Tailoring the Tails: Enhancing the Reliability of Probabilistic Load Forecasts

Roberto Baviera ROBERTO.BAVIERA@POLIMI.IT
Pietro Manzoni[*] PIETRO.MANZONI@POLIMI.IT
Politecnico di Milano, Department of Mathematics, Milan, Italy

Editor: Simone Vantini, Matteo Fontana, Aldo Solari, Henrik Boström and Lars Carlsson

Abstract

Quantifying predictive uncertainty regarding future electricity demand is the main goal of probabilistic load forecasting. A good probabilistic model is often identified with forecasted densities that are as concentrated ("sharp") as possible. However, this goal is frequently achieved by sacrificing forecast reliability, i.e. the statistical compatibility between forecasted densities and observed frequencies. In real-world applications, reliability is the crucial measure of model quality, especially when predicting distribution tails. We propose a new methodology for probabilistic load forecasting, introducing a novel loss function which allows an excellent balance between forecast sharpness and reliability. We apply the proposed modelling approach for predicting the electricity load on a benchmark dataset. Experimental results show that the obtained density forecasts are extremely reliable and also close to optimal in terms of sharpness and point accuracy.

Keywords: Probabilistic Load Forecasting, Neural Networks, Overconfidence

1. Introduction

Load forecasting is an essential task in modern power systems. Being able to generate accurate forecasts for future load is extremely important for several reasons, such as meeting technical constraints, achieving operational excellence and increasing financial opportunities both for system operators and end-users (see, e.g., Hong and Fan, 2016). In the last decade, probabilistic load forecasting has gained an ever-increasing interest among researchers and practitioners, by allowing management of the intrinsic uncertainty that characterises the dynamics of electricity demand.

In recent years, the explosion of Machine Learning (ML) techniques has significantly impacted probabilistic load forecasting. ML models have proven to be highly effective forecasting tools, achieving excellent predictive performance (see, e.g., Yang et al., 2018; Smyl and Hua, 2019; Baviera and Messuti, 2023). In particular, Artificial Neural Networks (ANNs) have been largely employed in probabilistic load forecasting because of their ability to capture and model the nonlinear dependencies observed in the electricity load behaviour (see, e.g., Vossen et al., 2018; Azzone and Baviera, 2021). However, it is known that ML models tend to be overconfident, meaning they frequently overestimate the likelihood of probable events, and underestimate the true predictive uncertainty (see, e.g., Guo et al., 2017). This behaviour is particularly risky in safety-critical disciplines such as load forecasting. Nowadays, utilities and system operators prioritise probabilistic forecasts being

[*] Corresponding author

reliable rather than merely accurate, due to the rapid evolution of power systems (see, e.g., Hong and Fan, 2016).

In this study, we focus on the issue of overconfidence in probabilistic load forecasting using Artificial Neural Networks. To our knowledge, overconfidence mitigation has only been explored in ANN literature within the domain of classification problems (see, e.g., Guo et al., 2017; Kristiadi et al., 2020; Wang et al., 2021). It has been proven that some ANNs used for classification tasks are always overconfident "far away from the data" (Kristiadi et al., 2020). Moreover, as pointed out by Wei et al. (2022), ANN-based classifiers often struggle with out-of-distribution (OOD) inputs, i.e. inputs sampled from a distribution different from that encountered during training. It is important to note that unlike image and document classification datasets, where data can be extensive, the time series data used in probabilistic load forecasting typically consists of a single realisation from the past. This characteristic makes forecasting models more vulnerable to OOD examples and, consequently, to overconfidence.

In point forecasting, accuracy is the most relevant objective for practical applications: the goal is to generate a point forecast $\hat{y}_t$ as close as possible to the realised value y_t. In probabilistic forecasting of time series, *reliability* – the frequency with which the realised value lies in the forecasted α-Prediction Interval (PI) – is the target quantity, and it is measured via the Empirical Coverage (see, e.g., Pinson et al., 2007). In particular, a prediction is considered overconfident if the forecasted distribution is too *sharp* around the point forecast. In this paper, we indicate a simple criterion for detecting overconfidence: a forecasting model is overconfident if the Empirical Coverage is smaller than α for every relevant confidence level α.

Mitigating the overconfidence that arises using ANN-based forecasters requires additional design thinking that is often overlooked in the ML literature. In many cases, overconfidence appears even to be incentivised, as maximising the *sharpness* of predictions is often perceived as the primary goal for a forecaster. In this paper, we design a new "robustified" loss function that, on the one hand, preserves accuracy, and on the other hand, increases reliability: mitigating overconfidence corresponds to enhancing reliability without affecting accuracy. We propose a quantitative approach, modifying a standard loss function by introducing an additional parameter responsible for controlling overconfidence.

Three are the main contributions of this study. First, we introduce a novel loss function for probabilistic forecasting, derived from an existing score, the Pinball Loss. Second, we present an application of this new loss function in probabilistic load forecasting with ANNs. With the proposed technique, we show that it is possible to increase the reliability of the predicted distributions and to enhance the modelling of their tails – a crucial aspect for managing operational and financial risks. Finally, we show that the increase in forecast reliability comes at no expense of forecast accuracy.

The rest of the paper is organised as follows. In Section 2, we describe the standard evaluation measures employed in probabilistic forecasting. In Section 3, we introduce the new loss function and discuss its role in mitigating overconfidence in predictive models. In Section 4, we present the experimental analysis, focusing on the applications of the proposed methodology in probabilistic load forecasting. Finally, Section 5 provides concluding remarks.

2. Evaluation of probabilistic forecasts: the standard approach

While measuring the accuracy of point forecasts is generally straightforward, evaluating the quality of probabilistic forecasts is a more complex task. Specific scores are employed for this purpose: depending on the output issued by the forecasting model, these scores are designed to evaluate either the quality of a forecasted quantile, or the quality of a forecasted probability distribution; in the literature, the expression *scoring function* is used to indicate scores of the former type, and the expression *scoring rule* to indicate scores of the latter type (see, e.g., Gneiting et al., 2007).
In this section, we discuss the most relevant scoring functions and scoring rules used in probabilistic forecasting.

2.1. From Pinball Loss to CRPS

According to the literature, the Pinball Loss is the fundamental scoring function used in probabilistic forecasting of univariate time series (see, e.g., Gneiting et al., 2007; Hong and Fan, 2016). It serves to measure the quality of forecasted quantiles and it is defined as follows: given a random variable Y and a level $\alpha \in (0,1)$,

$$\mathcal{P}(q, y\,; \alpha) := \begin{cases} \alpha(y-q) & \text{if } y \geq q\,, \\ (1-\alpha)(q-y) & \text{if } y < q\,, \end{cases} \tag{1}$$

where q is the forecast for the α-quantile of Y, and y is the corresponding realisation. This scoring function is well-known in the probabilistic forecasting literature and it is commonly selected as loss function in Quantile Regression (see, e.g., Könker, 2005). In particular, given an integrable random variable Y with CDF F, it holds that

$$\mathbb{E}\left[\mathcal{P}\left(F^{-1}(\alpha),\, Y\,;\, \alpha\right)\right] \;\leq\; \mathbb{E}\left[\mathcal{P}\left(q,\, Y\,;\, \alpha\right)\right] \qquad \forall q \in \mathbb{R}\,,$$

a property known as (strict) consistency (see, e.g., Fissler et al., 2023).

In forecasting practice, probabilistic forecasts are commonly provided in the form of PIs, rather than quantiles. The Pinball Loss can be extended to serve as a scoring function for PIs. We define a new scoring function $\mathcal{P}_{\mathrm{C}}$, which we call the *central* Pinball Loss: given a central α-PI $[l,u]$, we write

$$\mathcal{P}_{\mathrm{C}}(l, u, y;\, \alpha) := \mathcal{P}\left(l,\, y\,;\, \frac{1-\alpha}{2}\right) + \mathcal{P}\left(u,\, y\,;\, \frac{1+\alpha}{2}\right)\,. \tag{2}$$

The central Pinball Loss is obtained as the sum of the Pinball Loss of the quantiles at level $\frac{1-\alpha}{2}$ and $\frac{1+\alpha}{2}$, the two quantiles which characterise the central α-PI. Even this scoring function is consistent with respect to the class of integrable random variables – because each of the two terms in (2) is –, making it suitable to be a natural loss function for PIs.

When evaluating the quality of a probabilistic forecasting model, one considers the sample Pinball Loss of the forecasted α-quantiles of the time span $\{1,\ldots,T\}$. Given a set of forecasts for the α-quantile, the sample Pinball Loss is computed as

$$\mathrm{PL}_{\alpha} := \frac{1}{T}\sum_{t=1}^{T} \mathcal{P}(q_{t,\alpha},\, y_t\,;\, \alpha)\,, \tag{3}$$

being y_t the realisation at time t and $q_{t,\alpha}$ the corresponding forecast for the α-quantile. The computation of this quantity requires the selection of a specific confidence level α: in the load forecasting literature, the performances of different forecasters are typically evaluated considering just a few relevant values for α (for instance, 50% and 90%; see, e.g., Liu et al., 2015; Yang et al., 2019).

In recent years, it has also become common to evaluate probabilistic forecasts in terms of Average Pinball Loss (APL), a score obtained as the mean of Pinball Losses over the percentiles, i.e.

$$\mathrm{APL} := \frac{1}{99}\sum_{k=1}^{99} \mathrm{PL}_{\frac{k}{100}} \; . \tag{4}$$

APL gained momentum since the last editions of the GEFCom, where it was employed to rank predictive models and to assess probabilistic forecasts (Hong et al., 2016, 2019). APL can also be chosen as a loss function for models that output multiple quantiles, or entire probability distributions. Its employment for model training – and not just for performance evaluation – is quite recent in the literature; for instance, it has been considered by Wang et al. (2019) for training an LSTM network which predicts simultaneously all percentiles of the inferred density.

Many safety-critical applications, such as those in the energy sector, require forecasts in the form of continuous densities (see, e.g., Gilbert et al., 2023). In this case, forecasting models are designed to predict not quantiles or percentiles, but a vector of parameters $\underline{\theta}$ that completely characterises the CDF $F(x) = F(x; \underline{\theta})$ of the forecasted distribution. An example of a probabilistic forecaster of this kind is a model that predicts the mean and the standard deviation of a Gaussian density (see, e.g., Azzone and Baviera, 2021; Marcjasz et al., 2023).

To train these forecasting models, one can use as loss function the Continuous Ranked Probability Score (CRPS; cf. Matheson and Winkler, 1976; Berrisch and Ziel, 2023), which is defined as

$$\mathrm{CRPS}(F, y) := \int_{-\infty}^{+\infty} \left(F(x) - \mathbb{1}_{x \geq y}\right)^2 \mathrm{d}x = \tag{5}$$

$$= 2\int_0^1 \mathcal{P}\left(F^{-1}(\alpha), y\,;\,\alpha\right) \mathrm{d}\alpha \; . \tag{6}$$

As highlighted by the expression (6), this scoring rule represents the continuous counterpart of the APL, with the arithmetic average replaced by the integral average. The forecast q for each α-quantile – the first argument of the Pinball Loss in (1) – is selected as the α-quantile[1] of the CDF F. Moreover, with a change of variable, the CRPS can also be expressed in terms of the central Pinball Loss in (2), namely as

$$\mathrm{CRPS}(F, y) = \int_0^1 \mathcal{P}_{\mathrm{C}}\left(F^{-1}\left(\frac{1-\alpha}{2}\right), F^{-1}\left(\frac{1+\alpha}{2}\right), y\,;\,\alpha\right) \mathrm{d}\alpha \; . \tag{7}$$

[1] In the following, we consider a strictly monotone CDF F, so that any α-quantile is given by $F^{-1}(\alpha)$, i.e. by CDF inversion. Nevertheless, all results can be generalised for non-invertible CDFs.

In the energy forecasting literature, CRPS has mainly been employed as an evaluation metric, but recently it has also been adopted as a loss function for model training (see, e.g., Li et al., 2019; Wang et al., 2022).

The strength of CRPS is that it evaluates a whole probability distribution, providing an overall index of forecast quality. It is however important to underline that CRPS alone is not enough to measure reliability, and in particular to deal with overconfident forecasts, as we discuss in Section 3.

2.2. The Gaussian CRPS

ANNs are trained on data using iterative algorithms, making it crucial for loss functions to be easily computable. In the case of CRPS, calculating the integral in (5) quickly and accurately poses some relevant challenges. In the following proposition, we prove that this scoring rule admits a closed-form expression if F – the distribution we want to calibrate or forecast – is Gaussian. We denote the resulting scoring rule as $\mathcal{G}$-CRPS.

Proposition 1 *Let F be the CDF of a Gaussian random variable with mean μ and standard deviation σ. Then, CRPS reads*

$$\mathcal{G}\text{-CRPS}(\mu, \sigma, y) = \sigma \left[\frac{y-\mu}{\sigma} \left(2\mathcal{N}\left(\frac{y-\mu}{\sigma}\right) - 1 \right) + 2\varphi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}} \right], \tag{8}$$

where $\varphi(\cdot)$ and $\mathcal{N}(\cdot)$ represent the PDF and the CDF of a standard normal, respectively.

Proof See Gneiting et al. (2005), p.1102. ■

The obtained expression is very simple and it only makes use of the Gaussian PDF and CDF. We notice that $\mathcal{G}$-CRPS is symmetric with respect to the standardised error $\frac{y-\mu}{\sigma}$, a fact that is explained by the symmetry of the normal distribution. In this sense, this scoring rule does not discriminate on whether the realisation y is above or below the forecasted mean μ, but only on the magnitude of the standardised error $\left|\frac{y-\mu}{\sigma}\right|$.

3. The new loss function

Highly-parametrised ML models, such as ANNs, are particularly prone to overconfidence (see, e.g., Guo et al., 2017). In the context of probabilistic forecasting, this results in lack of generalisation on out-of-sample data and in underestimation of the true predictive uncertainty. As mentioned in Section 1, the issue of overconfidence is particularly relevant for safety-critical applications, such as energy systems.

In this section, we propose a new methodology to tackle overconfidence and to enhance the reliability of probabilistic forecasts. We introduce a novel "robustified" loss function, obtained by suitably modifying the CRPS, which can be employed for training ANNs.

3.1. Enhancing reliability: the general case

Nowadays it is standard practice to use CRPS – and its discretised version, the APL – to assess the quality of probabilistic forecasts, especially in the energy sector (see, e.g., Hong et al., 2016; Nowotarski and Weron, 2018). It would therefore seem natural to use CRPS as a loss function to train a well-performing forecasting model. Nevertheless, a good predictive model must both be accurate and ensure high out-of-sample reliability; as we show in Section 4, CRPS often falls short in achieving this second goal.

To enhance generalisation capability on unseen data, we introduce in the following $\text{CRPS}^{[\lambda]}$, a new loss function obtained by adding a regularisation term to the original CRPS, which is designed to counter predictive overconfidence. The derivation of this loss function requires two steps: first, we suitably modify the central Pinball Loss; then, we compute its integral average over all percentiles, as in (7).

Let us first focus on the structure of the central Pinball Loss. By writing explicitly the two components in (2), we can reformulate this scoring function as

$$\mathcal{P}_{\text{C}}(l,u,y;\,\alpha) = \frac{1-\alpha}{2}(u-l) + \begin{cases} (y-u) & \text{if } \; y > u\,, \\ 0 & \text{if } \; y \in [l,u]\,, \\ (l-y) & \text{if } \; y < l\,. \end{cases} \tag{9}$$

It is thus possible to notice that the central Pinball Loss is composed of two terms:

- the first one, $\frac{1-\alpha}{2}(u-l)$, pertains to the *sharpness* of the PI $[l,u]$;
- the second one, $(y-u)\mathbb{1}_{y>u} + (l-y)\mathbb{1}_{y<l}$, is a penalty applied when the realisation falls outside the PI, thereby assessing the *reliability* of the PI $[l,u]$.

As our goal is to mitigate overconfidence, we aim to increase the impact of the second term, the reliability penalty – or equivalently to lower that of the sharpness term. We modify the central Pinball Loss by introducing a coefficient $\lambda \in [0,1)$, and we define the novel loss function

$$\mathcal{P}_{\text{C}}^{[\lambda]}(l,u,y;\alpha) := (1-\lambda)\frac{1-\alpha}{2}(u-l) + (y-u)\mathbb{1}_{y>u} + (l-y)\mathbb{1}_{y>l} = \tag{10}$$

$$= \mathcal{P}_{\text{C}}(l,u,y;\alpha) - \lambda\,\frac{1-\alpha}{2}(u-l)\,, \tag{11}$$

The correction we propose serves as an adjustment for overconfidence, as it encourages models to prioritise reliability over sharpness. Specifically, as highlighted by expression (11), the loss function $\mathcal{P}_{\text{C}}^{[\lambda]}$ introduces a reward for wide PIs, which can be suitably controlled via the parameter λ.

As discussed in Section 2, for many practical applications it is required to train models that output entire probability distributions. As a second step, we use (7) to similarly modify CRPS, obtaining a new loss function for density forecasting. We define this “robustified” version of CRPS as

$$\text{CRPS}^{[\lambda]}(F,y) := \int_0^1 \mathcal{P}_{\text{C}}^{[\lambda]}\left(F^{-1}\left(\frac{1-\alpha}{2}\right),\, F^{-1}\left(\frac{1+\alpha}{2}\right),\, y\,;\,\alpha\right)\mathrm{d}\alpha\;. \tag{12}$$

Since $\text{CRPS}^{[\lambda]}$ is purposely designed as a loss function, it is essential for it to have a simple expression. Learning algorithms need to evaluate loss functions a large number of times during training, and the numerical integration of (12) does not ensure sufficient accuracy or computational speed. In the next subsection, we discuss a relevant case in which this new loss function admits an explicit formula.

3.2. Enhancing reliability: the Gaussian case

In the Gaussian case, it is possible to deduce a closed-form expression for CRPS$^{[\lambda]}$, which extends that presented in Proposition 1. Specifically, the following result holds.

Proposition 2 *Let F be the CDF of a Gaussian random variable with mean μ and standard deviation σ. Then, the new loss function reads:*

$$\mathcal{G}\text{-CRPS}^{[\lambda]}(\mu,\sigma,y) := \mathcal{G}\text{-CRPS}(\mu,\sigma,y) - \lambda\left(\frac{\sqrt{2}-1}{\sqrt{\pi}}\,\sigma\right), \tag{13}$$

where $\mathcal{G}$-CRPS is the original scoring rule defined in (8).

Proof The proof is analogous to that of Proposition 1. ■

The new loss function is an extension of the original one. It includes an adjustment term which is similar to the Lasso and Ridge penalties commonly employed in deep learning to increase generalisation capability (see, e.g., Aggarwal, 2018). This term is specifically tailored to affect only the σ component, aiming to mitigate overconfidence by encouraging the generation of more dispersed density forecasts.

4. Experimental Analysis

To test the effectiveness of the new loss function in mitigating overconfidence, we conduct an experimental analysis on a benchmark dataset. It contains the households' hourly electricity load, plus the dry-bulb and the dew-point temperatures, for the New England region. Data are published by the Independent System Operator of New England (ISO-NE), and this dataset was adopted for the Global Energy Forecasting Competition in 2017 (GEFCom2017 Hong et al., 2019). For this analysis, we work with data for the whole region, considering six calendar years, from January 2007 up to December 2012.

4.1. The model

We consider the predictive methodology described in Baviera and Manzoni (2022), which makes use of RNN(p) models, recurrent networks with a single hidden layer and multiple feedbacks of Jordan type (i.e. lagged feedbacks from the output layer to the input layer). These models are the nonlinear extension of ARX(p) models; thus, they are characterised by a high degree of interpretability.

We aim to generate probabilistic forecasts for the electricity load on an hourly scale. We consider a one-year-ahead time horizon, a problem that in the literature is known as mid-term forecasting (see, e.g., Hong and Fan, 2016). In the energy sector, this represents a challenging forecasting problem due to the coexistence of multi-scale seasonality (at a yearly, weekly and daily scale) and complex temporal dependencies.

The predictive methodology considers as target variable the logarithm of the electricity load, as standard in the literature (see, e.g., Benth et al., 2008, and references therein). It is a two-stage modelling scheme that considers an initial detrending and deseasonalisation of the time series and a more refined modelling of the residuals using an RNN.

As a first step, a General Linear Model (GLM) is employed to remove seasonal and trend components in the time series, using only calendar variables as regressors. Specifically, the regressors considered are: a linear trend, the first two Fourier terms for the day-of-the-year and the hour-of-the-day, and three dummy variables that identify Saturday, Sunday and holidays.

As a second step, the residuals of this linear regression are processed by the RNN(p) model, so as to capture the nonlinear dependencies involving calendar variables, weather variables and autoregressive effects. Each residual is modelled as a Gaussian random variable with mean μ_t and standard deviation σ_t, outputted by the RNN.

We start by training the baseline model, i.e. the model that uses the original $\mathcal{G}$-CRPS (8) as loss function for the RNN. Then, to enhance reliability and mitigate overconfidence, we utilise the new loss function $\mathcal{G}$-CRPS$^{[\lambda]}$ (13), selecting multiple values for the overconfidence parameter λ. Moreover, we also compare the results obtained with Maximum Likelihood Estimation (MLE), which corresponds to using the negative Gaussian log-likelihood as loss function (see, e.g., Goodfellow et al., 2016). The results are striking.

4.2. Results

We divide the dataset into a training set (data from 2007 to 2010), a validation set (2011) and a test set (2012). As standard in the literature, we consider training and test sets that span the full length of the year due to the significant yearly seasonality observed in the data (see, e.g., Hyndman and Fan, 2010).

We consider an RNN($\{1, 2, 24\}$) model due to the autocorrelation of the residuals. In this way, we aim to capture both hourly (i.e. 1 and 2 hours before) and daily (i.e. 24 hours before) autoregressive effects: when predicting the distributional parameters for the hour t, the network takes as autoregressive inputs the parameters forecasted for the hours t-1, t-2 and t-24. To train the RNNs, we split the original dataset into sequences with length equal to two days. Moreover, we use Adam as optimiser, and train the model using early stopping with a patience of 100 epochs. As standard in the literature, a sigmoid activation function is used in the hidden layer, while a linear one is used in the output layer (see, e.g., Goodfellow et al., 2016).

We perform a grid search to determine the best architecture for the RNN. We consider three main hyperparameters: the number of hidden neurons, considering 5, 10, 15 and 20 as possible values; the batch size, which can be 32, 64, or 128; and the learning rate, either 1e-4, 5e-4, 1e-3 or 5e-3. We perform a grid search for $\mathcal{G}$-CRPS, selecting the configuration that records the best accuracy in terms of MAPE. The best triplet of hyperparameters – corresponding to hidden neurons, batch size and learning rate – is found to be $[10, 32, 0.0005]$.

To test the effectiveness of the proposed approach, we then recalibrate the model with the selected optimal hyperparameters on the 2008-2011 data and analyse the results on the 2012 data. We compute the Empirical Coverage, denoted as EC(α), for the PIs with significance levels α ranging from 90% to 99% – those relevant for standard operations in the energy sector, where high reliability is required (see, e.g., Wang et al., 2017). To achieve more robust estimates, each model is trained ten times with a different random initialisation of the RNN weights. The mean of the ten Empirical Coverages is then considered, and the Standard Error (SE) is used as measure of variability. Different values of the overconfidence parameter λ are considered.

Figure 1 shows the main result: the parameter λ proves capable of increasing the reliability of the probabilistic forecasts. By changing the value of λ, it is possible to reduce overconfidence, obtaining more reliable forecasts. In particular, we observe that the MLE-trained model is outperformed in terms of predictive reliability at every considered confidence level α when λ is greater than 0.10.

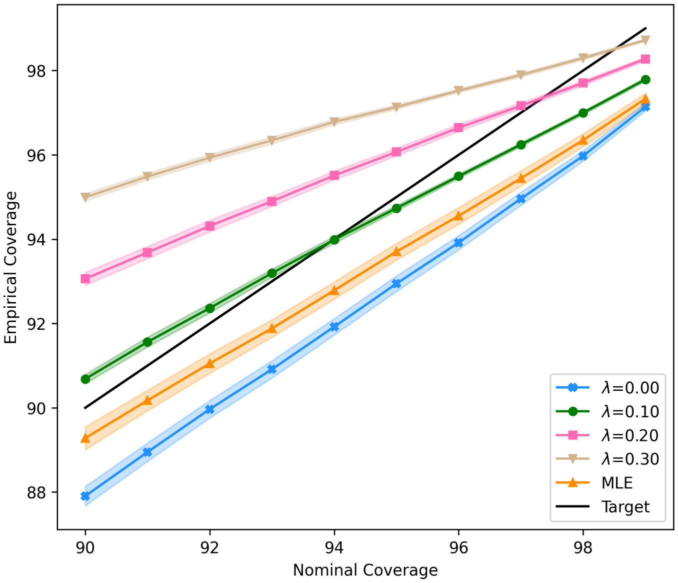

Figure 1: Coverage plot for PIs on the test set (2012). We plot the Empirical Coverage, i.e. the proportion of realisations falling inside the corresponding α-PI, against the Nominal Coverage level α. Ten repetitions with different random RNN seeds have been performed: the plots show the average coverage, with the shaded areas corresponding to ±1 SE.

For a quantitative analysis, we evaluate forecast quality using three standard performance metrics, namely MAPE, RMSE and APL. Moreover, we assess reliability by considering the Empirical Coverage at level $\alpha = 90\%$ and $\alpha = 95\%$, and the Average Absolute Coverage Error (AACE; see, e.g., Alfieri and De Falco, 2020), an index defined as

$$\text{AACE} := \frac{1}{Q} \sum_{q=1}^{Q} \left| \text{EC}(\alpha_q) - \alpha_q \right| , \tag{14}$$

where $\{\alpha_q\}_{q=1}^{Q}$ are the considered confidence levels: in our case, those between 90% and 99%. The results are reported in Table 1.

(*a*) $\mathcal{G}$-CRPS$^{[\lambda]}$

λ	MAPE [%]	RMSE [MWh]	APL [MWh]	EC(90%) [%]	EC(95%) [%]	AACE [%]
0.00	2.01 ± 0.01	409.58 ± 1.31	104.23 ± 0.33	87.90 ± 0.22	92.94 ± 0.18	2.03 ± 0.17
0.05	2.00 ± 0.01	408.45 ± 1.60	103.97 ± 0.39	89.21 ± 0.14	93.64 ± 0.17	1.26 ± 0.13
0.10	2.00 ± 0.01	406.49 ± 1.64	103.87 ± 0.43	90.68 ± 0.10	94.73 ± 0.06	0.59 ± 0.03
0.15	1.99 ± 0.01	406.11 ± 1.67	103.77 ± 0.42	91.87 ± 0.19	95.36 ± 0.11	0.91 ± 0.08
0.20	2.00 ± 0.01	406.91 ± 1.50	104.12 ± 0.40	93.06 ± 0.15	96.06 ± 0.09	1.44 ± 0.07
0.25	2.00 ± 0.01	407.78 ± 1.49	104.62 ± 0.42	94.06 ± 0.11	96.60 ± 0.07	1.94 ± 0.06
0.30	2.00 ± 0.01	408.67 ± 1.76	105.27 ± 0.46	94.99 ± 0.08	97.13 ± 0.05	2.46 ± 0.05

(*b*) MLE

MAPE [%]	RMSE [MWh]	APL [MWh]	EC(90%) [%]	EC(95%) [%]	AACE [%]
2.05 ± 0.01	423.59 ± 1.20	107.54 ± 0.39	89.28 ± 0.26	93.70 ± 0.19	1.33 ± 3.87

Table 1: Results on the test set (2012): MAPE, RMSE and APL for the considered loss functions, together with Empirical Coverage (EC) for the 90% and 95% PIs, and the Average Absolute Coverage Error (AACE). We report the average values and the SEs of the statistics, obtained as a result of ten repetitions of the training with different random initialisation.

In all cases, we obtain probabilistic forecasts with high reliability, achieving in particular an excellent coverage for $\lambda = 0.10$ and $\lambda = 0.15$. Moreover, the point forecasts are extremely accurate, with a MAPE well below the threshold of 2.50%, the value commonly used by practitioners to identify very good predictive models.

Compared to the benchmark MLE training, the loss function $\mathcal{G}$-CRPS$^{[\lambda]}$ demonstrates its capability to generate not only more reliable probabilistic forecasts but also more accurate point forecasts. Moreover, we notice that MAPE, RMSE and APL are not affected by the choice of λ: the adjustment does not have any impact in terms of the means μ_t of the predicted distributions, but only on their standard deviations σ_t.

Finally, the effect of the adjustment for overconfidence can be observed in Figure 2. The plot shows the forecasts for an entire week in August 2012, highlighting how the 95% PIs are modified when the coefficient λ is moved from $\lambda = 0.00$ to $\lambda = 0.10$. The PIs appear to be slightly enlarged, but not uniformly: the adjustment is more pronounced during the central hours of the day compared to the night hours, where less variability is present.

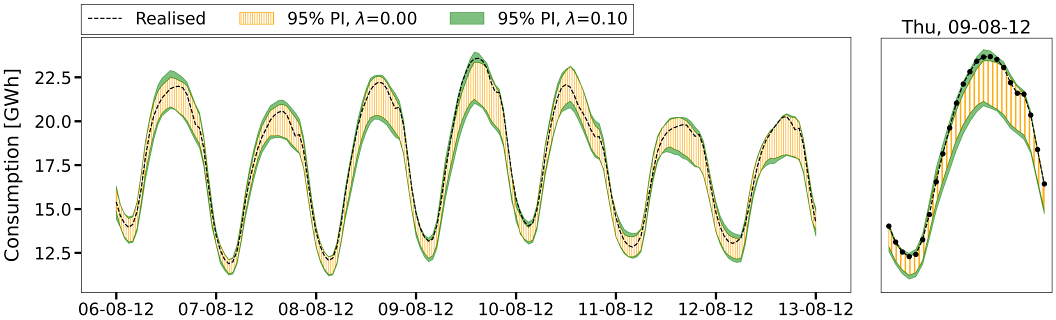

Figure 2: Density forecasts of hourly load on the test set (2012) for a week in August, when $\mathcal{G}$-CRPS$^{[0.00]}$ and $\mathcal{G}$-CRPS$^{[0.10]}$ are used for training the predictive model. The 95% PIs are plotted in the two cases, together with the realised electricity load. We observe that the PIs obtained with the new loss function are only slightly larger than the ones obtained with the standard technique. Nevertheless, as shown in the zoomed plot (*right*), this almost negligible modification allows the realised load to fall within the PIs during periods where, using the standard technique, they would have fallen outside of the PIs.

5. Conclusions

Artificial Neural Networks, although being highly accurate, often produce overconfident predictions potentially undermining decision-making in critical systems such as those in the energy sector. In the probabilistic load forecasting domain, very simple forecasting techniques, even if relatively inaccurate, are frequently employed because they provide reliable forecasts.

In this paper, we have presented a novel forecasting methodology aimed at reducing the predictive overconfidence. First, we have introduced a new loss function that incorporates an additional parameter λ. This loss function, called CRPS$^{[\lambda]}$, is designed to enhance the reliability of probabilistic forecasts, and allows fine-tuning of the calibration of distribution tails, which are crucial for decision-making in the energy sector. Second, we have tested the proposed approach on a benchmark dataset of electricity load, demonstrating that our methodology achieves exceptionally accurate and reliable results. Third, we have shown that the increase in forecast reliability – and in particular that of the tails, i.e. for high confidence levels – comes at no expense of forecast accuracy.

Acknowledgments

We are grateful to C. Lucheroni, S. Vantini, F. Ziel and two anonymous reviewers for helpful suggestions. The present research is part of the activities of “Dipartimento di Eccellenza 2023-2027”. The authors are members of the INdAM Research group GNCS.

References

Charu C. Aggarwal. *Neural Networks and Deep Learning.* Springer, 2018.

Luisa Alfieri and Pasquale De Falco. Wavelet-Based Decompositions in Probabilistic Load Forecasting. *IEEE Transactions on Smart Grid*, 11(2):1367–1376, 2020.

Michele Azzone and Roberto Baviera. Neural Network middle-term probabilistic forecasting of daily power consumption. *Journal of Energy Markets*, 1(14):1–26, 2021.

Roberto Baviera and Pietro Manzoni. Tree-Based Learning in RNNs for Power Consumption Forecasting, 2022. preprint, arXiv:2209.01378.

Roberto Baviera and Giuseppe Messuti. Daily middle-term probabilistic forecasting of power consumption in North-East England. *Energy Systems*, pages 1–23, 2023.

Fred Espen Benth, Jurate Saltyte Benth, and Steen Koekebakker. *Stochastic modelling of electricity and related markets.* World Scientific, 2008.

Jonathan Berrisch and Florian Ziel. CRPS learning. *Journal of Econometrics*, 237:105221, 2023.

Tobias Fissler, Michael Merz, and Mario V. Wüthrich. Deep quantile and deep composite triplet regression. *Insurance: Mathematics and Economics*, 109:94–112, 2023.

Ciaran Gilbert, Jethro Browell, and Bruce Stephen. Probabilistic load forecasting for the low voltage network: Forecast fusion and daily peaks. *Sustainable Energy, Grids and Networks*, 34:100998, 2023.

Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:243–268, 2007.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT Press, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern Neural Networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

Tao Hong, Pierre Pinson, Shu Fan, Hamidreza Zareipour, Alberto Troccoli, and Rob J. Hyndman. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913, 2016.

Tao Hong, Jingrui Xie, and Jonathan Black. Global Energy Forecasting Competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting*, 35(4): 1389–1399, 2019.

Rob J. Hyndman and Shu Fan. Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, 25:1142–1153, 2010.

Roger Könker. *Quantile Regression*. Cambridge University Press, 2005.

Agustinus Kristiadi, Matthias Hein, and Philipp Hennig. Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, 2020.

Tianyi Li, Yi Wang, and Ning Zhang. Combining Probability Density Forecasts for Power Electrical Loads. *IEEE Transactions on Smart Grid*, 11(2):1679–1690, 2019.

Bidong Liu, Jakub Nowotarski, Tao Hong, and Rafał Weron. Probabilistic load forecasting via quantile regression averaging on sister forecasts. *IEEE Transactions on Smart Grid*, 8(2):730–737, 2015.

Grzegorz Marcjasz, Michał Narajewski, Rafał Weron, and Florian Ziel. Distributional Neural Networks for electricity price forecasting. *Energy Economics*, 125:106843, 2023.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

Jakub Nowotarski and Rafał Weron. Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, 81:1548–1568, 2018.

Pierre Pinson, Henrik A. Nielsen, Jan K. Møller, Henrik Madsen, and George N. Kariniotakis. Non-parametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy*, 10(6):497–516, 2007.

Slawek Smyl and Grace Hua. Machine learning methods for GEFCom2017 probabilistic load forecasting. *International Journal of Forecasting*, 35(4):1424–1431, 2019.

Julian Vossen, Baptiste Feron, and Antonello Monti. Probabilistic Forecasting of Household Electrical Load Using Artificial Neural Networks. In *2018 IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, pages 1–6, 2018.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep Neural Networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34:11809–11820, 2021.

Huaizhi Wang, Haiyan Yi, Jianchun Peng, Guibin Wang, Yitao Liu, Hui Jiang, and Wenxin Liu. Deterministic and probabilistic forecasting of photovoltaic power based on deep convolutional Neural Network. *Energy conversion and management*, 153:409–422, 2017.

Ke Wang, Yao Zhang, Fan Lin, Jianxue Wang, and Morun Zhu. Nonparametric Probabilistic Forecasting for Wind Power Generation Using Quadratic Spline Quantile Function and Autoregressive Recurrent Neural Network. *IEEE Transactions on Sustainable Energy*, 13 (4):1930–1943, 2022.

Yi Wang, Dahua Gan, Mingyang Sun, Ning Zhang, Zongxiang Lu, and Chongqing Kang. Probabilistic individual load forecasting using Pinball Loss guided LSTM. *Applied Energy*, 235:10–20, 2019.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Fengand Bo An, and Yixuan Li. Mitigating Neural Network Overconfidence with Logit Normalization. In *International Conference on Machine Learning*, pages 23631–23644. PMLR, 2022.

Yandong Yang, Shufang Li, Wenqi Li, and Meijun Qu. Power load probability density forecasting using Gaussian process quantile regression. *Applied Energy*, 213:499–509, 2018.

Yandong Yang, Weijun Hong, and Shufang Li. Deep ensemble learning based probabilistic load forecasting in smart grids. *Energy*, 189:116324, 2019.