Wiley
International Journal of Energy Research
Volume 2025, Article ID 4460462, 24 pages
https://doi.org/10.1155/er/4460462

WILEY

Research Article

A Robust Approach to Extend Deterministic Models for the Quantification of Uncertainty and Comprehensive Evaluation of the Probabilistic Forecasting

Ajay Upadhaya[1], **Jan-Simon Telle**[1], **Sunke Schlüters**[1], **Mohammad Saber**[2] and **Karsten von Maydell**[1]

[1]*Energy Systems Technologies, DLR Institute of Networked Energy Systems, Oldenburg, Lower Saxony, Germany*
[2]*Department of Anesthesiology and Biomedical Engineering, Medical College of Wisconsin, Milwaukee, Wisconsin, USA*

Correspondence should be addressed to Ajay Upadhaya; ajay.upadhaya@dlr.de

Received 10 April 2024; Revised 12 December 2024; Accepted 9 January 2025

Academic Editor: Temitope Adefarati

Forecasting generation and demand forms the foundation of power system planning, operation, and a multitude of decision-making processes. However, traditional deterministic forecasts lack crucial information about uncertainty. With the increasing decentralization of power systems, understanding, and quantifying uncertainty are vital for maintaining resilience. This paper introduces the uncertainty binning method (UBM), a novel approach that extends deterministic models to provide comprehensive probabilistic forecasting and thereby support informed decision-making in energy management. The UBM offers advantages such as simplicity, low data requirements, minimal feature engineering, computational efficiency, adaptability, and ease of implementation. It addresses the demand for reliable and cost-effective energy management system (EMS) solutions in distributed integrated local energy systems, particularly in commercial facilities. To validate its practical applicability, a case study was conducted on an integrated energy system at a logistics facility in northern Germany, focusing on the probabilistic forecasting of electricity demand, heat demand, and PV generation. The results demonstrate the UBM's high reliability across sectors. However, low sharpness was observed in probabilistic PV generation forecasts, attributed to the low accuracy obtained by the deterministic model. Notably, the accuracy of the deterministic model significantly influences the accuracy of the UBM. Additionally, this paper addresses various challenges in popular evaluation scores for probabilistic forecasting with implementing new ones, namely a graphical calibration score, quantile calibration score (QCS), and percentage quantile calibration score (PQCS). The findings presented in this work contribute significantly to enhancing decision-making capabilities within distributed integrated local energy systems.

Keywords: decision-making; distributed integrated local energy system; energy management system (EMS); forecasting evaluation; probabilistic forecast; sector-coupling; uncertainty quantification

1. Introduction

Power systems are undergoing a profound transformation in the form of decentralization, along with the penetration of renewable energy sources, battery electric vehicles (BEVs), heat pumps, hydrogen, etc. This results in increasing uncertainties due to intermittent generation, as well as dynamic and less predictable demand [1–3]. These present challenges to power system planning and operation practices, such as in terms of energy management, economic dispatch, unit commitment, maintenance planning, etc. [4]. Traditional deterministic forecast techniques do not capture uncertainties, which leads to decision-makers being poorly advised [5]. A modern energy system, therefore, requires appropriate quantification of uncertainty to enable informed decision-making. In contrast to deterministic forecasts, the probabilistic forecasting method provides information about uncertainty and should, therefore, be investigated to facilitate the energy

transition [6–8]. Following the "Global Energy Forecasting Competitions" in 2012 and 2014 [9, 10], there has been a significant surge in research interest in probabilistic forecasting within the energy domain [10]. With the decentralization of power systems, a growing need has emerged in recent years for quantifying uncertainties in energy management systems (EMSs) within smaller facilities, including residential buildings and small- to medium-sized commercial buildings. These facilities require cost-effective EMS solutions that integrate probabilistic prediction techniques while being simple, convenient, and implementable with minimal data requirements.

Despite the potential benefits of probabilistic forecasting methods in offering uncertainty information, they are complex, data-intensive, and computationally demanding [3]. Additionally, a notable fraction of these methods rely on black-box models, which lack transparency, as highlighted in prior research [11]. Transparency is a crucial attribute in forecasting methodologies [1]. Furthermore, research still leans heavily toward deterministic forecasting [8, 12, 13]. While many probabilistic forecasting models directly generate forecast distributions without tying back to deterministic models, the literature concerning the combination of these approaches remains sparse. Typically, most literature employs the quantile regression averaging (QRA) method, which follows the approach of leveraging deterministic methods. This concept has been applied across various scenarios, including to electricity prices [14–16], electricity load [7, 17], solar PV generation [5], and wind power generation [3, 18]. However, the demand for multiple-point forecast models introduces challenges such as high data requirements, model complexities, and increased computational time. Furthermore, quantile regression (QR) itself demands extensive computational resources, necessitating separate model training for each quantile [19]. If the post-processing method adds significant complexity and effort on top of the development of the deterministic models, its practical value may be undermined.

Wang et al. [7] introduced a probabilistic forecast method that leverages existing point forecast methods by modeling the conditional forecast residual using QR to derive the probabilistic forecasts. Although this approach reduces reliance on multiple-point forecast models, it still necessitates complex post-processing modeling. Another study by Zhang et al. [20] employed copula theory to model the conditional forecast error for stochastic unit commitment in multiple wind farms. Dang et al. [21] utilized point forecasts from three deep neural network models and a similar-day load selection algorithm to facilitate short-term probabilistic load forecasting via quantile random forests (QRFs). Subsequently, QRF found diverse applications across various domains [19, 22–25]. Zhang, Quan, and Srinivasan [3] found QRF to be both more accurate and computationally efficient compared to QRA.

The empirical prediction intervals (EPIs) method, first introduced by William and Goodman [26], produces probabilistic forecasts around existing point forecasts based on the distribution of past point forecast errors within a time window [27]. This method has been implemented in various fields, including meteorology [28], economics [29], and energy [30]. However, the major limitation to EPIs is that the PIs are not conditional [27]. This leads to wider interval width due to unconditional uncertainty, making them less adaptive. Hence, it limits its wider application for decision-making in certain sectors and use cases. To quantify uncertainty in a wind power forecast for a wind farm in China, Huang et al. [31] presented a simplistic statistical approach that transforms point forecasts into interval forecasts by considering the conditional dependence between predicted values and prediction errors. Saber [32] then proposed three methods to transform point forecasts into probabilistic ones with relative ease of implementation. Compared to Huang et al. [31], Saber's methods employed historical weather data as conditionals. Saber's approach finds application in the quantification of uncertainties in U.S. electricity and natural gas consumption. Nevertheless, this approach has its limitations, as it is not readily applicable to all cases due to the possible unavailability of weather data, and the weather parameters as conditionals do not always have a high correlation with the forecasted outputs. Additionally, the approach requires several years of point forecasts and weather data for training, which may hinder its broader applicability. Furthermore, the approach has not been adequately tested in scenarios involving distribution level, where uncertainties are higher. The analog ensemble (AnEn) method was used to generate probabilistic wind power forecasts [33] and solar power forecasts [34]. The AnEn generates probabilistic forecasts using a set of past measured values corresponding to the most similar past deterministic forecasts of the predictors at the same lead time to a current point forecast (predictors). It computes the deviation from the current forecast and every similar past forecast at the same lead time for the predictor variables and selects the n number of forecasts with the lowest error values at each lead time; the corresponding past measured values are the ensembles of the AnEn forecast which constitutes probabilistic forecasts. One of the disadvantages of this technique is that it requires meteorological predictor variables and its forecasted values in the training set [34]. Also, the AnEn method is less adaptive and not extendable to sector-wise applications, as past observations at the same lead time will not always correlate to the current lead time, which may lead to unexpected interval width and inaccuracy of the model's output. The effectiveness of AnEn prediction is highly sensitive to the criteria used to define the similarity between the current situation and historical analogs [34]. Even slight alterations in these criteria can result in notable disparities in forecasted outcomes. Also, the conditional approach to generate probabilistic forecasting is still limited in this technique. Moreover, at various time stamps, the numbers of ensemble observations could be limited to get a proper distribution forecast for the current lead time.

Beyond addressing gaps in the probabilistic forecasting techniques, the lack of effective evaluation methods is a contributing factor to their limited adoption in forecasting applications [8]. Popular scores possess certain limitations: they

are data-dependent and prioritize sharpness over reliability. The pinball score and Winkler score (WS), for instance, yield higher scores for forecasts that are sharper and moderately reliable as opposed to less sharp but highly reliable forecasts [32, 35, 36]. Furthermore, mean PI width (MPIW) is data-dependent, making it unsuitable for comparing probabilistic forecasting across different datasets. Additionally, PI coverage probability (PICP) only measures the reliability of probabilistic forecasts by considering the PI without accounting for the reliability of each quantile bin. Reliability deviated from the expectation is not yet penalized by the currently available scoring rules. Moreover, graphical evaluation tools, although present in some of the literature in the form of probability integral transform (PIT) [37, 38], lack numerical scores, preventing the direct comparison of similar-looking PIT distributions [32].

There are significant benefits to combining the deterministic and probabilistic methods for uncertainty quantification [5]. Within this context, this paper emphasizes the exploration of the potential for leveraging deterministic forecast models to derive probabilistic forecasts. This research avenue will be the central focus of this paper, which will delve into the intricacies of this approach and its implications. Based on a comprehensive review of available probabilistic prediction techniques that leverage deterministic models, this paper aims to address existing knowledge gaps through a simplified and computationally efficient post-processing technique for uncertainty quantification. Inaccurate point forecasts lead to higher power system operating costs and inefficient use of renewable energy sources [5]. Therefore, the transition from point forecasts to probabilistic ones becomes vital for quantifying inherent uncertainties. Moreover, fostering the adoption of probabilistic forecasting in the energy domain necessitates more comprehensive scoring metrics to address existing gaps in the popular evaluation metrics.

This paper makes the following contributions:

1. It develops a simplified statistical probabilistic forecasting framework, named uncertainty binning method (UBM), that leverages deterministic models. The UBM framework is designed to be cost-effective, data-efficient, computationally fast, and easy to implement. The UBM can serve as an extended tool to convert point forecasts into probabilistic ones, thereby facilitating uncertainty quantification and aiding decision-making processes in power system planning and operation.
2. This paper implemented new evaluation scores, namely the graphical calibration measure (GCM), quantile calibration score (QCS), and percentage quantile calibration score (PQCS). They provide a holistic approach to probabilistic forecasting evaluation, effectively addressing deficiencies in popular scoring techniques.
3. The proposed method showcases robust performance across multiple sectors, including electricity, heat, and PV. A practical case study validates its application in an existing distributed integrated energy system of a logistics facility in northern Germany.

The remainder of this paper is organized as follows: In Section 2, the methodology of the forecasting framework is described in detail. In Section 2.2, performance evaluations of deterministic and probabilistic forecasting are discussed. In Section 3, a case study of the distributed integrated local energy system of a commercial logistics facility is presented. Section 4 presents the results, and Section 5 provides a comprehensive discussion. Finally, the paper concludes in Section 6 with an outlook on future work.

2. Methodology

This study presents the UBM for generating short- to medium-term probabilistic forecasts while leveraging a point forecast model. Historical point forecasts are divided into bins based on the forecasted value range, and quantiles of the forecast error empirical cumulative distribution function (ECDF) are computed for each bin. While generating probabilistic forecasts, the relevant bin is identified by comparing the new point forecast with bins forecast ranges, and the error values at predefined quantiles from the ECDF for the chosen bin are added to the new point forecast to produce probabilistic predictions. Further explanation of the UBM is detailed in Section 2.1. To evaluate the performance of both probabilistic and point forecasting results, various evaluation metrics are implemented, as discussed in Section 2.2.

2.1. UBM Framework. The UBM framework is designed to generate probabilistic forecasts using output from deterministic forecast models, effectively bridging the gap between deterministic and probabilistic forecasting methods. First, a deterministic model processes historical measured data and related factors to generate point predictions. The forecast errors are calculated by comparing them with the actual measurements as given in Equation (1).

$$e(t) = \widehat{y}(t) - y(t), \tag{1}$$

where $e(t)$ denotes the point forecast error, $\widehat{y}(t)$ denotes the (historical) point forecast value and $y(t)$ denotes the actual measurement value at time t, respectively.

Subsequently, the point forecasts and the forecast errors are used as historical training data for the UBM. To enhance adaptability to uncertainties and evolving conditions, the model is trained dynamically by iteratively adding new point forecast values to the training data. The training data are used as input features for describing the conditional distribution of errors into various clusters based on the point prediction range. Figure 1 illustrates the division of the example dataset into clusters. Essentially, the data are sorted on the point forecast values and subsequently divided into a predefined number of clusters based on the point forecast range. The ECDF of the forecast errors for each cluster is then computed. Further explanations regarding the formation of clusters are detailed in Section 2.1.2. Following this, the point predictions generated by the deterministic model during the forecasting period of the UBM are compared with the point prediction range of clusters derived from the

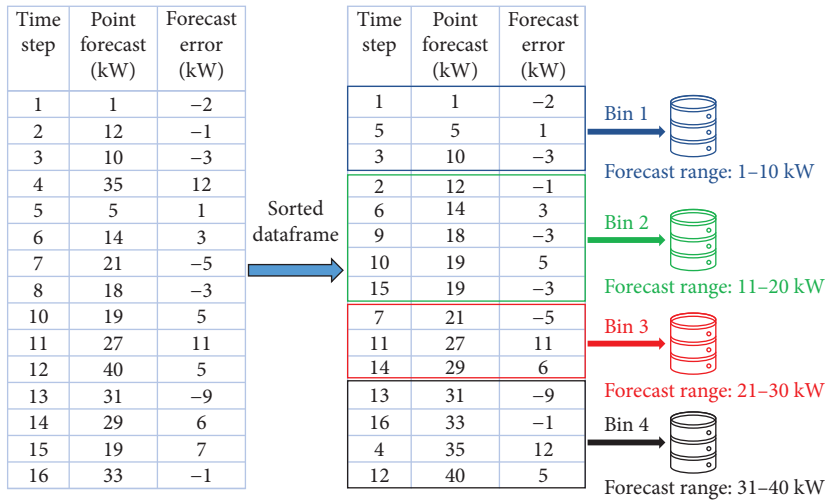

FIGURE 1: Cluster division using an example dataset (adapted from [32]).

training, facilitating the selection of the appropriate cluster. Finally, the error values at different quantiles of the error's ECDF for the selected cluster are combined with the point prediction value at each timestamp to generate the probabilistic predictions.

In the following, let t_0 denote the current time at which a new prediction is calculated. The algorithmic process steps of the framework are divided into three stages: the deterministic stage, training stage, and forecasting stage, as presented via the process flowchart in Figure 2. For each forecasting period $t_1, \ldots, t_N$, all three phases are executed. It is pre-requisite to have sufficient historical deterministic forecasts $\widehat{y}(t_i)$, $i<0$ and measured data $y(t_i), i<0$ to generate probabilistic predictions.

2.1.1. Phase I: Deterministic Stage. In the first phase, a deterministic model is implemented to generate the point forecasts $\widehat{y}(t_1), \ldots, \widehat{y}(t_N)$, where N denotes the number of predicted time steps. The UBM is compatible with any deterministic model, be it statistical or machine learning-based. In this study, the statistical method known as the personalized standard load profile (PSLP) is implemented to generate point forecasts for electricity and heat demands, as well as PV generation. Detailed explanations and applications of the PSLP for load forecasting can be found in [39–41]. Additionally, this work integrates and builds upon a recent study [42] that extended PSLP for heat demand and PV generation forecasting. The PSLP is expected to perform poorly for PV forecasting; however, it serves as an example to demonstrate the UBMs performance when a poor performance of the point forecast model is observed.

The historical measured data are collected to train the PSLP model. These data are categorized according to daytype (weekdays, Saturday, and Sunday) and season classifications (summer, transition, and winter), similar to SLPs derived by the German Association of Energy and Water Industries (BDEW) [39]. The daytype classification is not used for PV generation forecasts as it has a very low correlation. Forecasts are then generated by aggregating historical values within each category using statistical measures such as the mean, median, and maximum.

The PSLP training data grows as the model iterates to make it more adaptable to changes in the load profile. That means that the next day incorporates the measured values from the preceding day, and so on. Additionally, a rolling forecast was implemented to limit the training window with the maximum historical days from the day of the forecast, as discussed in [41, 42]. That means the training window slides as it iterates over time, while continuously updating the training data to include recent measurements while excluding older ones beyond the specified window. This approach ensures that the training window does not exceed a specified maximum historical period from the forecast date.

2.1.2. Phase II: Training Stage. Training of the UBM is carried out in this phase. First, the training data are clustered based on historical point forecast values. A predefined number of clusters (K) of equal width are defined as follows. Let $\mathcal{H} = \{t_{-1}, \ldots, t_{-M}\}$ denote the set of historical time points used for training, and let a and b denote minimal and maximal forecasted value, respectively, that is the following:

$$a = \min\{\widehat{y}(t) : t \in \mathcal{H}\},$$

$$b = \max\{\widehat{y}(t) : t \in \mathcal{H}\}.$$

The classification intervals $I_k, k = 1, \ldots, K$ are then defined as follows:

$$I_k = [a + (k-1)w, a + kw], \tag{2}$$

where $w = (b-a)/K$ denotes the bin width of the intervals. The bins are then defined as follows:

$$B_k = \{t \in \mathcal{H} : \widehat{y}(t) \in I_k\}, \; k = 1, \ldots, K. \tag{3}$$

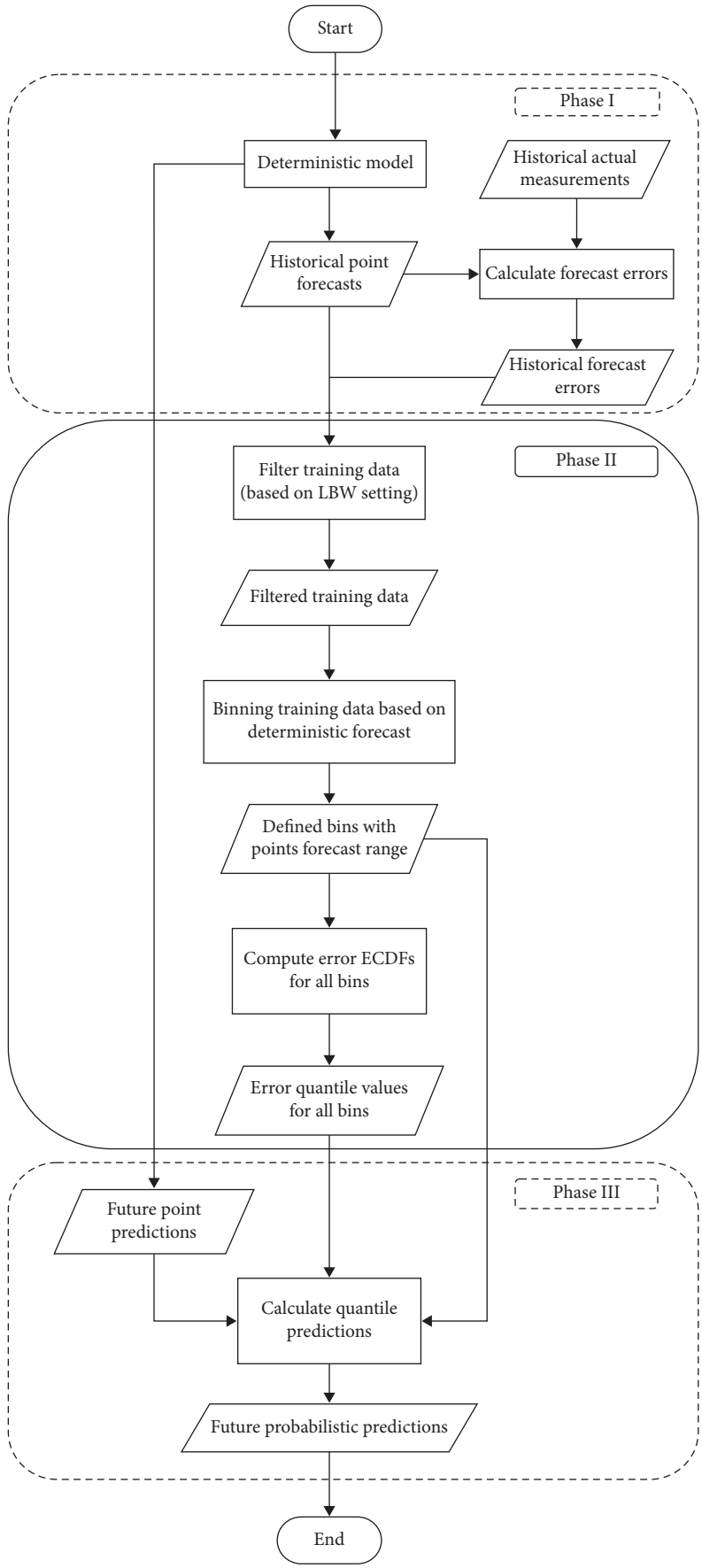

Figure 2: The UBM algorithm process flowchart (adapted ideas from [32]). UBM, uncertainty binning method.

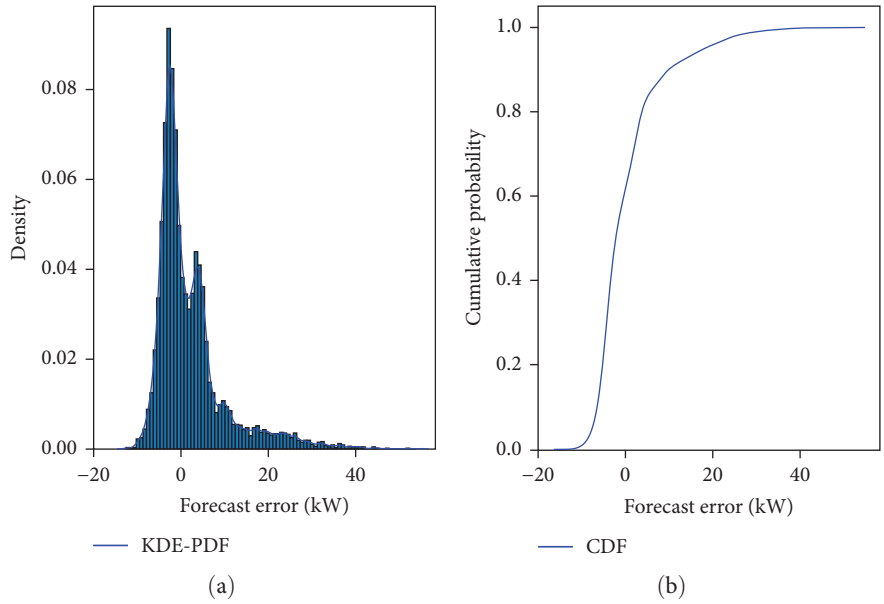

FIGURE 3: (a) Cluster's error histogram distribution and PDF using KDE and (b) cluster's error ECDF plot. ECDF, empirical cumulative distribution function; KDE, kernel density estimation; PDF, probability density function.

This binning process is repeated after each forecasting period, meaning that the point forecast intervals (I_k) used to define clusters change dynamically as new point forecasts are added to the training dataset at $\mathscr{H}$. In the next step, the point forecast errors $e(t)$, as given by Equation (1), are allocated to their respective bins (B_k). Histograms and probability density function (PDF) of point forecast errors are then generated for each bin (B_k) to gain insights into the spread, skewness, and presence of any outliers in the forecast error distribution. A PDF is fitted for each bin (B_k)using a nonparametric technique known as kernel density estimation (KDE) [43], as shown in Figure 3a and given by Equation (4).

$$f_k(x) = \frac{1}{n_k h_k} \sum_{t \in B_k} \mathscr{K}\left(\frac{x - e(t)}{h}\right), \tag{4}$$

where $f_k(x)$ is the density function at point x for bin k, $n_k = |B_k|$ is the number of data points, h_k is the bandwidth parameter that controls the smoothness of the resulting density curve, $\mathscr{K}$ is the kernal function (Guassian in this study), $e(t)$ are the individual errors.

The bandwidth h_k in Equation (4) is determined via Scott's rule [44] $h_k = n_k^{-1/(d+4)}$, where $d = 1$ denotes the dimensionality of the data.

Figure 3a shows histogram distribution of the errors and fitting of PDF using KDE, and Figure 3b shows the ECDF of the errors. The point forecast errors at predefined quantiles from the distribution are extracted from the ECDF. In this study, the ECDFs for each bin are obtained using the nearest rank method (NRM), a nonparametric approach. To determine the value of a given percentile q, it selects the corresponding value from the sorted data, such that a proportion q of the data points is smaller than the selected value. A detailed explanation of NRM for this study is given in the following:

Let $e_j, j = 1, \ldots, n_k$ denote the ordered error values for the kth bin, that is, $e_j = e(t_j), j = 1, \ldots, n_k$ with $B_k = \{t_1, \ldots, t_{n_k}\}$ such that $e_1 \leq \ldots \leq e_{n_k}$.

For a given percentiles $q = \{0.1, 0.2, \ldots, 0.9\}$, the rank R is calculated as follows:

$$R = q \cdot (n_k), \tag{5}$$

where $n_k = |B_k|$ is the number of data points for the kth bin.

The error percentile value $\varepsilon_k(q)$ for the kth bin at percentile q is then given by Equation (6), which accounts for the contributions of the lower and upper values to the overall estimation of the percentile value if R is not an integer.

$$\varepsilon_k(q) = \begin{cases} e_R & \text{if } R \text{ is an integer} \\ e_{\lfloor R \rfloor} \cdot (\lceil R \rceil - R) + e_{\lceil R \rceil} \cdot (R - \lfloor R \rfloor), & \text{otherwise} \end{cases}. \tag{6}$$

The error values at each percentile q for all the bins B_k given by Equation (6) are later used in the forecasting stage to convert point predictions into distributions of forecasts.

The UBM training time series keeps growing as the model iterates over the forecasting time period $t_1, \ldots, t_N$ to simulate the real case scenario. For example, at the current

time t_0 and the forecast horizon of 24 h, with the time resolution of 15 min, the measurement and point forecast values for the forecasting period $(t_1, \ldots, t_{96})$ are included in the training set to generate probabilistic predictions for the next forecasting period $(t_{97}, \ldots, t_{192})$. This means that the UBM is trained after every forecasting period with the new measurements and point forecasts added as part of the training set. This approach enables the capture of any uncertainties in future predictions and adapts to changing conditions. Additionally, in order to prevent an infinite growth of the training data set, the lookback window (LBW) feature is implemented and can be used to set a fixed predefined training window. This should be used as per the requirement and type of data profile to be forecasted. Activating the LBW feature provides the possibility of limiting the training window (with sliding) as the model iterates over the forecasting period $t_1, \ldots, t_N$, thus capturing the seasonal trends more accurately and avoid considering old data which are no longer relevant. This feature is implemented in this work to evaluate its potential impact on the performance of the UBM model on all three data profiles. The training windows of 180, 150, 120, 90, and 60 days are considered for analysis. For instance, if the LBW of 180 days is chosen, then the training set for a current time $t = 0$ will contain the historical point forecast values $\widehat{y}(t_i)$ for $i = -1, \ldots, -(180 \cdot m)$ where m represents the number of data points per day.

2.1.3. Phase III: Forecasting Stage. In this final phase, probabilistic predictions are generated using the trained UBM model. First, a forecasting time period is selected on the basis of the forecast horizon. In this work, the forecast horizon of 24 h was chosen, with the time resolution of 15 min. For instance, at the current time t_0, the first forecasting time period spans $t_1, \ldots, t_{96}$. During this period, new point forecasts $\widehat{y}(t_i)$ for $i = 1, \ldots, 96$ are obtained from the deterministic model. These forecasts are compared with the point forecast intervals (I_k) of all bins B_k as determined in Phase II to identify the appropriate bin at each timestamp.

Once the appropriate bin is selected, the model retrieves the error values $\varepsilon_k(q)$ at predefined quantiles q of ECDF for the k_{th} bin. These error quantile values represent the distribution of forecast errors for the specific interval of historical point forecast values (I_k). The error values $\varepsilon_k(q)$ at different quantiles q are then added to the new point forecast $\widehat{y}_{t_i}$ to generate a distribution of predictions $\widehat{y}_{t_i}(q)$ at time t_i, for $i > 0$ as given by Equation (7). This involves taking each error quantile and adding it to the point forecast value to create a range of possible forecast outcomes, effectively converting point prediction to probabilistic prediction.

$$\widehat{y}_{t_i}(q) = \widehat{y}_{t_i} + \varepsilon_k(q), \tag{7}$$

where $\widehat{y}_{t_i}(q)$ is quantile forecast at percentile q and time t_i, $\widehat{y}_{t_i}$ is point forecast value at the time t_i, $\varepsilon_k(q)$ is quantile point forecast error obtained from ECDF of the selected cluster k.

The process described above is then repeated for each timestamp within the forecasting period to generate a probabilistic prediction curve.

2.2. Performance Evaluation

2.2.1. Deterministic Forecast Evaluation. This work uses widely recognized evaluation metrics that are commonly used in the deterministic forecasting literature, as highlighted in [45–48]. The mean absolute error (MAE), mean square error (MSE), root MSE (RMSE), mean absolute percentage error (MAPE), and mean absolute scaled error (MASE) are used as the point forecast evaluation indices as given by the following equations:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \widehat{y}_i)|, \tag{8}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2, \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}, \tag{10}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|(y_i - \widehat{y}_i)|}{y_i}, \tag{11}$$

$$\text{MASE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y}_i|}{\frac{1}{n-m}\sum_{j=m+1}^{n}|y_j - y_{j-m}|}, \tag{12}$$

where y_i and y_j are measured values in the time series at time i and j, respectively, $\widehat{y}_i$ is the predicted value at time i, y_{j-m} is the naive forecast at previous time $j - m$, n is the total number of data points and m is the previous day for the naive forecast.

For PV forecasting, the MAPE metric encounters limitations due to the prevalence of zero measured data. This leads the MAPE calculation in Equation (11) to be undefined. To tackle this problem, the MASE is also implemented, which essentially compares the accuracy of the model with the naive forecast approach. In the MASE calculations for heat forecasts, a naive forecast obtained from the previous week, that is, 7 days ago, is used, whereas for electricity and PV forecasts, the naive forecast from the previous day is used.

2.2.2. Probabilistic Forecast Evaluation. Three main aspects are considered, namely "reliability" or "calibration," "sharpness," and "resolution" while evaluating the performance of the probabilistic forecasting [21, 38]. Reliability measures the credibility of the probabilistic forecast model in capturing the actual values within the PI. Sharpness measures the spread of interval width or concentration of predictive distribution. Resolution evaluates the model's effectiveness in minimizing sharpness while maintaining reliability within an acceptable range. These aspects are assessed using six metrics in this study for a comprehensive evaluation of probabilistic forecasting. All the numerical evaluation metrics are summarized in Table 1.

a. PICP: The PICP measures reliability within the PI, represented in percentage (%). For a good forecast, PICP is expected to be closely aligned with the PI

Table 1: Summary of all the numerical probabilistic forecasting evaluation metrics.

Score	PICP	MPIW	WS	QCS	PQCS
Definition and purpose	Measures the reliability by calculating the percentage of actual values falling within the prediction interval	Measures the sharpness of the prediction interval	Considers both reliability and sharpness, rewarding narrower widths but penalizing when observed values fall outside	Numerical score for GCM, which measures reliability at each quantile bin	Numerical score for GCM, which measures reliability at each quantile bin and overcomes the limitation of QCS, i.e., data size-dependent
Unit	Percentage (%)	Unit of entity	None	None	Percentage (%)
Interpretation	Higher values indicate better reliability	Lower values indicate better sharpness	Lower scores indicate better overall forecast quality	Lower values indicate better consistency	Lower values indicate better consistency
Limitations	Scale-dependent, do not consider reliability at each quantile bin, overestimation of reliability is not penalized	Scale-dependent	Biased toward sharpness, Sensitive to outliers	Limited when comparing datasets of different sizes, Produces invalid scores during uniform data periods (e.g., zero power generation)	Produces invalid scores during uniform data periods (e.g., zero power generation)
Strengths	Good and popular score for measuring overall reliability	Good and popular score for measuring overall sharpness	Takes both reliability and sharpness while scoring	Provides numerical scoring to the visualization GCM score, considers reliability at each quantile bin, not biased toward sharpness over reliability	Overcomes the limitation of QCS being data size-dependent, considers reliability at each quantile bin, not biased toward sharpness over reliability

Abbreviations: GCM, graphical calibration measure; MPIW, mean prediction interval width; PICP, prediction interval coverage probability; PQCS, percentage quantile calibration score; QCS, quantile calibration score; WS, Winkler score.

[35, 49]. PICP is expressed as Equation (13):

$$\text{PICP} = \frac{1}{N}\sum_{i=1}^{N} C_i, \tag{13}$$

where C_i is given by the following:

$$C_i = \begin{cases} 1 & \text{if } y_i \in [L_i, U_i], \\ 0 & \text{if } y_i \notin [L_i, U_i] \end{cases},$$

where y_i is the measured values in the time series at time i, L_i, and U_i are the lower and upper predictive percentiles at time i, N is the total number of data points, and C_i the coverage factor.

b. MPIW: The MPIW measures the sharpness of the probabilistic forecasts. However, it is scale-dependent; that is, MPIW is not favorable for comparing probabilistic forecast results for different datasets. A higher PICP and a lower MPIW are desirable, but both conflict with each other [36]. A tradeoff between the PICP and MPIW must be made while evaluating probabilistic forecasts. The MPIW is expressed as Equation (14):

$$\text{MPIW} = \frac{1}{N}\sum_{i=1}^{N} (U_i - L_i), \tag{14}$$

c. WS: The WS considers both reliability and sharpness for evaluation [8]. A low score indicates better probabilistic forecasting. For a central $(1-\alpha)$ PI, where $\alpha \in (0, 1)$, WS is expressed as Equation (15):

$$\text{WS}_{i,\alpha} = \begin{cases} \delta & \text{if } L_{i,\alpha} \le y_i \le U_{i,\alpha}, \\ \delta + \dfrac{2\left(L_{i,\alpha} - y_i\right)}{\alpha} & \text{if } y_i < L_{i,\alpha}, \\ \delta + \dfrac{2\left(y_i - U_{i,\alpha}\right)}{\alpha} & \text{if } y_i > U_{i,\alpha} \end{cases}, \tag{15}$$

where y_i is the measured value at time i, $L_{i,\alpha}$, and $U_{i,\alpha}$ are the lower and upper predictive percentiles at level $\alpha/2$ and $1 - \alpha/2$ and δ is the PI width given by the difference between the upper and lower predictive percentiles $(U_{i,\alpha} - L_{i,\alpha})$.

The score rewards narrower PI widths and penalizes if the measured values fall outside the PI [50]. However, the WS is more biased toward sharpness than reliability, that is, it gives a better score for sharper and moderately reliable probabilistic forecasts compared to less sharp and highly reliable ones. It is also sensitive to outliers. Furthermore, the score is scale-dependent.

d. QCS: The QCS assesses reliability with consideration to each quantile bin and penalizes any deviation from expected reliability. In this study, the quantile bins are formed with equal width (10%) percentile ranges (i.e., 0%–10%, 11%–20%,…, 91%–100%). A low QCS value is desired, and the perfect forecast is obtained when the QCS is 0, that is, when the reliability at each quantile bin matches exactly the expected reliability. There is no upper limit for the QCS. It basically rewards when the frequency of observed values (O_i) matches the expected frequency (E_i) at each quantile bin (i) and penalizes deviation from expected frequency (E_i), that is, when high or low sharpness forecasts are obtained for example [32]. The QCS is independent of scale; however, it faces limitations when comparing datasets of different sizes. Additionally, the score becomes invalid for applications with prolonged data uniformity, such as zero power generation in PV forecasting. In such scenarios, all quantile forecasts may converge to the same value (e.g., zero), causing every point to be valid across all quantile bins. This results in significantly inflated O_i for each bin, leading to a disproportionately high QCS due to the penalization for deviations from E_i.

The QCS is represented by Equation (16).

$$\text{QCS} = \frac{1}{n}\sum_{i=1}^{n} \frac{(E_i - O_i)^2}{E_i}, \tag{16}$$

where n is the number of quantile bins (e.g., 10 in this study), E_i and O_i are the expected frequency and the observed frequency, respectively, for each quantile bin (i).

e. PQCS: To overcome the limitation of the QCS, the PQCS is implemented, which is expressed as a percentage (%). It is scale- and size-independent; however, like QCS, it becomes invalid for applications with prolonged data uniformity. Lower PQCS values are preferred, with the optimal score being 0%. The PQCS is calculated as Equation (17).

$$\text{PQCS} = \frac{1}{n}\sum_{i=1}^{n} \frac{|E_i - O_i|}{E_i} \times 100. \tag{17}$$

f. GCM: The GCM is a graphical evaluation tool to assess reliability and sharpness [32]. The GCM plots the bar charts for the actual percentage of observed values (blue-colored bars) in each quantile bin, comparing them to the expected percentage (red line), as shown in Figure 4. Any deviation of actual observed values from the expected observed values for each quantile bin is penalized, resulting in a higher score for its numerical metrics (QCS and PQCS), effectively addressing both under- and overestimations in forecast reliability. Furthermore, the shape of the GCM plot provides insight into the forecast's sharpness. The perfectly calibrated probabilistic forecasting is

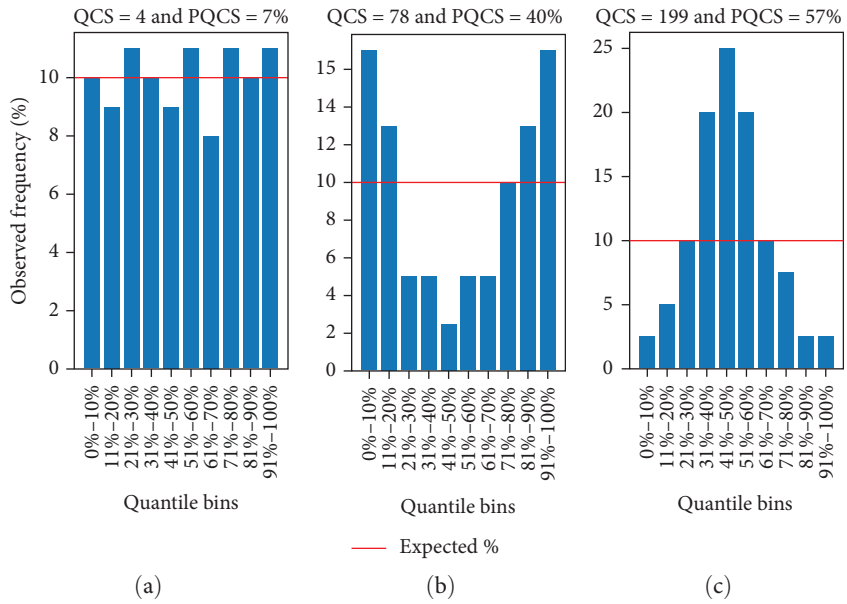

FIGURE 4: GCM plots for three different cases: (a) nearly uniform, (b) high sharpness, and (c) low sharpness. The QCS and PQCS values for each of these cases are shown in the figure. GCM, graphical calibration measure; PQCS, percentage quantile calibration score; QCS, quantile calibration score.

obtained when the GCM exhibits a uniform distribution shape, that is, when QCS is 0 and PQCS is 0%. Conversely, the GCM with a triangular shape distribution, as shown in Figure 4c, indicates low sharpness and U-shaped distributions, just as Figure 4b, indicate high sharpness probabilistic predictions, and Figure 4a shows a nearly uniform shaped distribution with low QCS and PQCS values. Both high and low-sharpness forecasts are penalized by the QCS and PQCS because high-sharpness is often unreliable across quantile bins, and low-sharpness probabilistic forecasts are not ideal for decision-making [32]. Figure 4 shows how the QCS and PQCS scores change with the shape of the GCM, offering distinct scores for each of these cases. Moreover, they are not biased toward sharpness over reliability.

3. Case Study

The commercial logistics facility located in northern Germany is considered for the demonstration of the proposed method. This logistics center represents the integration (or "sector–coupling") of electricity, heat, cooling, and transport. A detailed description of the integrated energy system at the logistics facility can be found in [51]. To verify the effectiveness and accuracy of the method proposed in this paper, the measurements from the ElogZ [52] project are utilized. Electricity and heat meters were installed in different subdistributions of the logistics facility. For this case study, the electricity and heat demand of each subsystem were aggregated. Space heating and the demand for domestic hot water are categorized under the heat sector. To meet these demands, the system utilizes two cascades of air heat pumps, a heat-water buffering storage unit, and two gas boilers to handle peak demand. The cooling needs for the building are met using heat pumps/chillers and a cold water storage unit. The cooling demand is considered in the context of the heating demand for this study. Additionally, servers are equipped with individual air conditioning systems. All electricity needs for the logistics center are supplied by the low-voltage grid. This includes the electricity requirements of the office building, warehouses, dormitories, and other facilities within the center, as well as the energy needs for refrigerated trailers (conditioning, precooling, and maintaining of cold chains) associated with the transport sector. Figure 5a,b depicts the office building (business center) and dormitory, respectively. Recently, rooftop PV panels were installed in the warehouse to boost local electricity generation for the facility, as shown in Figure 5c. Due to the lack of PV generation measurement data for this study, PV generation was simulated with a system size of 200 using the Python library pvlib [53], utilizing publicly available weather data obtained from open DWD [54]. The assumption made in the simulation for the PV modules is that they are oriented half to the east and half to the west, with an inclination angle of 10°. Battery and power electronics for control are installed in the dormitory, as depicted in Figure 5d. Air source heat pump systems are also integrated within this facility. Both deterministic and probabilistic forecasts are generated as an essential component of the EMS for the given distributed integrated energy system. The observation, forecasting, and waiting periods for both methods are detailed in Table 2. All of the data used in this study, including measured values and generated point and probabilistic forecasted output results, were considered at a 15-min resolution and represented in kW. The forecasting framework (UBM) and its evaluation are implemented in Python programing language.

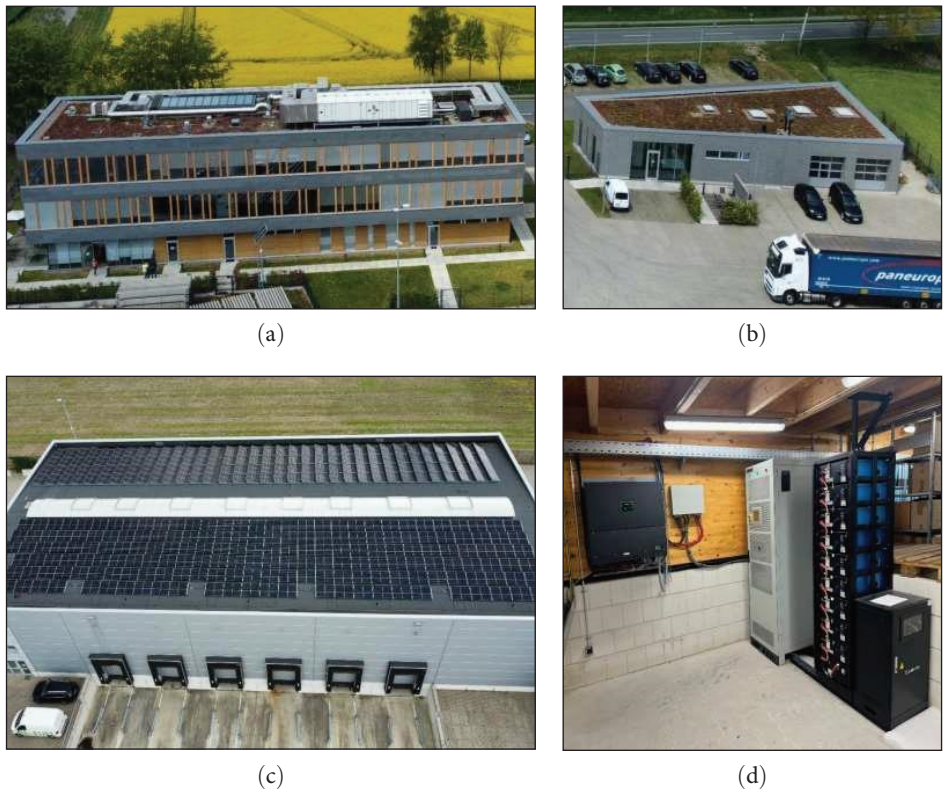

FIGURE 5: Logistics facility [52]: (a) office building, (b) dormitory, (c) warehouse with rooftop PV, and (d) battery storage and power electronics.

TABLE 2: Dataset description providing observation period, number of waiting days, and observation period for both point forecast (PSLP) and probabilistic forecast (UBM).

Forecasting parameters	Deterministic forecast (PSLP)	Probabilistic forecast (UBM)
Observation period	5 Sep 2021–30 Aug 2022	26 Sep 2021–30 Aug 2022
NWDs	21 days	7 days
Forecasting period	26 Sep 2021–30 Aug 2022	5 Oct 2021–30 Aug 2022

Abbreviations: NWDs, number of waiting days; PSLP, personalized standard load profile; UBM, uncertainty binning method.

TABLE 3: PSLP scores averaged (mean) for the whole forecasting period.

Sector	Scores				
	MAE (kW)	MSE (kW)	RMSE (kW)	MAPE (%)	MASE
Electricity	6.34	78.29	8.71	13.18	0.65
Heat	6.4	76.65	8.52	26.55	0.6
PV	9.9	441.76	20.62	—	0.93

Abbreviations: MAE, mean absolute error; MAPE, mean absolute percentage error; MASE, mean absolute scaled error; MSE, mean square error; PSLP, personalized standard load profile; RMSE, root mean square error.

4. Results

4.1. Deterministic Forecast. In this section, the forecast results and accuracy of the PSLP for electricity, heat, and PV profiles are presented and compared. After setting the number of waiting days (NWDs) to 21 as the basis for the PSLP, point forecasts were generated for a period between 26 September 2021 and the end of August 2022, as shown in Table 2. For the same period, the mean of the evaluation scores for the electricity demand, heat demand, and PV generation were obtained, as displayed in Table 3. Figure 6 illustrates the PSLP forecasts and its comparison with the actual values for a random day (8 June 2022). The mean MAE for the electricity (6.34 kW) and heat demand forecasts (6.4 kW) are within a similar range, whereas the PV generation forecasts exhibit a higher value of 9.9 kW.

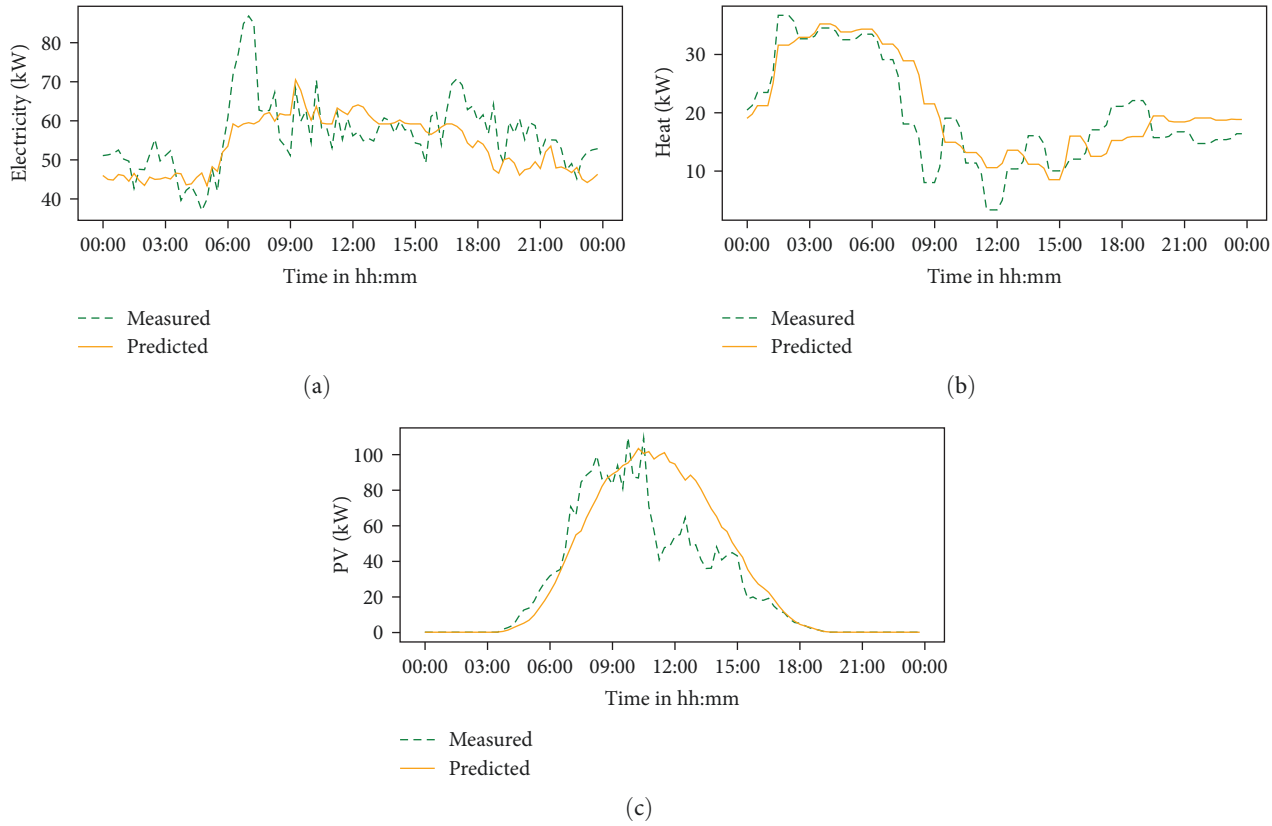

FIGURE 6: PSLP forecasting results on a random day (08 June 2022) for (a) electricity, (b) heat, and (c) PV. PSLP, personalized standard load profile.

Moreover, a high MSE value of 441.76 kW for the PV forecasts indicates that there are significant forecast errors on numerous timestamps throughout the forecasting period. The lowest MAPE score was obtained for electricity (13.18%), followed by the heat (26.55%). With respect to MASE, it can be observed that the PSLP performed better than the naive forecast for all of the sectors, as it was below 1.

The hourly forecasting error distribution for the entire forecasting period in a boxplot format is shown in Figure 7. The box represents the interquartile range (IQR), which encompasses the values between the 25th and 75th percentiles. The black line inside the box represents the median, which corresponds to the 50th percentile (the median). The upper and lower whiskers extend from the box by a maximum of 1.5 times the IQR. Figure 7 illustrates that the PV predictions exhibit wider error spread distributions during the daytime hours compared to heat and electricity demand forecasts. That means the PSLP resulted in poor forecasting for PV generation, as expected, which could impact decision-making in EMS. In such cases, probabilistic forecasting using UBM could be beneficial for informed decision-making, as elaborated upon in Section 5.

4.2. Probabilistic Forecast. In this section, the forecast results and accuracy of the UBM for the electricity, heat, and PV profiles are presented. In this study, the NWD for the UBM was 7 days, resulting in the generation of probabilistic forecasts from 3 October 2021 to the end of August 2022, as shown in Table 2. The probabilistic forecasts compromise percentiles ranging from 10% to 90%, representing an 80% PI and forming the basis for its performance evaluation. Later, the forecasts were also evaluated with different PI levels of 40% and 60%. Several forecast evaluation metrics (PICP, MPIW, GCM, QCS, and PQCS) are used to assess its performance, as discussed in Section 2.2.2. The PICP and MPIW were calculated daily, while the QCS and PQCS were computed monthly to ensure an adequate number of data points for their calculation. The QCS and PQCS metrics are not applicable to PV, as the predictions at different percentiles often coincide with the measured values, particularly during periods of zero power generation, as exemplified in Figure 8.

4.3. Probabilistic Forecast Results Without LBW Feature. The mean of the evaluation metrics presented in Table 4 were computed for the probabilistic forecasting period (Table 2), with the LBW feature being inactive. This means that the training dataset progressively expands over the forecasting period without being limited by the training window. Monthly PICP, MPIW, WS, QCS, and PQCS were averaged over the entire forecasting period across sectors for different number of bins, as shown in Table 4. Figure 8 shows the probabilistic output results from the UBM for a random day (8 June 2022). On this day, it was observed that the UBM performed with high reliability across all sectors; that is, the majority of the measured values, denoted by green

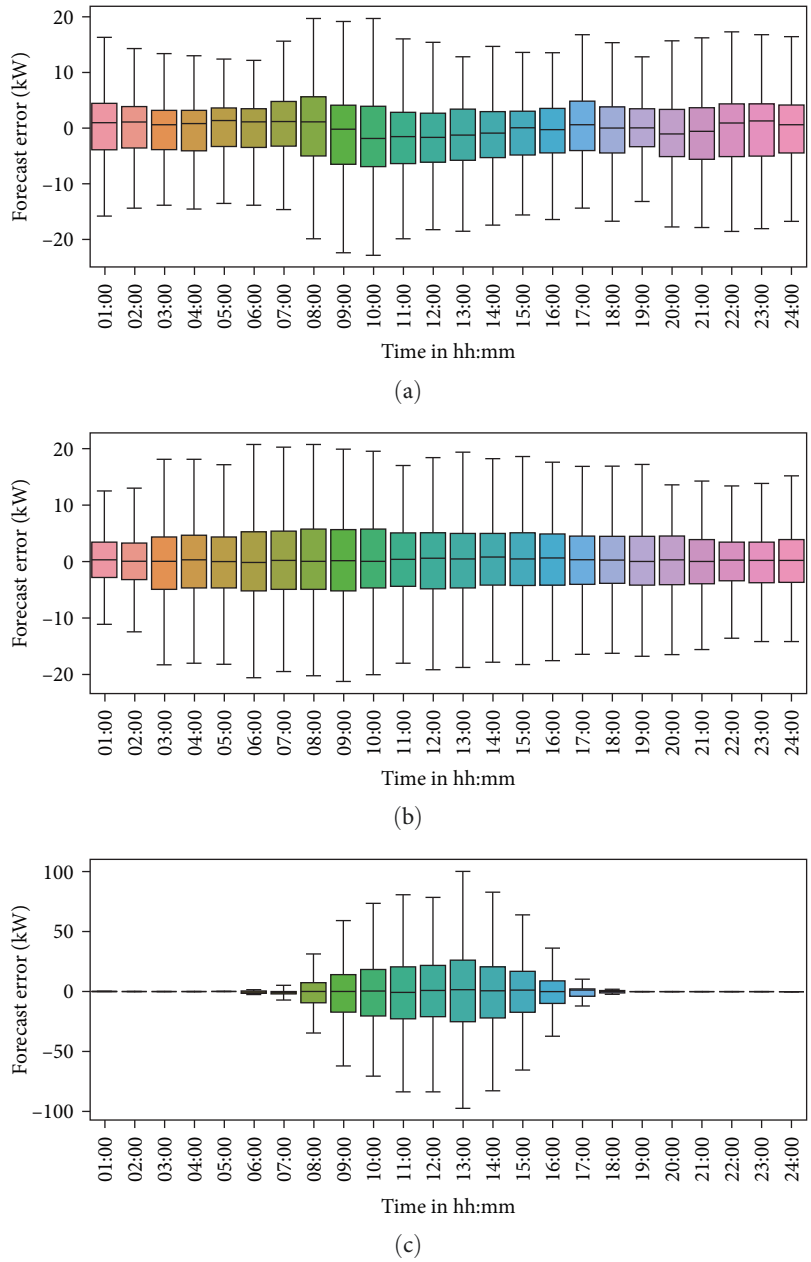

Figure 7: Hourly PSLP forecasting error boxplot distributions for (a) electricity, (b) heat, and (c) PV. PSLP, personalized standard load profile.

crosses, fall within the PI. However, the UBM exhibited low-sharpness PV forecasts, especially during peak generation hours.

From Table 4, the analysis reveals that the UBM model consistently generated reliable probabilistic forecasts with different number of bins across all sectors, with the average PICP being closely aligned with the PI level (80%). For the electricity forecast, the highest PICP (80.32%) was obtained with three bins, but at the same time highest MPIW (19.71 kW), that is, the lowest sharpness, was observed. Considering the WS, QCS, and PQCS, the best forecast for electricity was obtained with seven bins. For the same, the WS, QCS, and PQCS are 30.54%, 20.99%, and 20.95%, respectively. For the heat demand forecast, the highest PICP was obtained with three bins; however, this also resulted in the lowest sharpness, that is, the highest MPIW of 22.42 kW. While the lowest WS (31.54) was observed with nine bins. Considering QCS and PQCS, the best forecast for the heat sector was obtained with seven bins, yielding QCS and PQCS values of 23.31% and 22.53%, respectively. For the PV forecast, the best UBM accuracy was achieved with 12 bins, which corresponded to the lowest WS of 48.58. However, it is worth noting that this also resulted in the lowest sharpness, as indicated by the highest MPIW among the bin configurations. Based on the overall scores in Table 4, the optimal number of bins for electricity, heat, and PV were determined

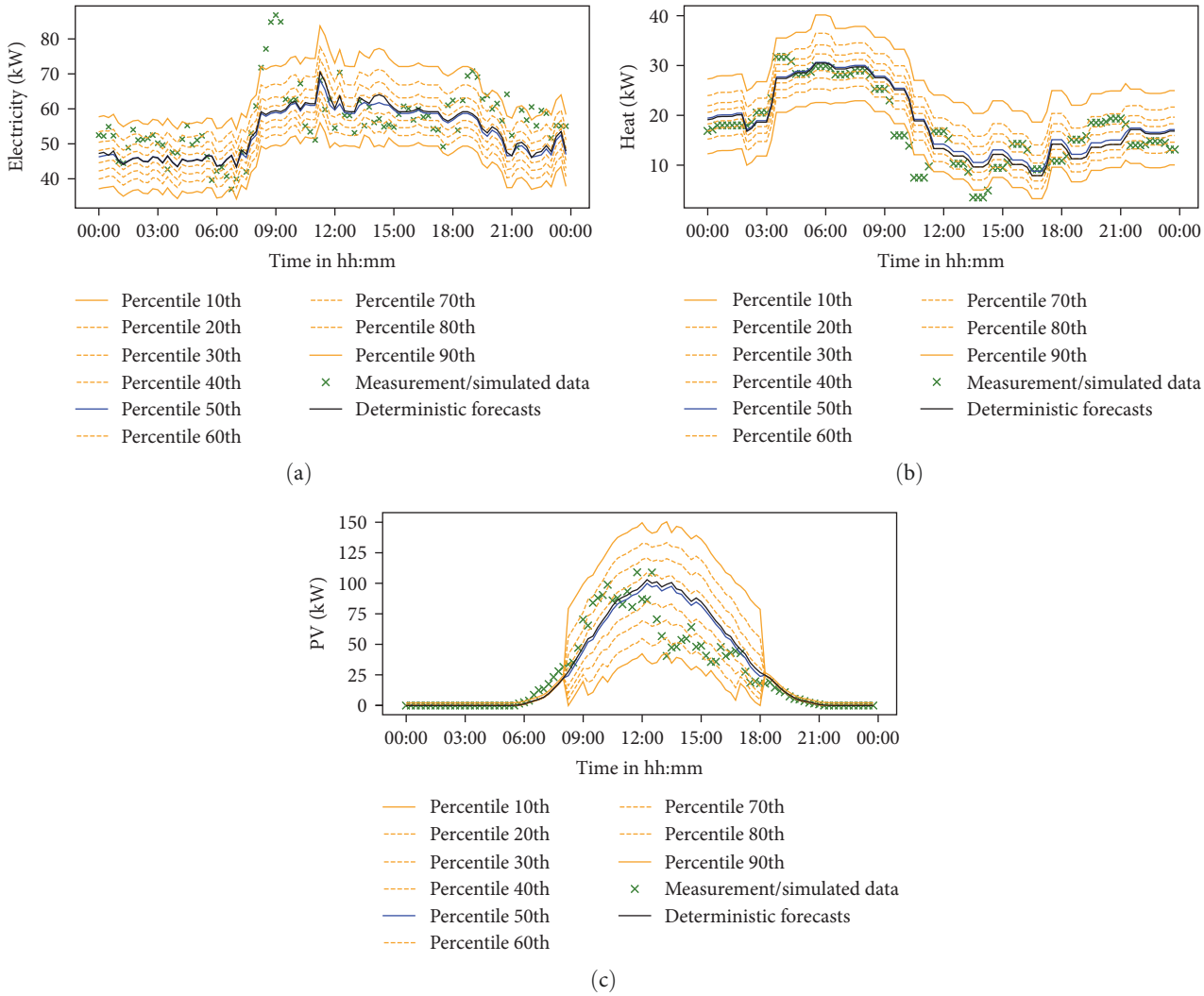

FIGURE 8: UBM forecasting results on a random day (08 June 2022) for (a) electricity, (b) heat, and (c) PV. UBM, uncertainty binning method.

TABLE 4: UBM scores averaged (mean) for the whole forecasting period across sectors for different number of bins.

Sector	Number of bins	Scores				
		PICP (%)	MPIW (kW)	WS	QCS	PQCS (%)
Electricity	3	80.32	19.71	30.81	21.23	21.13
	7	79.52	19.59	30.54	20.99	20.95
	9	78.95	19.48	30.57	23.63	22.48
	12	79.04	19.52	30.66	22.62	21.88
Heat	3	82.32	22.42	31.64	27.43	23.95
	7	81.45	22.29	31.59	23.31	22.53
	9	81.55	22.33	31.54	29.88	25.25
	12	80.89	22.05	31.88	30.03	25.72
PV	3	81.32	26.06	59.30	—	—
	7	80.62	30.09	50.99	—	—
	9	80.46	31.40	49.45	—	—
	12	81.37	32.54	48.58	—	—

Abbreviations: MPIW, mean prediction interval width; PICP, prediction interval coverage probability; PQCS, percentage quantile calibration score; QCS, quantile calibration score; UBM, uncertainty binning method; WS, Winkler score.

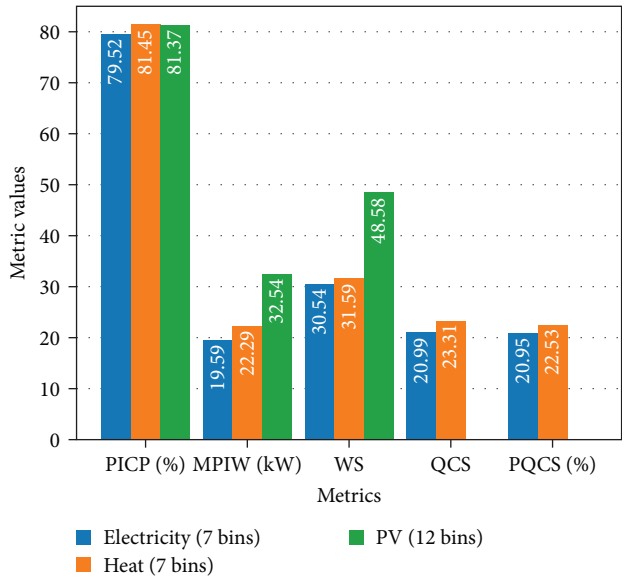

FIGURE 9: UBMs best results of each sector for the selected number of bins. UBMs, uncertainty binning methods.

to be 7, 7, and 12, respectively. Consequently, the results presented are derived from this selection of the number of bins for each sector. Figure 9 illustrates the comparison of scores across different sectors based on their respective optimal bin configurations.

From Figure 9, it can be observed that UBM achieved the highest PICP value of 81.45% for heat forecasts, whereas the lowest PICP value was obtained for electricity forecasts. With respect to the MPIW, the UBM excels in producing sharp forecasts for electricity (19.59 kW) and heat demand (22.29 kW) compared to PV generation (32.54 kW). A high interval width (i.e., low sharpness) was observed for PV generation, which can be interpreted from a high MPIW. Conversely, the UBM achieved the highest sharpness (i.e., lowest MPIW) for the electricity profile with a value of 19.59 kW. The lowest WS was observed for electricity (30.54), followed by heat (31.59), but exhibited notably high WS for PV (48.58). Considering the variability of PV generation and inaccuracy of the PSLP forecasts output, a high MPIW and WS were observed, as expected. Based on the QCS and PQCS, the highest calibrated forecast was observed for electricity, with values of 20.99 and 20.95%, respectively, compared to the heat sector, with values of 23.31% and 22.53%, respectively.

The performance of the UBM, as measured by PICP, MPIW, WS, QCS, and PQCS, on the electricity, heat, and PV data profiles for each month throughout the forecasting period, is depicted in Figure 10, with the LBW feature being inactive. The black dashed line in the first subplot depicts PI (i.e., 80%) to indicate the deviation from the expected PICP. Notably, higher PICP values were observed in the spring and summer months compared to the winter months. During the same period, higher MPIW values for PV were observed, possibly due to higher fluctuation in the data profile. The MPIW for the electricity experienced a slight rise during the spring and summer, while the opposite trend was observed for a heat profile. The trend of the WS can be interpreted by combining the trends of the PICP and MPIW. The lowest WS for PV (24.79 kW) was observed in December 2021, while the lowest WS for electricity (23.62 kW) and heat (22.83 kW) was observed in June 2022 and August 2022, respectively. Moreover, the monthly QCS and PQCS values were observed for the electricity and heat profiles. The lowest QCS and PQCS values for electricity (2.2% and 6.8%, respectively) were observed in October 2021. The lowest QCS and PQCS values for the heat forecasts were 11.8 in July 2022 and 14.2% in October 2021, respectively. The highest QCS and PQCS, that is, the lowest reliability at each bin for the electricity profiles, were observed in May 2022 with values of 44% and 33.2%. On the other hand, for the heat profile, the highest QCS (47.5) and PQCS (35.1%) were observed in March 2022 and November 2021, respectively. Overall, the UBM demonstrated relatively good accuracy for electricity compared to the heat sector based on WS, QCS, and PQCS. The GCM plots for the electricity in April and May 2022 are shown in Figure 11 as an example. It was observed that both the QCS and PQCS penalize deviations from the expected percentage. This is evident in the way the values of QCS and PQCS change with the shape of the GCM, making them valuable scores for further investigating the accuracy of the probabilistic forecasts. They offer a more comprehensive reliability assessment, along with insights into the sharpness of probabilistic forecasting.

The results presented above were based on 80% PI as a benchmark for evaluating the probabilistic forecast. In the following, forecasts are evaluated across different PI levels, including 40% and 60%, using the optimal bins for each sector, as given in Table 5. A PI of 100% is excluded from consideration because it encompasses all possible outcomes, making it overly broad and uninformative for practical decision-making in energy management. It fails to provide actionable insights or the precision needed to assess risks or optimize resource allocation, which is essential for the intended application. The QCS and PQCS remain unchanged across different PI levels, as they are computed considering the entire quantile range from 0% to 100% with a fixed bin width of 0.1th or 10th percentile. Consequently, these metrics are not included in Table 5. It can be observed that the PICP remains closely aligned with the PI levels for each sector, except for the PV with a 40% and 60% PI, which resembles over-coverage, reflected in a PICP value of 67.33% and 74.41%, respectively. From Table 5, it is evident that with increased PI levels, the interval width also increases, as indicated by MPIW values, to cover the expected real values within the quantile range. This leads to also increase in WS, as the score emphasizes sharpness, as discussed in Section 2.2.2.

4.4. Probabilistic Forecast Result With LBW Feature. To show the effect of the LBW feature on the UBM performance, evaluation metrics were calculated for electricity, heat, and PV with varying sliding training windows (180, 150, 120, 90, and 60 days) with an optimal number of bins and PI of 80%, as shown in Table 6. For all the data profiles, reducing the training window led to a slight decrease in the MPIW, WS, QCS, and PQCS. However, it should be noted that PICP also

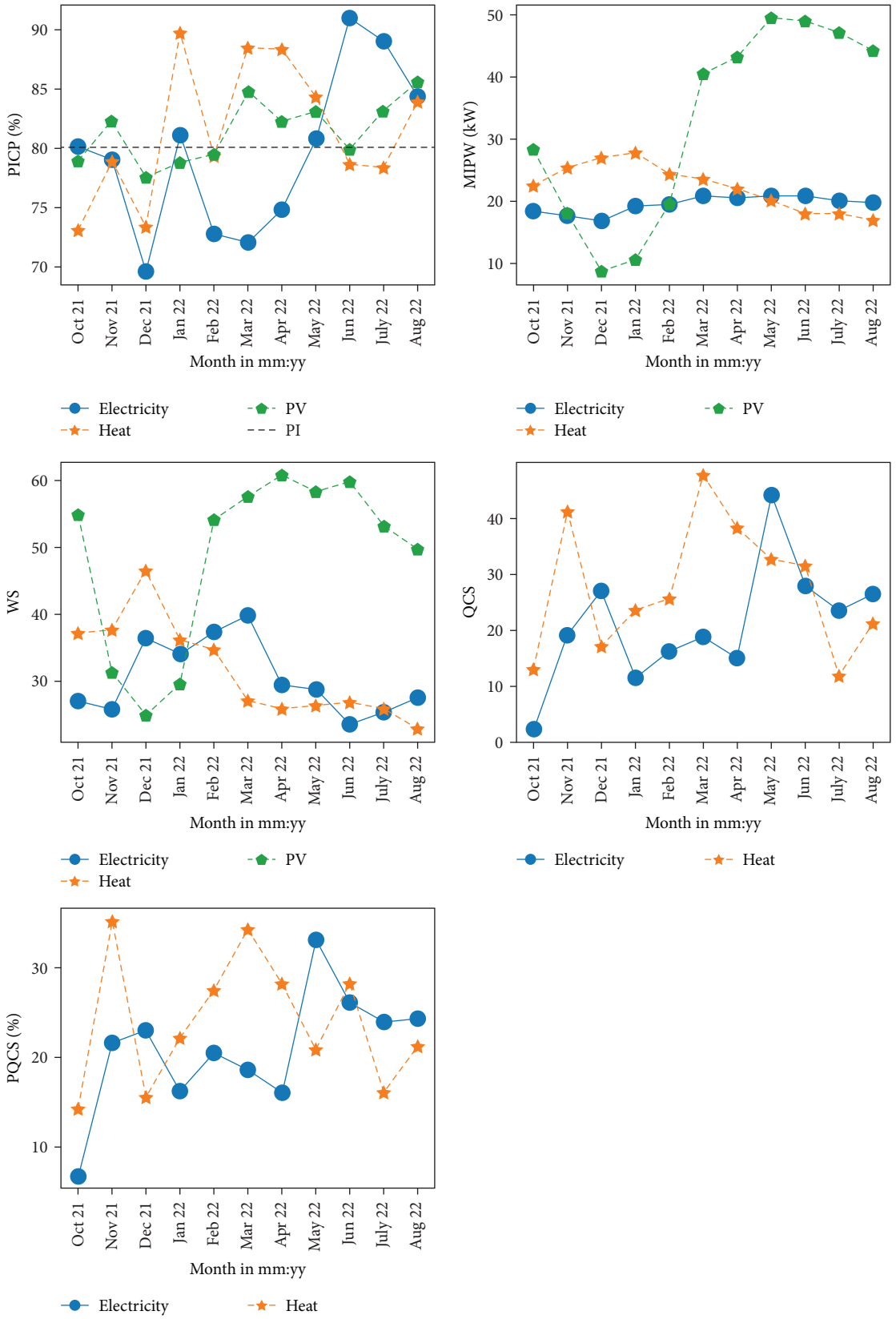

Figure 10: Scoring metrics calculated for each month in the forecasting period with LBW feature inactive. LBW, look back window.

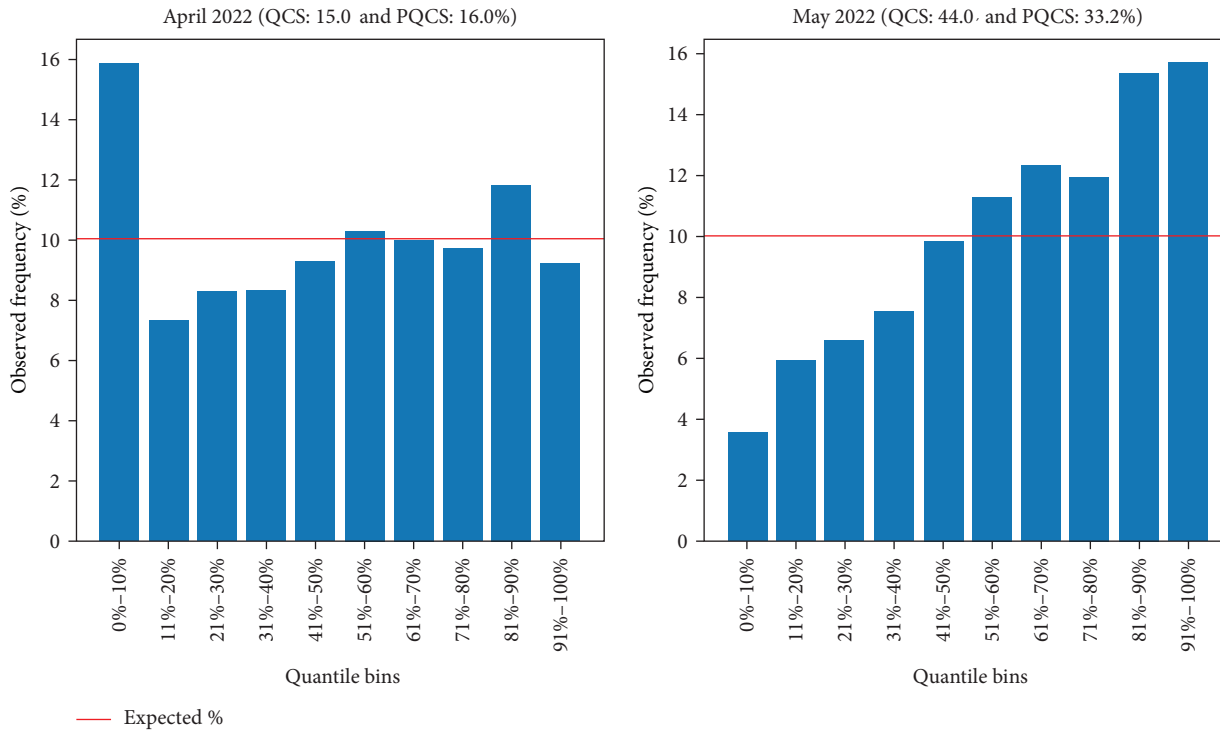

FIGURE 11: GCM plots for electricity demand on (a) April 2022 and (b) May 2022. GCM, graphical calibration measure.

TABLE 5: UBM scores averaged (mean) for the whole forecasting period across sectors with their selected optimal number of bins for different PI.

Sector	Scores			
	PI (%)	PICP (%)	MPIW (kW)	WS
Electricity	40	39.7	7.55	18.76
	60	59.28	12.36	23.13
	80	79.52	19.59	30.54
Heat	40	42.99	8.3	19.34
	60	63.22	13.9	24.01
	80	81.45	22.29	31.59
PV	40	67.33	13.96	29.67
	60	74.41	22.07	36.53
	80	81.37	32.54	48.58

Abbreviations: MPIW, mean prediction interval width; PI, prediction interval; PICP, prediction interval coverage probability; UBM, uncertainty binning method; WS, Winkler score.

decreases simultaneously. The decrease in these metrics can be attributed to the smaller training dataset size, leading to lower data variability over shorter time frames. Consequently, this results in a reduced spread of error distributions, subsequently lowering evaluation metrics values. In the case of electricity, the lowest MPIW (19.67 kW), QCS (18.25), and PQCS (18.71%) were obtained with a training window of 60 days. However, the PICP was lowest (78.13%) compared to the other training windows. On the other hand, the lowest WS (30.54) was obtained with a training window of 90 days. The choice between achieving a higher reliability or a higher sharpness depends on the specific application. In the case of the heat profile, the lowest WS (31.02) was obtained with a 90-day training window, whereas the lowest QCS (18.96) and PQCS (19.97) scores were obtained with 60 days. For the PV forecasts, the highest PICP and lowest WS was achieved with a 150-day training window. However, overall, it can be said that with shorter training window results in better performance of probabilistic forecasting, considering WS, QCS, and PQCS, which take both reliability and sharpness into account.

Table 6: UBM mean scores for the varying training window with LBW feature active.

Training horizon	Scores				
	PICP (%)	MPIW (kW)	WS	QCS	PQCS (%)
Electricity					
180 days	79.65	19.75	30.66	21.16	21.12
150 days	79.77	19.81	30.61	20.15	20.63
120 days	79.55	19.8	30.6	18.33	19.39
90 days	79.17	19.8	30.54	17.88	18.91
60 days	78.13	19.67	30.6	18.25	18.71
Heat					
180 days	80.0	21.46	31.33	25.55	23.16
150 days	79.67	21.2	31.27	23.95	22.78
120 days	78.95	20.81	31.11	21.55	21.12
90 days	78.37	20.6	31.02	22.67	21.45
60 days	78.3	20.58	31.09	18.96	19.97
PV					
180 days	81.32	31.36	48.85	—	—
150 days	81.32	31.33	47.98	—	—
120 days	81.23	30.74	48.69	—	—
90 days	81.37	30.25	48.37	—	—
60 days	81.11	29.6	48.52	—	—

Abbreviations: LBW, look back window; MPIW, mean prediction interval width; PICP, prediction interval coverage probability; PQCS, percentage quantile calibration score; QCS, quantile calibration score; UBM, uncertainty binning method; WS, Winkler score.

5. Discussions

Due to the increased uncertainty and requirement of low-cost EMS at the distribution level, a simple and robust forecasting framework (UBM) was introduced in this study. It was found to be computationally fast, reliable, adaptable across sectors, low feature engineering, and relies on easily accessible data. The UBM generates probabilistic forecasting by levering deterministic models, in this case, PSLP. The electricity demand, heat demand, and PV generation forecasting were produced for the distributed integrated energy system at a logistics facility in northern Germany. This work also emphasizes the limitations of commonly used evaluation metrics for probabilistic forecasting and illustrates how a more comprehensive evaluation can be achieved by incorporating new metrics such as GCM, QCS, and PQCS.

The PSLP was evaluated using popular metrics (MAE, MSE, RMSE, MAPE, and MASE) for all the sectors. The mean of these metrics across the entire forecasting period was calculated, and it was found that the PSLP performed better for electricity and heat demand forecasts but exhibited high forecasting errors for PV generation, as expected. This shortcoming can be attributed to the inherent variability in PV generation, as illustrated in Figure 12, where generation over three consecutive days highlights the significant variability and notable forecast errors incurred by the PSLP model. This will directly affect probabilistic forecasting generated by the UBM.

The UBM performance was evaluated with a different number of bins for each sector. The optimal bins for electricity, heat, and PV were found to be 7, 7, and 12 bins. The

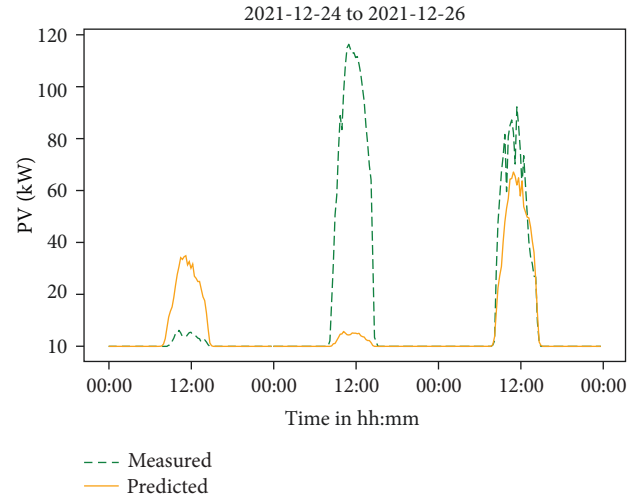

Figure 12: Variability of PV generation in 3 consecutive days.

UBM demonstrated highly reliable probabilistic forecasting across sectors. However, it fails to provide a sharp PV generation forecast. Notably, low sharp probabilistic forecasts were observed mostly during the PV peak hours. This is mainly due to the high variability of PV generation itself, which consequently leads to high forecast errors generated by the PSLP model. This, in turn, results in a wider spread of point forecast errors during the peak PV generation hours, as depicted in Figure 7c. This corresponds to the wider error ECDFs for the clusters in Phase II of the UBM model. Consequently, this leads to wider interval widths (i.e., low sharpness)

Table 7: Comparison of UBM with other probabilistic forecasting methods that also leverage deterministic models.

Method	Limitations	Strengths	Sources
QRA	• Requires point forecasts from multiple deterministic models • High computational complexity • High model complexity • Black box QR model	• Extensively covered in the literature • Generates probabilistic forecasting with good reliability and sharpness • Diverse applications, including electricity prices, electricity load demand, PV, and wind generation • Nonparametric	[14–16]
QRF	• Uses point forecast from multiple deterministic models [3] • Medium-to-high computational complexity • High model complexity • Black box model	• Extensively covered in the literature • Found to be more accurate and computationally efficient compared to QRA [3] • Diverse application • Nonparametric	[3, 22, 23]
EPIs	• Prediction intervals are not conditional • Lacks adaptability • Moderately complex	• Single deterministic model is sufficient to construct the prediction interval • Have been applied in various fields, including meteorology, economics, and energy • Can use nonparametric distribution of forecast error • Low computational complexity • Simplistic approach and easy to implement	[26–28]
AnEn	• Along with the forecasted values, it needs meteorological predictor variables (conditionals) • Less adaptive to sector-wise and diverse application • Less accurate • Highly sensitive to the criteria used to define the similarity between current situation and historical analogs • In several lead times, the number of ensemble observation values could be limited to get the proper distribution forecast	• Single deterministic model is sufficient • Robust approach • Low-to-medium computational complexity • Nonparametric assumptions can be considered	[33, 34]
UBM	• Found to be generating low sharpness probabilistic forecast for highly variable entities and low accuracy of the point forecast model (e.g., PV generation) • Newly developed and has to be explored with further research	• Only uses single-point forecast model • Only uses point forecast as conditional • Highly transparent approach (non-black box) • Low model complexity • Robust model and has the ability to capture uncertainty accurately • Easy to implement • Moderately accurate results • Extremely low computational • Nonparametric distribution approach • Can be used in diverse application	This study

Abbreviations: AnEn, analog ensemble; EPIs, empirical prediction intervals; QR, quantile regression; QRA, quantile regression averaging; QRF, quantile random forest; UBM, uncertainty binning method.

in the generated probabilistic forecast in Phase III, as shown, for example, in Figure 8c.

Low sharpness in probabilistic forecasting can pose challenges for effective decision-making when compared to real behavior. Decision-makers rely on forecasts not only to assess reliability but also to ensure the intervals are sufficiently narrow to facilitate actionable insights. Forecasts with low sharpness imply broad uncertainty bands, which may be interpreted as a lack of confidence in the forecast's precision. In practical terms, this can lead to overly conservative decisions, such as over-provisioning resources or failing to optimize the deployment of assets like energy storage systems. For instance, in energy management, a forecast with wide PIs might hinder effective grid balancing or delay response strategies, thus

impacting the cost-effectiveness and efficiency of operations. Nevertheless, even when a deterministic model demonstrates significant forecasting errors, as in the case of PV, probabilistic forecasting emerges as highly valuable from a decision-making perspective. It provides critical insights into the degree of uncertainty associated with the model's forecasted outputs. Depending solely on inadequate point forecasts can pose substantial challenges across a multitude of applications within a power system. Therefore, the UBM is found to be a highly valuable tool, capable of converting inaccurate point forecasts into reliable probabilistic ones and providing richer information about uncertainty for decision-making in energy management.

In future work, the implementation of more accurate point forecast models for PV generation could be explored to achieve more precise probabilistic forecasting results from the UBM, however, this is out of scope for this work. Since the UBM's performance relies heavily on the accuracy of the underlying deterministic model, any limitations or inaccuracies in the point forecast can directly impact the quality of the probabilistic forecast. Therefore, improving the base deterministic model for PV generation is essential to enhance the sharpness and overall reliability of the UBM's output. By integrating advanced forecasting models that better capture the high variability and complex patterns of PV generation, such as hybrid machine learning approaches or enhanced statistical models, the UBM could potentially generate tighter PIs and reduce the spread of errors observed during peak generation periods. This would ultimately lead to more actionable and trustworthy forecasts, facilitating more effective decision-making and resource management within power systems.

Overall, a correlation between the accuracy of the deterministic model (PSLP) and the UBM model was also observed. As the point forecasts and its errors are used as training data for the UBM, their impact on the model's performance becomes evident. A higher variability in the data profile and lower accuracy of the deterministic model results in low sharpness probabilistic predictions. Typically, the UBM compromises sharpness by prioritizing the PICP to be closer to or higher than the PI. In order to obtain a better PICP, the interval width must simultaneously increase. On the other hand, reducing the PI can improve the sharpness, but this does not necessarily enhance the overall accuracy of the probabilistic forecasts. But, the UBM at all time tries to achieve reliable forecasts, even if sharpness is compromised. This investigation sheds light on how the performance of the deterministic model influences the sharpness and reliability of the probabilistic forecasts produced by the UBM. Ultimately, this insight supports informed decision-making in EMS.

UBM was also evaluated with different PI levels. PICP values were found to be aligned with PI levels being considered for all the sectors, except for PV, with PI of 40% and 60%. It is usually expected to have a PICP closer to PI; therefore, overestimating the PICP beyond the expected PI level indicates that the model is overestimating the level of uncertainty or variability in the PV generation, leading to wider intervals that cover more of the actual data points than anticipated. This over-coverage can be problematic in energy management because it may lead to inefficient decision-making, such as overestimating the reserve capacity needed or over-allocating resources to account for higher-than-expected variability. With the increase in PI levels, the interval width also increased as expected, ensuring that the forecasted intervals covered the real values within the quantile range.

Moreover, the UBM implemented the LBW feature, which basically limits and slides the training window as the model iterates over the time. It was observed that reducing the training window decreases sharpness (MPIW and WS) but has a slight effect on reliability (PICP). Overall, it was observed that with shorter training window results in better performance of the UBM. Across all given data profiles, the effect of changing the training window on the performance of the UBM was not significant. However, even a slight improvement in forecasting results counts. Slightly higher differences were observed in electricity and heat demand forecasts based on the QCS and PQCS. Although the effect on the data profiles in this study was not substantial, it could be more significant for other data profiles. Therefore, testing the LBW feature to assess its impact on the UBM, depending on the data profile, can be useful for improving the accuracy of the UBM. As a guideline for selecting an appropriate training window, observations from this study suggest the following: for data profiles with pronounced seasonal variations, such as PV generation and heat demand, a training window of 30–90 days is recommended. This range ensures that the training data reflects season-specific characteristics, as using historical data from a season like summer to construct forecast distributions for winter could lead to reduced sharpness. Conversely, electricity demand profiles may accommodate a larger training window (e.g., over 90 days), but this depends on the nature of the connected load. In the future, as the number of heat pumps increases, winter demand is expected to grow relative to summer, amplifying seasonal differences. In such scenarios, a shorter training window (e.g., 30–90 days) would again be advisable. However, it is important to note that shorter training windows result in fewer data points within each cluster, potentially hindering the generation of accurate error distributions. Overall, the selection of the training window should be tailored to the specific characteristics of the case at hand.

Table 7 compares UBM with several known models or approaches that leverage the deterministic model to generate probabilistic forecasts. The GCM, QCS, and PQCS overcome some of the shortcomings of popular metrics. The GCM offers a graphical evaluation, which is not commonly found in the literature on probabilistic forecasting evaluation. PIT is a similar technique as GCM that has been described in other literature [30, 31], but it lacks numerical representation, which is addressed by QCS and PQCS in this study. The QCS and PQCS are independent of the data scale, in contrast to MPIW and WS. This supports the comparison of model's performance based on different datasets. Furthermore, QCS and PQCS are not biased toward sharpness over reliability, as in the case with WS. Unlike PICP, which only measures the

reliability within the PI, QCS, and PQCS assess reliability with consideration to each quantile bin and check its deviation from expected reliability. That is, they penalize any over and underestimation of reliability. However, these metrics cannot be applied to PV, as the predictions at various percentiles align with the measured values, particularly during periods of zero power generation, leading to significantly inflated observed frequency (O_i) for each bin. This results in disproportionately high QCS and PQCS due to the penalization for deviations from expected frequency (E_i). An alternative approach could involve excluding such occurrences when calculating QCS and PQCS, enabling their use in evaluating probabilistic forecasting of PV generation. In future work, the adaptation of QCS and PQCS for PV generation should be prioritized to provide a fair evaluation framework. Developing modified versions of these metrics or applying conditional calculations would enhance their applicability. This would support more nuanced insights into forecasting performance and better guide decision-making. There is no single metric that can evaluate all aspects of probabilistic forecasting without limitations. Therefore, in addition to popular metrics (PICP, MPIW, and WS), GCM and its numerical scores (QCS and PQCS) can provide a more comprehensive and intuitive performance assessment.

There are numerous possibilities for enhancing the accuracy of the UBM, primarily by considering more precise deterministic models, especially in the case of PV forecasting. Besides the deterministic methods, alternative binning approaches can be explored for the UBM training. This work uses the Python "cut" function [55], which requires to predefined the number of clusters and discretizes the data with equal-width bins. An alternative approach is the Python "qcut" function [56], which discretizes arrays into equally sized bins. This enables the function's algorithm to generate clusters based on the data profile itself while maintaining an equal number of data points in each bin. Such flexibility in generating clusters should be data-driven rather than being preset to predefined clusters. There are other popular automatic clustering algorithms, such as K-means and fuzzy C-means clustering [57–59] that could be implemented to train the UBM. Furthermore, hyperparameter optimization could enhance the UBMs performance further.

6. Conclusions and Outlook

This work introduces an approach to generating probabilistic forecasting by extending a deterministic model. Given the increasing significance of uncertainty quantification in the energy domain, the proposed UBM emerges as a valuable tool for decision-making in EMS. It is shown to be a transparent and efficient method that leverages deterministic models while offering simplicity and high computational speed, harnesses readily available data, requires minimal training data, involves minimal feature engineering, offers rapid computational capabilities, and achieves reasonable accuracy. This approach caters to the evolving landscape of EMS requirements, particularly in smaller-scale settings like buildings and mid-sized facilities, where efficiency and affordability are the main factors. The UBM was rigorously validated for forecasting electricity demand, heat demand, and PV generation. A practical case study was conducted using an existing distributed integrated local energy system at a logistics facility in northern Germany.

The statistical PSLP method was implemented to generate point forecasts, which were subsequently used as input features for the UBM. For its evaluation, MAE, MSE, RMSE, MAPE, and MASE were considered. The PSLP model demonstrated good performance in forecasting the electricity and heat demand but exhibited limitations in accurately predicting PV generation. As UBM uses point forecasts from the PSLP model as input features for generating probabilistic forecasts, the accuracy of the PSLP impacted the UBMs overall accuracy. This impact was particularly noticeable in the interval width, as evident through MPIW and WS. In the given case study, sensitivity analysis was carried out to determine the optimal number of bins on the performance of the UBM, which resulted in the selection of 7, 7, and 12 bins for electricity, heat, and PV, respectively. Following the selection of optimal bins, it was observed that the UBM achieved notably good reliability across all sectors, with PICP values of 79.52%, 81.45%, and 81.37 % for electricity, heat, and PV, respectively. With respect to sharpness, the UBM showed better performance on electricity and heat demand over PV generation. A high MPIW (32.54 kW) and WS (48.58) were observed for PV generation forecasts, which can be attributed to the low accuracy of the PSLP model due to the high variability of the data profile itself. Hence, it was observed that the accuracy of the point forecast model directly impacts that of the probabilistic forecasts produced by the UBM. However, it has the capability to transform inaccurate deterministic forecasts into reliable probabilistic ones, providing richer information on uncertainty and, consequently, supporting decision-making in EMS. To address the limitations of popular evaluation scores, the GCM, QCS, and PQCS were implemented, providing a more comprehensive performance evaluation of probabilistic forecasting. Both the QCS and PQCS scores were found to be better for electricity, with values of 20.99% and 20.95%, respectively, compared to heat, with the QCS value of 23.31 and PQCS value of 22.53 %. However, the scores were found to be unsuitable for evaluating PV forecasts. The UBMs performance was also evaluated at different PI levels (40%, 60%, and 80%). The PICP was found to be closely aligned with PI levels for each sector, except for PV, with 40% and 60% PI. The interval width was found to be increased with an increase in PI levels, as expected, resulting in higher MPIW and WS. Moreover, performance evaluations were conducted with varying sliding training windows. It was observed that the reduction in the training window yielded an improvement in the probabilistic forecast scores except PICP.

Future work should focus on testing the UBM with more accurate deterministic models, especially for PV. Moreover, K-means clustering, fuzzy C-means clustering, qcut, or other more advanced clustering techniques should be explored for training the UBM. A detailed comparison of the UBM with other available probabilistic forecasting techniques should be

investigated in further work. The adaptation of the GCM, QCS, and PQCS for PV probabilistic forecasts is another avenue to be considered. Further detailed examination is required on how point forecast errors propagate to the accuracy of the UBM. In the future, the real application of the UBM for operational optimization can be tested, for example, by optimizing EV charging schedules. Although the potential applications of the UBM within the distributed integrated local energy systems are abundant, these possibilities remain avenues for exploration in future research.

Nomenclature

AnEn: Analog ensemble
BEVs: Battery electric vehicles
ECDF: Empirical cumulative distribution function
ELogZ: Energieversorgungskonzepte für Klimaneutrale Logistikzentren
EMS: Energy management system
EPI: Empirical prediction interval
GCM: Graphical calibration measure
IQR: Interquartile range
KDE: Kernel density estimation
LBW: Look back window
MAE: Mean absolute error
MAPE: Mean absolute percentage error
MASE: Mean absolute square error
MPIW: Mean prediction interval width
MSE: Mean square error
NRM: Nearest rank method
NWDs: Number of waiting days
PDF: Probability density function
PI: Prediction interval
PICP: Prediction interval coverage probability
PIT: Probability integral transform
PQCS: Percentage quantile calibration score
PSLP: Personalized standard load profile
PV: Photovoltaic
QCS: Quantile calibration score
QR: Quantile regression
QRA: Quantile regression averaging
QRF: Quantile random forest
RMSE: Root mean square error
SLP: Standard load profile
UBM: Uncertainty binning method
WS: Winkler score.

Data Availability Statement

The authors do not have institutional permission to share data or codes.

Disclosure

Responsibility for the content of this publication lies with the author. More information regarding the ELogZ can be found at www.elogz.de.

Conflicts of Interest

The authors declare conflicts of interest.

Funding

The project Energieversorgungskonzepte für klimaneutrale Logistikzentren (ELogZ) on which this article is based was funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK) under the funding code 03EN1015F.

Acknowledgments

The authors gratefully acknowledge PANEUROPA Transport GmbH for the provision of time series data, pictures, and knowledge of its logistics facility under the project "Energieversorgungskonzepte für klimaneutrale Logistikzentren (ELogZ)."

References

[1] J. Xie, T. Hong, T. Laing, and C. Kang, "On Normality Assumption in Residual Simulation for Probabilistic Load Forecasting," *IEEE Transactions on Smart Grid* 8, no. 3 (2017): 1046–1053.
[2] C. Voyant, G. Notton, S. Kalogirou, et al., "Machine Learning Methods for Solar Radiation Forecasting: A Review," *Renewable Energy* 105 (2017): 569–582.
[3] W. Zhang, H. Quan, and D. Srinivasan, "Parallel and Reliable Probabilistic Load Forecasting via Quantile Regression Forest and Quantile Determination," *Energy* 160 (2018): 810–819.
[4] C. Kang, Y. Wang, Y. Xue, G. Mu, and R. Liao, "Big Data Analytics in China's Electric Power Industry: Modern Information, Communication Technologies, and Millions of Smart Meters," *IEEE Power and Energy Magazine* 16, no. 3 (2018): 54–65.
[5] F. Mei, J. Gu, J. Lu, et al., "Day-Ahead Nonparametric Probabilistic Forecasting of Photovoltaic Power Generation Based on the LSTM-QRA Ensemble Model," *IEEE Access* 8 (2020): 166138–166149.
[6] G. I. Nagy, G. Barta, S. Kazi, G. Borbély, and G. Simon, "GEFCom2014: Probabilistic Solar and Wind Power Forecasting Using a Generalized Additive Tree Ensemble Approach," *International Journal of Forecasting* 32, no. 3 (2016): 1087–1093.
[7] Y. Wang, N. Zhang, Y. Tan, et al., "Combining Probabilistic Load Forecasts," *IEEE Transactions on Smart Grid* 10, no. 4 (2019): 3664–3674.
[8] T. Hong and S. Fan, "Probabilistic Electric Load Forecasting: A Tutorial Review," *International Journal of Forecasting* 32, no. 3 (2016): 914–938.
[9] T. Hong, P. Pinson, and S. Fan, "Global Energy Forecasting Competition 2012," *International Journal of Forecasting* 30, no. 2 (2014): 357–363.
[10] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 and Beyond," *International Journal of Forecasting* 32, no. 3 (2016): 896–913.
[11] E. Lucas Segarra, G. Ramos Ruiz, and C. Fernández Bandera, "Probabilistic Load Forecasting for Building Energy Models," *Sensors* 20, no. 22 (2020): 6525.

[12] O. Grothe, F. Kächele, and F. Krüger, "From Point Forecasts to Multivariate Probabilistic Forecasts: The Schaake Shuffle for Day- Ahead Electricity Price Forecasting," *Energy Economics* 120 (2023): 106602.
[13] Meer V. dDW, J. Widén, and J. Munkhammar, "Review on Probabilistic Forecasting of Photovoltaic Power Production and Electricity Consumption," *Renewable and Sustainable Energy Reviews* 81 (2018): 1484–1512.
[14] B. Uniejewski, R. Weron, and F. Ziel, "Variance Stabilizing Transformations for Electricity Spot Price Forecasting," *IEEE Transactions on Power Systems* 33, no. 2 (2018): 2219–2229.
[15] B. Uniejewski, "Smoothing Quantile Regression Averaging: A New Approach to Probabilistic Forecasting of Electricity Prices," arXiv preprint arXiv: 2302.00411, 2023
[16] R. Weron, "Electricity Price Forecasting: A Review of the State-of-the-Art With a Look into the Future," *International Journal of Forecasting* 30, no. 4 (2014): 1030–1081.
[17] B. Liu, J. Nowotarski, T. Hong, and R. Weron, "Probabilistic Load Forecasting via Quantile Regression Averaging on Sister Forecasts," *IEEE Transactions on Smart Grid* 8, no. 2 (2015): 730–737.
[18] Y. Zhang, K. Liu, L. Qin, and X. An, "Deterministic and Probabilistic Interval Prediction for Short-Term Wind Power Generation Based on Variational Mode Decomposition and Machine Learning Methods," *Energy Conversion and Management* 112 (2016): 208–219.
[19] L. Zhang, S. Lu, Y. Ding, et al., "Probability Prediction of Short-Term User-Level Load Based on Random Forest and Kernel Density Estimation," *Energy Reports* 8 (2022): 1130–1138.
[20] N. Zhang, C. Kang, Q. Xia, and J. Liang, "Modeling Conditional Forecast Error for Wind Power in Generation Scheduling," *IEEE Transactions on Power Systems* 29, no. 3 (2014): 1316–1324.
[21] S. Dang, L. Peng, J. Zhao, J. Li, and Z. Kong, "A Quantile Regression Random Forest-Based Short-Term Load Probabilistic Forecasting Method," *Energies* 15, no. 2 (2022): 663.
[22] H. Aprillia, H.-T. Yang, and C.-M. Huang, "Statistical Load Forecasting Using Optimal Quantile Regression Random Forest and Risk Assessment Index," *IEEE Transactions on Smart Grid* 12, no. 2 (2021): 1467–1480.
[23] E. Freeman and G. Moisen, *An Application of Quantile Random Forests for Predictive Mapping of Forest Attributes* (U. S. Department of Agriculture, Forest Service, 2015).
[24] K. Vaysse and P. Lagacherie, "Using Quantile Regression Forest to Estimate Uncertainty of Digital Soil Mapping Products," *Geoderma* 291 (2017): 55–64.
[25] S. P. Vasseur and J. L. Aznarte, "Comparing Quantile Regression Methods for Probabilistic Forecasting of NO2 Pollution Levels," *Scientific Reports* 11, no. 1 (2021): 11592.
[26] W. H. Williams and M. L. Goodman, "A Simple Method for the Construction of Empirical Confidence Limits for Economic Forecasts," *Journal of the American Statistical Association* 66, no. 336 (1971): 752–754.
[27] Y. S. Lee and S. Scholtes, "Empirical Prediction Intervals Revisited," *International Journal of Forecasting* 30, no. 2 (2014): 217–234.
[28] "National Hurricane Center Forecast Verification," NOAA National Hurricane Center, 2016, https://www.nhc.noaa.gov/verification/verify6.shtml.
[29] E. Britton, P. Fisher, and J. Whitley, *The Inflation Report Projections: Understanding the Fan Chart* (The Bank of England, England, tech. rep, 1998).
[30] L. H. Kaack, J. Apt, M. G. Morgan, and P. McSharry, "Emperical Prediction Intervals Improve Energy Forecasting," *PNAS* 114 (2017).
[31] H. Huang, R. Jia, J. Liang, J. Dang, and Z. Wang, "Wind Power Deterministic Prediction and Uncertainty Quantification Based on Interval Estimation," *Journal of Solar Energy Engineering* 143, no. 6 (2021): 061010.
[32] M. Saber, *Quantifying Forecast Uncertainty in the Energy Domain*, (PhD thesis, (Marquette University, Marquette, USA, 2017).
[33] S. Alessandrini, L. Delle Monache, S. Sperati, and J. N. Nissen, "A Novel Application of an Analog Ensemble for Short-Term Wind Power Forecasting," *Renewable Energy* 76 (2015): 768–781.
[34] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, "An Analog Ensemble for Short-Term Probabilistic Solar Power Forecast," *Applied Energy* 157 (2015): 95–110.
[35] C. Sigauke, M. M. Nemukula, and D. Maposa, "Probabilistic Hourly Load Forecasting Using Additive Quantile Regression Models," *Energies* 11, no. 9 (2018): 2208.
[36] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Lower Upper Bound Estimation Method for Construction of Neural Network- Based Prediction Intervals," *IEEE Transactions on Neural Networks* 22, no. 3 (2011): 337–346.
[37] A. Harvey and G. Sucarrat, "Evaluating Density Forecasts With Applications to Financial Risk Management," *Computational Statistics & Data Analysis* 76 (2014): 320–338.
[38] M. B. Bjerregård, J. K. Møller, and H. Madsen, "An Introduction to Multivariate Probabilistic Forecast Evaluation," *Energy and AI* 4 (2021): 100058.
[39] M. Hinterstocker, Roon vS, and M. Rau, *Bewertung der Aktuellen Standardlastprofile Österreichs Und Analyse Zukünfnftiger Anpas- Sungsmöglichkeiten Im Strommarkt* (Symposium Energieinnovation, *[Evaluation of Austria's Current Standard Load Profiles and Analysis of Future Adjustment Options in the Electricity Market]*, 2014).
[40] J. S. Telle, N. Maitanova, T. Steens, B. Hanke, Maydell vK, and M. Grottke, *Combined PV Power and Load Prediction for Building-Level Energy Management Applications* (IEEE, 2020): 1–15.
[41] T. Steens, J. S. Telle, B. Hanke, et al., "A Forecast-Based Load Management Approach for Commercial Buildings Demonstrated on an Integration of BEV," *Energies* 14, no. 12 (2021): 3576.
[42] J. S. Telle, A. Upadhaya, P. Schönfeldt, T. Steens, B. Hanke, and K. von Maydell, "Probabilistic Net Load Forecasting Frame-Work for Application in Distributed Integrated Renewable Energy Systems," *Energy Reports* 11 (2024): 2535–2553.
[43] B. Du, S. Huang, J. Guo, H. Tang, L. Wang, and S. Zhou, "Interval Forecasting for Urban Water Demand Using PSO Optimized KDE Distribution and LSTM Neural Networks," *Applied Soft Computing* 122 (2022): 108875.
[44] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization* (John Wiley & Sons, 1992).
[45] J. M. González-Sopeña, V. Pakrashi, and B. Ghosh, "An Overview of Performance Evaluation Metrics for Short-Term Statistical Wind Power Forecasting," *Renewable and Sustainable Energy Reviews* 138 (2021): 110515.
[46] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F. J. Martinez-de-Pison, and F. Antonanzas-Torres, "Review of Photovoltaic Power Forecasting," *Solar Energy* 136 (2016): 78–111.
[47] R. J. Hyndman and A. B. Koehler, "Another Look at Measures of Forecast Accuracy," *International Journal of Forecasting* 22, no. 4 (2006): 679–688.

[48] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," (OTexts (2018).
[49] X. Sun, Z. Wang, and J. Hu, "Prediction Interval Construction for Byproduct Gas Flow Forecasting Using Optimized Twin Extreme Learning Machine," *Mathematical Problems in Engineering* 2017, no. 1 (2017): 5120704, 12.
[50] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association* 102, no. 477 (2007): 359–378.
[51] B. Steden, J. S. Telle, N. Wollek, S. Schlüters, and J. Marx-Gómez, "Distributed Sector Coupled Low Carbon Energy Supply for Logistics Properties-Concept for the Integration of Heating, Electricity and Transport," *EnviroInfo* (2021).
[52] Energieversorgungskonzepte Für Klimaneutrale Logistikzentren, "Energy Supply Concepts for Climate-Neutral Logistics Centres," : 2019–2024, https://elogz.de/.
[53] W. F. Holmgren, C. W. Hansen, and M. A. Mikofski, "Pvlib Python: A Python Package for Modeling Solar Energy Systems," *Journal of Open Source Software* 3, no. 29 (2018): 884.
[54] Deutscher Wetterdienst - Open Data, "WebPage";-Free Provision of Spatial Data of the DWD via the DWD's Open Data Server," (2023).
[55] Pandas, *Python Cut Function From Pandas Library* (Webpage, 2023).
[56] Pandass, *Python Qcut Function from Pandas Library* (Webpage, 2023).
[57] A. Ashabi, S. B. Sahibuddin, and M. Salkhordeh Haghighi, "The Systematic Review of K-Means Clustering Algorithm," in *International Conference on Networks, Communication and Computing* (2020): 13–18
[58] S. Askari, "Fuzzy C-Means Clustering Algorithm for Data With Unequal Cluster Sizes and Contaminated With Noise and Outliers: Review and Development," *Expert Systems with Applications* 165 (2021): 113856.
[59] Y. Sinambela, S. Herman, A. Takwim, and S. R. Widianto, "A Study of Comparing Conceptual and Performance of K-Means and Fuzzy C-Means Algorithm (clustering Method of Data Mining) of Consumer Segmentation," *Jurnal Riset Informatika* 2, no. 2 (2020): 49–54.