datasets provide more data for parameter estimation, improving the models' ability to capture the underlying data structure. Consequently, log-likelihood values can be higher, and AIC scores – balancing fit and model complexity – tend to improve as the penalty for additional parameters becomes less significant with increased data. These metrics indicate that models fitted to larger datasets can be more accurate overall, even if the stricter KS test evaluations reject some models for minor discrepancies.

Our findings highlight the necessity of using multiple metrics to evaluate the GoF comprehensively. While the KS test pass rates for smaller datasets may reflect reduced sensitivity rather than genuinely better fits, the superior log-likelihood and AIC results indicate better alignment with the data's underlying structure. Combining metrics like the KS test, log-likelihood, and AIC ensures a balanced assessment of model performance, particularly when comparing datasets of varying sizes. This multi-faceted approach provides a better understanding of model behavior across different temporal resolutions.

With all that said, and despite the intuitive expectation that a finer clustering would lead to better results, our findings revealed that DH outperformed SDH for the UK and REC1[7] datasets (with the exceptions of datasets 9 and 20 of REC1 and a less prominent exception for dataset 21 of REC1), and SDH outperformed MDH for the UK datasets. It is important to note that in real-world applications, more data could be collected over time, potentially altering these findings. However, for REC2 datasets and REC1 datasets 9 and 20, the opposite trend was observed. This suggests that, despite the limited amount of data in these cases, the optimization routine (used for parameter estimation) produced more accurate distribution fittings for the finer clustering. It suggest that none of the 20 tested models possessed sufficient flexibility to adequately capture the data patterns inherent in the coarser clustering categories, particularly when contrasted with the improved performance observed under finer clustering. This observation is further supported by the fact that the most complex models were the only ones to achieve some success for these datasets particularly under the DH category.

While coarser clustering outperformed finer clustering in the aforementioned cases, it is important to note that this approach also led to a higher rate of KS test failures. This is likely due to the increased heterogeneity within broader clusters, making it more challenging for distributions to achieve statistically significant fits. Therefore, a practical strategy for practitioners is to initially apply a coarser clustering approach to identify general consumption patterns and assess model performance. If the KS test fails or the fitted distributions do not adequately capture the data, refining the clustering granularity for these cases tends to help achieve better results by isolating

[7]We did not analyze REC1 datasets under the MDH category, as the resulting clusters would contain too few data points (e.g., clusters with only four data points), making such analysis infeasible.

more homogeneous consumption patterns. However, the user must consider the amount of data available since finer clustering reduces the number of data points per cluster, which can hinder the ability to obtain reliable parameter estimates and stable model fits.

7.2. Models

Computational resources are crucial when selecting models to fit data, especially when conducting robust evaluations such as those performed in this work. The accuracy of bootstrap-based KS tests, for instance, is influenced by the number of resampling iterations, which can substantially increase computational demands. While a detailed runtime analysis was not the primary focus of this study – due to the use of multiple machines and implementations – the significant variations in computational cost among models cannot be ignored. During our experiments, we observed a substantial computational burden associated with fitting and evaluating certain models, particularly the WMMs and beta distributions. For a general understanding of runtime differences among the tested models, readers can refer to Table A.16 in the Appendix A, which provides the runtime of experiments conducted on the Dragon1 machine for the UK datasets under the MDH category.

The fitting speed of a model depends on several factors, including the complexity of the model, the optimization procedure, and the data itself. Although our WMM implementation may not be fully optimized, models like GMMs generally fit faster due to the simplicity of the Gaussian PDF, efficient closed-form updates in the EM algorithm, and reduced reliance on numerical optimization compared to WMMs. These factors make GMMs a more accessible choice than WMMs when computational resources are limited. Thus, the user's decision to use a particular model should weigh resource availability against the model's expected performance.

It is also challenging to provide definitive recommendations because, for certain datasets, multiple models performed similarly. However, we attempt to outline some general guidelines for model selection based on the comprehensive set of metrics employed in this study, including fitting success, KS test outcomes, log-likelihood, AIC, and LLR tests.

Overall, the mixture models (both GMMs and WMMs) demonstrated strong performance across all tested scenarios, regardless of dataset group or clustering granularity. This success can likely be attributed to their multimodal nature, which allows them to effectively model energy consumption patterns with multiple peaks. In our experiments, we tested models with 2 and 3 components, enabling them to capture patterns with up to 2 or 3 distinct peaks. Users can extend this approach by opting for models with additional components to accommodate even more complex patterns. However, it is essential to balance this flexibility with considerations of overfitting and computational cost, as increasing the number of components may lead to excessive complexity and higher resource demands.

Among the three clustering categories, the DH category is the easiest one for us to provide some recommendations. Under the DH category, characterized by coarser clustering granularity, our results suggested that MMs, particularly GMMs, effectively balance computational efficiency and model performance. While WMMs delivered superior results in several cases, their higher computational demands make GMMs a more practical choice, especially if computational resources are a limitation for the user. Therefore, GMMs emerge as a reliable choice under DH, delivering strong results across all evaluation metrics while remaining computationally feasible. However, if computational resources are not a concern, WMMs can be a strong alternative.

While MMs continued to perform robustly under SDH and MDH categories, particularly in terms of passing the KS test, their dominance becomes less pronounced as clustering granularity increases. When additional metrics, such as log-likelihood and AIC, are considered, the performance landscape became more diverse, with several models showing competitive results in spe-

cific scenarios. The visual assessment provided in Section 6.5 illustrates well how several models achieved good results in specific situations. However, the overall weak performance of certain models (such as Gumbel, and 1- and 2-parameter Rayleigh), even under the MDH category, provides a clear indication of their limitations in capturing the data complexity. These models struggled to provide adequate fits across multiple cases, making them unsuitable for general recommendations. Despite its generally poor performance, the normal distribution remains a convenient option for quick testing – particularly under finer clustering granularities – due to its simplicity and low computational cost during training and evaluation.

Another important finding is that despite its popularity in other works related to building energy consumption modeling [4, 5, 7], the 2-parameter Weibull distribution may not be a reliable choice for general use, particularly considering that its computational cost for fitting and evaluation is relatively high. For instance, GMMs often provided better results while being significantly more computationally efficient. The findings in [4] suggested that the 2-parameter Weibull distribution outperforms the 2-parameter log-normal, but our results do not support this conclusion either. In our evaluation, both the 2-parameter log-normal and its 3-parameter counterpart performed better in many cases, particularly when considering both GoF and computational efficiency together. Furthermore, several other canonical distributions – such as the exponentially modified normal, 2- and 3-parameter inverse Gaussian, 2- and 3-parameter gamma, and the 3-parameter Weibull – frequently outperformed the 2-parameter Weibull distribution. It highlights the importance of re-evaluating established practices in light of comprehensive performance assessments across diverse datasets.

When selecting between a simpler and a more complex version of a model (e.g., norm vs. GMM2 vs. GMM3 or gamma2 vs. gamma), our overall LLR test results also indicate that the advantage of using more complex models increases as clustering granularity becomes coarser.

Thus, our overall recommendation for model selection begins with choosing the appropriate clustering granularity, which should be guided by the amount of available data as well as the specific requirements and constraints of the system. Once the granularity is determined, the selection of the probability model should follow accordingly. GMMs consistently emerge as a strong default choice, offering a favorable balance between performance and computational cost across all clustering levels. However, when finer clustering is feasible – enabled by larger datasets or higher temporal resolution – canonical distributions become a more viable and efficient alternative. Among these, the log-normal distribution stands out as a particularly effective option due to its good performance and low computational cost.

7.3. Data

Our analysis of the UK, REC1, and REC2 datasets revealed both shared patterns and notable differences in the effectiveness of probability distribution models for energy consumption modeling. One of the consistent trends across all three datasets was the positive impact of increased clustering granularity on the KS test success (despite it can be due the decrease on the number of instances available). However, despite this general trend, differences in model performance emerged when evaluating log-likelihood and AIC values. Specifically, for the UK and REC1 datasets, coarser clustering (DH) generally outperformed finer clustering (SDH and MDH), whereas the opposite trend was observed for REC2 datasets. These contrasting behaviors indicate that while increased granularity can enhance distribution fitting in some cases, the optimal level of clustering depends on the nature of the dataset.

We also identified internal patterns within each dataset group. For instance, in terms of fitting success, certain models – such as lognorm2, lognorm, invgauss2, gamma2, and Rayleigh1 –

demonstrated consistent behavior across REC1 datasets. Under the DH clustering category, these models only achieved fitting success in specific cases, such as weekday 1 (all hours) and weekday 7 (all hours except hour 22). Additionally, lognorm achieved fitting success for weekday 6 at hour 23 across all REC1 datasets. This pattern suggests that REC1 datasets share some common temporal characteristics, making these time periods more favorable for fitting. Interestingly, similar trends emerged under the SDH clustering category, reinforcing the coherence among certain temporal clusters within REC1. However, REC2 datasets displayed a clear division between energy profiles A and B and energy profiles C, D, and E. While these models struggled to fit the latter three datasets, they encountered fewer challenges with datasets A and B. A similar distinction was observed with the Weibull2 model, which had good fitting success on datasets A and B but faced difficulties with datasets C, D, and E. These findings suggest that within each dataset group, energy consumption patterns exhibit both shared characteristics and considerable variation.

While our results highlight these variations in energy patterns across datasets, they do not provide conclusive evidence that diversity is necessarily greater between dataset groups (UK vs. REC1 vs. REC2) than within them. Instead, our observations indicate that regardless of dataset origin, energy profiles contain clusters where model fitting is significantly easier than in others, independent of clustering granularity. This underscores the importance of carefully selecting clustering strategies and model evaluation criteria to account for the complexity of energy consumption patterns.

8. Conclusion

This study presents a structured and comprehensive extension of prior work on electricity consumption modeling in buildings. Our contributions are fivefold: (1) we benchmark 20 probability distributions – including both simpler and more complex forms of canonical distributions, additional right-skewed distributions, mixture models, and KDE as a non-parametric alternative; (2) we introduce three levels of temporal clustering granularity to accommodate varying data availability and support adaptable modeling strategies; (3) we conducted a diverse dataset analysis, incorporating both residential and business electricity consumption profiles from the UK and Belgian RECs; (4) we introduce a robust evaluation framework that integrates multiple statistical metrics – log-likelihood, AIC, likelihood-ratio tests, and a bootstrap-adjusted KS test—for rigorous model comparison; and (5) we provide practical guidance for model selection based on clustering granularity, model complexity, and computational cost. These contributions address key methodological and practical gaps in the literature.

Our findings emphasize the critical role of clustering granularity in determining model performance for electricity consumption modeling. Mixture models – particularly Gaussian Mixture Models (GMMs) – demonstrated consistently strong results across all datasets and clustering levels. In coarser clusters (DH), GMMs provided an optimal trade-off between accuracy and computational cost. At the same time, Weibull Mixture Models (WMMs), despite occasionally achieving better fits, were a less practical choice due to their higher computational demands.

As clustering granularity increases (SDH and MDH), the dominance of mixture models diminishes. Model performance becomes more context-dependent, with several canonical distributions performing adequately in specific scenarios. However, models such as Gumbel and Rayleigh often underperformed and are not recommended for general use. Notably, the widely used 2-parameter Weibull may not be a reliable choice given its performance in our experiments and its relative complexity.

We recommend selecting the appropriate clustering granularity based on data availability and system constraints, and then choosing the probability distribution accordingly. GMMs serve as a robust general-purpose option across all levels, but when finer clustering is possible, canonical distributions – especially the log-normal – offer an effective and efficient alternative.

Future work could explore dynamic clustering strategies that adapt resolution based on the characteristics of the dataset. Another promising direction is the development of hybrid or composite models – such as mixtures with additional components or combinations of different distribution types – to improve fitting accuracy in challenging clusters. Expanding the dataset scope to include more diverse consumption profiles, particularly from non-residential sectors like schools, hospitals, and industrial facilities, would also enhance the generalizability of the findings.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Grammarly and ChatGPT-4o in order to improve the writing and make the text clearer and more concise. After using Grammarly and ChatGPT-4o, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Acknowledgments

We would like to thank the financial supports from the Service Public de Wallonie Recherche under Grant No. 2110107 - SERENITY2 by WIN2WAL. We gratefully acknowledge the Brussels-Capital Region - Innoviris (Brussels Public Organisation for Research and Innovation) for financial support under grant number 2021-RDIR-49b. This work was also partially supported by the AIDE project funded by the Belgian SPF BOSA under the program "Financing of projects for the development of artificial intelligence in Belgium" with reference number 06.40.32.33.00.10. We would also like to thank WeSmart for providing us with data sets. Computational resources have been provided by the *Consortium des Équipements de Calcul Intensif* (CÉCI), funded by the *Fonds de la Recherche Scientifique de Belgique* (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region. Finally, we would like to extend our sincere gratitude to Professor Kim Mens for his time and advice in reviewing this paper.

Appendix A. Extra results

Table A.13: Percentage of KS tests with p-values exceeding the significance level for the **UK** datasets, based only on successful model fittings.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	**0**	0.0	1.9	2.4	1.8	1.9	2.4	4.2	3.8	4.8	0.0	0.0	0.0	1.8	2.4	13.1	32.7	48.8	57.7	10.1	9.5
	1	0.0	3.6	3.6	6.0	2.4	2.4	1.8	1.8	2.4	1.2	0.0	0.0	0.0	0.6	11.3	40.5	53.6	61.3	22.6	18.5
	12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.6	16.1	61.3	67.9	0.0	0.0
	13	0.0	1.2	1.2	0.6	1.2	1.2	1.2	1.2	1.2	0.0	0.6	0.6	1.2	1.2	8.3	37.5	66.7	77.4	32.7	28.0
	14	1.8	6.6	6.5	8.3	7.8	8.3	1.8	1.8	1.8	0.0	0.0	0.0	0.6	1.2	8.3	33.3	58.3	76.2	13.7	12.5
	15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.0	26.8	56.5	64.3	13.1	12.5
	16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	34.5	4.2	11.9	0.0	0.0
	17	0.0	0.0	0.0	1.2	0.0	0.0	1.2	0.6	1.2	0.0	0.0	0.0	1.2	1.2	0.0	9.5	38.1	51.8	0.6	0.6
	18	0.0	0.6	0.6	0.0	2.4	3.6	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	13.1	53.0	60.7	0.0	0.0
	20	0.0	3.0	3.0	4.8	3.0	3.0	2.4	3.0	4.2	2.4	0.0	0.0	0.6	0.6	7.7	20.8	58.9	69.6	8.3	8.3
	Mean	0.2	1.7	1.7	2.3	1.9	2.1	1.3	1.2	1.5	0.4	0.1	0.1	0.5	0.7	7.4	26.5	49.9	59.9	10.1	9.0
	Std	0.6	2.2	2.2	3.0	2.4	2.6	1.3	1.4	1.8	0.8	0.2	0.2	0.7	0.8	7.0	10.9	17.8	18.6	11.0	9.4
	Rank	12.1	8.1	8.7	8.1	8.1	8.1	8.6	8.4	7.8	12	12.3	12.4	10.4	9.7	6.1	2.8	2.1	1.1	5.4	5.9
SDH	**0**	7.0	19.7	20.5	25.4	19.1	19.0	25.7	17.3	25.3	8.8	4.4	4.5	14.4	20.5	48.7	70.7	49.0	58.9	23.7	22.0
	1	1.9	14.9	15.2	17.4	13.6	14.4	14.0	10.6	12.6	9.2	5.2	5.2	9.1	11.5	32.4	60.6	55.4	65.3	27.2	25.1
	12	0.1	5.8	6.8	4.5	4.5	6.0	3.7	1.6	4.2	1.0	0.7	0.7	1.8	3.1	19.2	44.3	36.2	57.1	2.5	2.2
	13	5.2	11.5	11.8	15.0	10.0	10.0	18.3	8.4	10.6	5.1	3.7	3.9	10.3	11.5	45.1	67.4	60.1	76.3	43.6	42.6
	14	7.1	28.8	30.1	26.9	27.0	28.4	22.9	19.3	23.1	10.1	4.3	4.5	14.4	19.8	40.2	67.6	52.5	75.9	28.7	27.1
	15	4.5	13.1	13.8	18.2	11.6	11.8	12.6	9.0	11.6	7.1	3.0	3.0	7.1	8.2	29.2	59.2	44.3	58.2	25.3	23.8
	16	0.1	2.1	2.1	2.7	2.2	2.4	4.6	0.9	2.2	0.9	0.4	0.4	1.6	2.4	10.9	39.6	23.1	35.7	5.4	4.9
	17	2.2	6.9	7.4	9.5	7.0	7.3	14.4	7.5	11.2	1.0	3.0	3.0	6.5	9.7	12.8	43.5	39.4	55.1	5.7	5.2
	18	0.0	19.0	19.5	9.1	20.1	22.0	17.3	11.0	15.8	0.3	0.0	0.0	8.5	13.4	9.1	33.2	50.6	59.2	0.9	0.7
	20	5.8	21.8	23.2	21.1	19.4	20.1	19.0	15.7	16.8	12.6	4.8	4.8	12.2	14.3	38.2	58.6	49.3	67.4	24.1	23.4
	Mean	3.4	14.4	15.0	15.0	13.5	14.1	15.3	10.1	13.3	5.6	3.0	3.0	8.6	11.4	28.6	54.5	46.0	60.9	18.7	17.7
	Std	2.9	8.2	8.5	8.4	7.9	8.2	7.1	6.1	7.3	4.6	1.9	1.9	4.5	6.0	14.7	13.2	10.8	11.6	14.2	13.7
	Rank	18.6	9.4	8.3	8.5	10.8	9.8	8.8	14.1	10.7	17.1	18.7	19.0	15.3	11.7	4.9	1.8	2.9	1.3	8.3	9.4
MDH	**0**	28.8	48.0	51.6	53.6	47.4	49.7	44.8	48.2	55.6	34.4	27.8	27.9	42.9	59.0	72.3	84.0	48.9	60.4	41.9	40.7
	1	16.0	37.7	38.9	40.7	36.6	37.3	45.2	36.6	47.0	26.3	20.1	20.0	33.0	42.3	60.7	80.7	58.3	67.4	39.1	36.9
	12	3.3	18.2	21.1	19.8	17.4	21.1	21.1	12.9	31.7	6.8	5.0	5.0	10.4	27.7	44.0	64.3	26.2	44.9	10.1	9.4
	13	20.9	32.2	34.2	39.1	30.0	31.3	48.0	31.5	42.6	20.0	20.2	20.1	37.3	45.3	68.7	84.2	54.0	75.8	52.7	50.4
	14	19.3	51.2	53.4	50.4	48.8	51.7	48.7	42.7	52.9	29.4	20.3	20.3	38.9	52.8	70.3	84.9	49.9	72.2	41.6	39.8
	15	15.8	36.5	38.9	43.1	36.0	37.1	36.0	31.2	43.8	23.9	17.9	17.7	29.7	41.9	60.3	80.8	41.6	55.4	39.4	37.1
	16	6.3	14.5	15.9	19.8	13.8	15.4	21.7	12.3	24.8	8.3	6.8	6.9	11.9	22.2	35.0	65.5	24.6	44.7	16.3	15.5
	17	12.6	29.0	31.1	37.0	27.5	29.3	39.4	24.6	44.0	18.0	15.5	15.5	25.7	39.1	50.5	71.4	41.5	62.6	23.7	22.4
	18	3.2	42.1	43.8	29.4	40.2	41.9	42.4	48.7	47.1	7.9	4.4	4.4	36.4	47.7	49.2	66.7	50.8	61.2	7.8	7.5
	20	21.4	44.8	46.8	46.7	44.4	46.0	43.3	37.1	48.1	31.0	24.1	24.2	35.1	47.4	68.4	83.1	45.7	66.7	40.4	39.2
	Mean	14.8	35.4	37.6	37.9	34.2	36.1	39.1	32.6	43.8	20.6	16.2	16.2	30.1	42.5	57.9	76.6	44.2	61.1	31.3	29.9
	Std	8.4	12.1	12.3	11.8	12.0	12.0	10.0	13.0	9.3	10.2	8.2	8.2	11.1	10.9	12.7	8.5	11.1	10.4	15.5	14.9
	Rank	19.5	11.0	8.2	8.5	12.8	10.1	9.4	12.8	5.1	17.1	18.6	18.6	14.5	6.0	2.8	1.0	6.2	2.3	11.9	13.5

Table A.14: Percentage of KS tests with p-values exceeding the significance level for the **REC1** datasets, based only on successful model fittings.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	**0**	0.0	2.1	4.2	0.0	4.3	1.2	1.2	2.1	1.2	0.6	0.0	0.0	0.0	0.6	22.6	49.4	8.3	19.0	3.0	2.4
	10	3.0	12.8	12.5	7.1	6.4	4.2	4.8	4.3	3.6	3.0	0.0	0.6	7.2	8.9	30.4	55.4	19.0	25.0	26.2	25.0
	12	0.0	19.1	18.8	2.4	10.6	4.8	4.2	6.4	3.0	1.8	2.1	1.2	4.0	4.2	25.0	42.3	19.0	21.4	2.4	2.4
	13	0.0	2.1	6.3	0.0	4.3	1.8	0.6	0.0	0.6	0.0	0.0	0.0	0.0	0.6	5.4	23.2	16.1	16.7	0.0	0.0
	18	0.0	2.1	2.1	0.6	2.1	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	28.0	38.1	20.2	22.0	0.6	0.6
	20	0.6	0.0	0.0	0.0	0.0	0.0	1.8	0.0	0.6	0.0	4.3	1.2	0.0	0.0	28.6	47.6	15.5	25.0	13.7	12.5
	21	8.3	17.0	16.7	22.6	19.1	16.1	22.6	21.3	20.8	14.9	19.1	10.7	13.1	21.4	47.6	50.0	19.6	21.4	29.2	27.4
	22	0.0	6.4	6.3	7.1	8.5	6.0	2.4	2.1	1.8	2.4	0.0	0.6	3.5	1.8	18.5	54.2	16.1	22.6	4.2	3.0
	8	0.0	14.9	14.6	1.8	14.9	4.8	0.0	0.0	0.0	0.6	0.0	0.0	0.8	3.0	14.3	36.9	22.6	22.6	0.6	0.6
	9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.3	31.5	1.8	1.2	7.1	6.0
	Mean	1.2	7.7	8.1	4.2	7.0	3.9	3.8	3.6	3.2	2.3	2.6	1.4	2.9	4.0	23.2	42.9	15.8	19.7	8.7	8.0
	Std	2.7	7.5	7.0	7.1	6.3	4.8	6.8	6.6	6.3	4.5	6.0	3.3	4.4	6.7	11.9	10.4	6.3	7.0	10.8	10.3
	Rank	14.8	7.6	7.9	10.1	7.5	10.3	10.3	11.3	11.9	13.4	13.7	14.6	12.1	10.6	2.9	1.0	4.8	3.5	7.7	8.4
SDH	**0**	8.5	21.1	24.3	20.7	23.0	22.2	20.0	16.5	22.9	10.1	9.1	7.3	11.4	25.1	40.5	54.8	26.3	43.9	10.9	10.0
	10	13.7	45.2	49.3	39.4	45.7	41.5	38.5	44.0	40.6	18.3	16.1	14.1	32.9	49.1	54.5	74.6	42.0	54.3	30.2	28.1
	12	10.1	43.1	47.1	32.6	40.2	39.0	39.4	51.8	39.1	23.1	18.6	16.8	35.1	45.1	56.8	75.0	41.7	56.0	22.9	21.9
	13	4.5	16.3	18.0	19.5	19.0	17.9	19.8	28.3	20.8	4.8	4.7	3.6	19.6	24.6	38.1	54.6	38.5	49.7	10.7	10.1
	18	8.9	16.7	19.9	15.9	18.2	18.0	19.5	21.4	15.2	6.8	5.3	4.3	14.4	19.3	38.8	49.3	41.8	58.2	12.6	12.4
	20	18.9	37.2	37.2	30.8	25.8	27.7	28.1	30.7	33.0	18.8	21.6	17.3	30.1	35.6	64.7	83.8	24.1	28.9	34.8	34.5
	21	29.6	45.4	47.1	50.7	43.3	41.5	40.7	40.0	49.4	34.1	33.2	30.2	33.0	52.4	62.4	75.3	45.2	51.3	45.4	44.0
	22	8.8	34.3	36.4	25.4	32.6	32.4	26.5	38.1	30.4	15.5	9.7	9.8	25.9	29.2	55.7	61.9	41.7	54.0	18.3	16.5
	8	8.3	36.3	41.2	28.4	40.8	41.2	27.5	44.2	29.2	14.3	9.8	11.3	26.4	35.9	53.7	68.3	43.2	58.2	17.3	16.2
	9	16.4	38.2	38.4	35.6	28.0	28.4	35.5	30.7	37.6	20.5	24.8	19.2	22.9	36.6	50.7	79.5	13.4	13.4	36.5	35.0
	Mean	12.8	33.4	35.9	29.9	31.7	31.0	29.5	34.6	31.8	16.6	15.3	13.4	25.2	35.3	51.6	67.7	35.8	46.8	24.0	22.9
	Std	7.3	11.3	11.4	10.4	10.3	9.6	8.4	11.0	10.3	8.5	9.3	8.0	8.1	11.0	9.5	11.9	10.7	14.6	12.1	11.9
	Rank	19.1	8.1	5.9	10.7	9.8	10.8	11.1	8.0	9.0	16.8	17.7	19.0	13.5	6.5	2.6	1.1	8.1	5.1	12.7	14.2

Table A.15: Percentage of KS tests with p-values exceeding the significance level for the **REC2** datasets, based only on successful model fittings.

Cat.	D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
DH	A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	5.4	48.2	82.7	0.0	0.0
	B	0.0	3.6	3.6	3.6	3.0	3.0	1.2	2.4	3.0	6.0	0.0	0.0	0.0	0.0	17.9	23.8	55.4	56.0	20.8	19.0
	C	0.0		0.0	0.0		0.0	0.0		0.0	0.0		0.0	0.0	0.0	19.2	50.3	0.0	0.0	0.0	0.0
	D	0.0		0.0	0.0		0.0	0.0		0.0	0.0		0.0	0.0	0.0	0.0	4.2	0.0	0.0	0.0	0.0
	E	0.0		0.0	0.0		0.0	0.0		0.0	0.0		0.0	0.0	0.0	1.2	4.8	1.8	0.6	0.0	0.0
Mean		0.0	1.8	0.7	0.7	1.5	0.6	0.2	1.2	0.7	1.2	0.0	0.0	0.0	0.0	7.6	17.7	21.1	27.9	4.2	3.8
Std		0.0	2.5	1.6	1.6	2.1	1.3	0.5	1.7	1.3	2.7	0.0	0.0	0.0	0.0	9.9	20.0	28.2	39.0	9.3	8.5
Rank		8.6	3.2	7.2	7.2	3.8	7.8	8.4	4.4	7.6	6.8	4.8	8.6	8.6	8.6	6.0	4.2	4.6	4.6	6.2	6.4
SDH	A	1.9	2.1	2.2	2.5	0.9	1.0	2.1	2.1	2.1	0.0	0.1	0.1	1.8	1.2	13.5	23.8	43.6	42.7	5.2	4.9
	B	14.1	19.1	19.8	24.1	17.5	17.4	24.1	16.7	19.8	11.0	9.4	9.5	18.2	19.3	42.1	60.9	46.3	53.9	36.6	35.7
	C	0.0	6.7	2.6	1.3	13.3	0.9	0.9	13.3	0.9	0.3	0.0	0.1	9.7	0.7	11.8	25.8	1.3	0.9	0.1	0.1
	D	0.0	28.8	20.2	8.9	34.6	5.8	2.8	28.8	2.2	3.1	3.8	1.6	26.5	3.3	14.7	19.2	4.2	4.3	0.6	0.6
	E	0.0	0.0	0.0	1.0	0.8	0.3	1.8	5.9	0.9	0.0	0.8	0.1	2.6	0.9	16.1	22.7	8.7	9.6	0.3	0.3
Mean		3.2	11.3	9.0	7.6	13.4	5.1	6.3	13.4	5.2	2.9	2.8	2.3	11.7	5.1	19.7	30.5	20.8	22.3	8.6	8.3
Std		6.2	12.3	10.1	9.8	14.0	7.2	10.0	10.4	8.2	4.7	4.0	4.1	10.5	8.0	12.7	17.2	22.2	24.3	15.8	15.4
Rank		17.2	9.2	9.2	7.6	9.0	12.6	10.0	6.8	11.4	16.8	16.0	17.4	8.4	12.4	4.2	2.4	5.4	5.4	11.4	11.8
MDH	A	10.6	12.3	13.2	15.8	12.3	12.8	20.9	14.7	17.3	7.5	7.4	7.4	14.0	17.3	28.9	37.6	39.1	42.1	15.2	15.0
	B	33.5	41.7	43.3	48.1	40.3	40.9	48.6	40.6	49.8	31.5	29.2	29.3	40.3	47.6	66.9	83.2	50.4	52.3	53.2	52.1
	C	0.9	9.7	6.5	5.2	16.6	3.9	2.9	37.7	3.6	2.8	5.7	2.1	20.1	2.6	16.4	26.8	5.3	5.3	1.9	1.8
	D	3.7	49.7	42.0	15.5	59.4	13.3	10.6	69.5	10.1	9.4	23.7	6.3	58.4	11.7	30.2	36.8	14.5	14.8	5.1	4.9
	E	5.0	13.3	9.5	10.5	17.3	6.6	12.7	31.9	10.0	5.3	10.1	5.7	16.8	8.5	28.8	43.8	28.6	32.1	6.7	6.6
Mean		10.7	25.3	22.9	19.0	29.2	15.5	19.1	38.9	18.2	11.3	15.3	10.2	29.9	17.5	34.2	45.7	27.6	29.3	16.4	16.1
Std		13.2	18.9	18.2	16.8	20.1	14.8	17.7	19.8	18.3	11.6	10.5	10.9	19.0	17.6	19.1	21.9	18.2	19.3	21.1	20.7
Rank		18.8	9.0	9.8	9.4	8.6	13.4	10.0	6.0	10.6	17.2	13.2	18.2	8.0	11.8	4.4	2.6	6.6	5.2	12.6	14.0

Table A.16: Runtime in hours for experiments in the Dragon 1 machine – UK datasets under the MDH category.

D.	norm	lognorm2	lognorm	exponnorm	invgauss2	invgauss	beta	gamma2	gamma	Gumbel	Rayleigh1	Rayleigh	Weibull2	Weibull	GMM2	GMM3	WMM2	WMM3	KDE scott	KDE silv.
1	0.06	12.61	53.75	181.05	158.17	200.81	454.70	191.15	272.24	5.08	3.11	3.07	163.66	223.70	35.91	45.88	448.54	960.55	26.30	26.49
12	0.07	12.55	18.38	298.07	207.23	119.25	553.15	198.92	309.16	5.12	3.05	3.13	139.58	313.55	36.83	46.06	447.09	962.41	29.71	29.87
16	0.07	12.46	32.37	168.50	196.95	231.59	474.56	195.76	316.73	5.28	3.12	3.15	151.73	241.80	34.99	62.67	449.05	919.94	28.44	28.46
17	0.07	12.48	35.64	234.40	159.86	164.14	505.97	169.63	283.18	5.11	3.10	3.12	133.57	231.60	38.82	47.30	442.77	918.66	27.02	27.83
18	0.07	13.17	43.76	369.95	116.64	119.13	547.99	240.85	329.43	5.13	2.96	2.93	254.20	330.83	37.98	45.25	400.87	891.11	27.82	30.47

References

[1] K. Li, Z. Ma, D. Robinson, J. Ma, Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering, Applied Energy 231 (C) (2018) 331–342. doi:10.1016/j.apenergy.2018.0.
URL https://ideas.repec.org/a/eee/appene/v231y2018icp331-342.html

[2] W. Charytoniuk, M. Chen, P. Kotas, P. Van Olinda, Demand forecasting in power distribution systems using nonparametric probability density estimation, IEEE Transactions on Power Systems 14 (4) (1999) 1200–1206. doi:10.1109/59.801873.

[3] S. Arora, J. W. Taylor, Forecasting electricity smart meter data using conditional kernel density estimation, Omega 59 (2016) 47–59, business Analytics. doi:https://doi.org/10.1016/j.omega.2014.08.008.
URL https://www.sciencedirect.com/science/article/pii/S0305048314001546

[4] J. Munkhammar, J. Rydén, J. Widén, Characterizing probability density distributions for household electricity load profiles from high-resolution electricity use data, Applied Energy 135 (2014) 382–390. doi:https://doi.org/10.1016/j.apenergy.2014.08.093.
URL https://www.sciencedirect.com/science/article/pii/S0306261914009167

[5] J. Munkhammar, J. Widén, J. Rydén, On a probability distribution model combining household power consumption, electric vehicle home-charging and photovoltaic power production, Applied Energy 142 (2015) 135–143. doi:https://doi.org/10.1016/j.apenergy.2014.12.031.
URL https://www.sciencedirect.com/science/article/pii/S0306261914012884

[6] U. P. N. led Low Carbon London project, Smartmeter energy consumption data in london households (2014).
URL https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[7] Michalková, Mária, Pobočíková, Ivana, Sedliačková, Zuzana, Modelling the electricity consumption in a manufacturing company through probability distribution, MATEC Web Conf. 357 (2022) 08004. doi:10.1051/matecconf/202235708004.
URL https://doi.org/10.1051/matecconf/202235708004

[8] S. Eliason, Maximum Likelihood Estimation: Logic and Practice, Quantitative Applications in the Social Sciences, SAGE Publications, 1993.
URL https://books.google.be/books?id=c1SsBwAAQBAJ

[9] G. J. McLachlan, D. Peel, Finite mixture models, in: Wiley Series in Probability and Statistics, 2000.
URL https://api.semanticscholar.org/CorpusID:124985575

[10] M. Aitkin, G. T. Wilson, Mixture models, outliers, and the em algorithm, Technometrics 22 (3) (1980) 325–331. arXiv:https://www.tandfonline.com/doi/pdf/10.1080/00401706.1980.10486163, doi:10.1080/00401706.1980.10486163.
URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1980.10486163

[11] R. D'Agostino, Goodness-of-Fit-Techniques, Mathematics & Statistics, Routledge, 1986. URL https://doi.org/10.1201/9780203753064

[12] A. Maydeu-Olivares, C. Forero, Goodness-of-Fit Testing, Vol. 7, Elsevier Science, 2010, pp. 190–196. doi:10.1016/B978-0-08-044894-7.01333-6.

[13] X. Kang, J. An, D. Yan, A systematic review of building electricity use profile models, Energy and Buildings 281 (2022) 112753. doi:10.1016/j.enbuild.2022.112753.

[14] D. B. Crawley, L. K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher, M. J. Witte, et al., Energyplus: creating a new-generation building energy simulation program, Energy and buildings 33 (4) (2001) 319–331.

[15] J. P. Zimmermann, End-use metering campaign in 400 households in sweden assessment of the potential electricity savings, Contract 17 (September) (2009) 5–2743.

[16] A. Zeimbekakis, E. D. Schifano, J. Yan, On misuses of the kolmogorov–smirnov test for one-sample goodness-of-fit, The American Statistician (just-accepted) (2024) 1–18.

[17] J. Xu, X. Kang, Z. Chen, D. Yan, S. Guo, Y. Jin, T. Hao, R. Jia, Clustering-based probability distribution model for monthly residential building electricity consumption analysis, in: Building Simulation, Vol. 14, Springer, 2021, pp. 149–164.

[18] E. Carpaneto, G. Chicco, Probability distributions of the aggregated residential load, in: 2006 International Conference on Probabilistic Methods Applied to Power Systems, 2006, pp. 1–6. doi:10.1109/PMAPS.2006.360235.

[19] A. Cagni, E. Carpaneto, G. Chicco, R. Napoli, Characterisation of the aggregated load patterns for extraurban residential customer groups, in: Proceedings of the 12th IEEE Mediterranean Electrotechnical Conference (IEEE Cat. No. 04CH37521), Vol. 3, IEEE, 2004, pp. 951–954.

[20] F. Gonzalez-Longatt, J. Rueda, I. Erlich, W. Villa, D. Bogdanov, Mean variance mapping optimization for the identification of gaussian mixture model: Test case, in: 2012 6th IEEE International Conference Intelligent Systems, 2012, pp. 158–163. doi:10.1109/IS.2012.6335130.

[21] I. Erlich, G. K. Venayagamoorthy, N. Worawat, A mean-variance optimization algorithm, in: IEEE Congress on Evolutionary Computation, IEEE, 2010, pp. 1–6.

[22] A. Jordan, Y. Chen, Parametric probability density function characterization of single household peak loads, in: Proceedings of eSim 2024: 13th Conference of IBPSA-Canada, Vol. 13 of eSim, IBPSA-Canada, Alberta, Canada, 2022.
URL https://publications.ibpsa.org/conference/paper/?id=esim2024_159

[23] L. H. Bandória, B. Cortes, M. C. de Almeida, Statistical characterization of electricity use profile: Leveraging data analytics for stochastic simulation in a smart campus, Energy and Buildings 324 (2024) 114934. doi:https://doi.org/10.1016/j.enbuild.2024.114934.
URL https://www.sciencedirect.com/science/article/pii/S0378778824010508

[24] P. Stoica, Y. Selen, Model-order selection: a review of information criterion rules, IEEE Signal Processing Magazine 21 (4) (2004) 36–47.

[25] W. Stute, W. G. Manteiga, M. P. Quindimil, Bootstrap based goodness-of-fit-tests, Metrika 40 (1993) 243–256.

Declaration of interests

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

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: