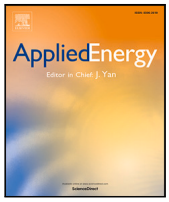
Applied Energy 361 (2024) 122900

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy



A novel day-ahead regional and probabilistic wind power forecasting framework using deep CNNs and conformalized regression forests

Jef Jonkers [a,*], Diego Nieves Avendano [a,b], Glenn Van Wallendael [a,b], Sofie Van Hoecke [a,b]

[a] IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium
[b] IDLab, imec, Ghent, Belgium

ARTICLE INFO

Dataset link: here

Keywords:
Regional wind power forecasting
Quantile forecasting
Convolutional Neural Networks (CNN)
Prediction distribution
Conformal predictive distribution
Quantile regression forest

ABSTRACT

Regional forecasting is crucial for a balanced energy delivery system and for achieving the global transition to clean energy. However, regional wind forecasting is challenging due to uncertain weather prediction and its high dimensional nature. Most solutions are limited to single-turbine or farm/park forecasting; therefore, this work proposes a day-ahead regional wind power forecasting framework using deep Convolutional Neural Networks (CNN) with context-aware turbine maps and Conformal Quantile Regression (CQR) to generate quantile forecasts with valid coverage.

Additionally, this work introduces the use of the Split Conformal Predictive System (SCPS) to generate valid prediction distributions, which has not yet been proposed for wind power forecasting in general. As well as a new method to generate calibrated prediction distributions based on SCPS and Quantile Regression Forests (QRF). This new method, named Split Conformal Distribution Regression Forests (SCDRF), allows for conditional conformal predictive distribution that increases efficiency compared to SCPS while maintaining valid coverage. SCDRF, together with CNNs and context-aware turbine maps, outperforms the existing models on the evaluated dataset, reducing the pinball loss by 5.89% while having more flexibility due to the generation of prediction distributions that can be used to generate any quantile prediction without retraining the model.

1. Introduction

Reducing carbon intensity in electricity generation is crucial to the ongoing battle against climate change. This is because the power generation sector accounts for more than 40% of energy-related carbon dioxide (CO_2) emissions.[1] Therefore, it is imperative to exert significant effort to attain net-zero emissions by 2050 to give the world a chance to limit the global temperature rise to 1.5 °C.[2] One of the critical areas in pursuing this objective is the transition to utilizing clean and renewable energy sources, such as wind power, to generate electricity. However, the increasing reliance on wind-generated electricity comes with several challenges, such as maintaining balance within the electricity grid and dealing with volatile energy prices, which leads to a rise in the overall energy cost. Accurate and reliable energy production forecasts can contribute to a more sustainable energy mix, decreasing operating costs, mitigating reserve shortfalls, and limiting the extent of wind curtailment [1]. However, precisely those accurate and reliable predictions of the power outputs of wind energy production are a complex task since wind energy production depends on weather conditions with a stochastic nature.

This work focuses on a crucial type of forecast, namely the day-ahead forecast, which provides predictions of energy output a day before actual generation, which is necessary for enabling the effective operation of the electrical grids. These markets use a matching principle, i.e., a forward contract, where a seller promises to deliver/produce a certain amount of energy and a buyer promises to receive/consume the same amount. Failure to adhere to the terms of the forward contract results in financial penalties. These forward contracts are particularly challenging for photovoltaic (PV) and wind energy producers due to the stochastic nature of these energy sources, as these are dependent on the weather conditions, increasing the chance of financial penalties.

Whereas forecasting approaches for day-ahead power generation for renewable energy sources are an extensively studied subject [2–7], the majority of approaches focus on forecasting for single turbine or farms [7], while regional forecasting methods have received relatively limited attention despite their pivotal role in maintaining a balanced energy system [2–18]. Regional forecasts help energy producers with strategic economic decisions, such as, turning on or off a hydropower system or gas power plant in case of additional or reduced energy needs. System

* Corresponding author.
E-mail address: jef.jonkers@ugent.be (J. Jonkers).

[1] https://www.iea.org/reports/world-energy-outlook-2021/overview.
[2] https://www.iea.org/reports/net-zero-by-2050.

https://doi.org/10.1016/j.apenergy.2024.122900
Received 5 December 2023; Received in revised form 26 January 2024; Accepted 18 February 2024
Available online 21 February 2024

operators can also use these regional forecasts to balance the electrical grid so that the energy consumed matches the amount fed into the grid.

We consider the EEM20 dataset as a starting point for this work, which was made available in a competitive format at the EEM20 conference. To our knowledge, this dataset is the only publically available dataset that considers regional forecasting spanning multiple price regions and gives access to grid-like numerical weather prediction (NWP) data and the specifications of 4004 turbines located in the Swedish area. An interesting challenge the EEM20 dataset exhibits is the increase in operational wind turbines in each region by expanding existing wind farms and creating new wind farms at entirely different locations. This challenge can cause a distribution shift and consequently deteriorate model performance over time. This work tries to resolve this by introducing the turbine map, which gives spatial context to a deep and dense convolutional network. These turbine maps convert the changing and sparse turbine capacity information into a dense grid and connect the NWPs with information about the wind turbines at the specific grid points.

In addition to enhancing regional deterministic forecasting and addressing shifts in distributions, we also assess various methodologies for generating probabilistic forecasts. These uncertainty quantifications can improve decision-making processes within the electric power industry [4,19]. According to Zhang et al. [3], probabilistic forecasting can be categorized into density forecasting, quantile forecasting, and interval prediction. Our standpoint is that probability density forecasting surpasses the other types of probability forecasting due to its inherent ability to transform from a density forecast seamlessly into quantile and interval forecasts. Conversely, the opposite transformation is unfeasible. Consequently, our work proposes an approach that allows the prediction of conditional probability distribution, which imparts considerably more information than quantile forecasts and proves invaluable in decision-making scenarios. For instance, it enables us to determine the probability of power output being above or below a certain production threshold.

While the EEM20 forecasting competition and its associated literature concentrate on quantile forecasting and evaluate forecast approaches using the pinball loss function, it has been noted that this approach leans towards narrower forecasts, sacrificing reliability. Given our intention to benchmark our system against existing methodologies, we will conduct both density and quantile forecasts, subjecting them to evaluation using the pinball loss function. Moreover, our evaluation will encompass forecast reliability, assessed by quantifying the coverage of the quantile predictions, as unreliable probability predictions hold minimal value.

To achieve credible forecasts, we will assess a range of machine-learning techniques in conjunction with conformal prediction. Specifically, we propose applying Conformal Quantile Regression (CQR) and the Conformal Predictive System (CPS) for quantile and density forecasting. Notably, our work pioneers the use of CQR for regional wind power prediction and introduces CPS to wind power prediction in general, enabling the generation of prediction distributions.

In summary, the main contributions of this paper are as follows:

1. The use of deep and dense CNN architectures and the introduction of turbine maps address the distributional shift by supporting the increase of existing wind farms' capacity and the creation of new wind farms. Additionally, it allows the use of a single model for multiple regions and implicitly augments the training dataset. Essentially, these turbine maps make the model context-aware.
2. A proposed quantile forecasting approach for day-ahead regional wind forecasting that outperforms the current state-of-the-art on the EEM20 dataset.
3. The application of CQR to regional wind power forecasting to achieve adaptive calibrated quantile forecasts.
4. The use of CPS to generate prediction distributions, which can be extremely useful in the decision-making process of different actors in the electric power industry.
5. The introduction of a new adaptive CPS algorithm, Split Conformal Distribution Regression Forests (SCDRF), which combines CPS with the philosophy of Quantile Regression Forests (QRF) [20].

The remainder of the paper is as follows. Section 2 discusses related works on regional forecasting and conformal prediction in wind power forecasting. Section 3 discusses and analyzes the available data. Next, Section 4 describes the proposed forecasting framework and models that are evaluated. Section 5 discusses the performance of the presented models and compares them with the state-of-the-art models on the EEM20 dataset. Finally, Sections 7 and 6 discuss future work and present the conclusion, respectively.

2. Related work

Deterministic forecasting methods provide a single-spot forecast and strive for the highest possible accuracy. Many ways are proposed in the literature to handle deterministic power forecasting of wind energy sources. The literature generally divides these methods into three main approaches: the physical approach, the statistical/machine learning approach, and the hybrid approach, which combines the physical and statistical/machine learning approaches [2,4,6,7]. Another way to classify wind power forecasting is based on the time horizon of the forecasts, i.e., the time span between the moment when the forecast is generated and the future point in time for which the prediction is intended. Santhosh et al. [6] mention four classes based on the forecasting time scale: very short-term (a few seconds to 30 min ahead), short-term (30 min to day-ahead), medium-term (day- to month-ahead) and long-term (more than month-ahead). This work will focus on methods that handle short- to medium-term forecasting, 12-to 72-hours-ahead (day-ahead forecasts), which has, for example, applications in the operational security of the day-ahead electricity market. Currently, most of the literature agrees that the hybrid approach outperforms the physical and statistical/machine learning approach [21–23], and also, the proposed forecasting approach in this work is a hybrid one.

Forecasting approaches for day-ahead wind power production are a heavily studied subject. However, most literature only covers approaches to forecasting the power output of a single farm or turbine [2–7]. The first regional forecasting approaches viewed the task as a univariate time series problem [8–10]. Treating the day-ahead forecast as a univariate problem neglects the fact that the outcome is highly dependent on the weather. More advanced multivariate machine learning methods have recently been proposed with NWP for regional wind power forecasting. Some methods, which are called a top approach, leverage grid-like NWPs features [18,24–28] while others, referred to as an upscaling or bottom-up approach [15], use NWPs at specific wind farms [11–17]. Another way of categorizing different regional forecasting approaches is the availability of the power output of single turbines or farms during training. This availability enables the use of a bottom-up or upscaling approach where forecasts are made for individual farms and then aggregated to form a regional forecast, with the possibility of correcting for correlated errors [11,13–15,17,18]. However, as these individual power time series are not always available, we focus, in our approach, on a top approach that uses grid-like NWP for regional wind power prediction. Another argument pro grid-like NWP features is that more information is embedded than in the NWPs per farm. In the top approaches, deep learning methods, such as CNNs and Vision Transformers (ViT), are currently not used while we believe that these methods could better handle the curse of dimensionality, which is present because of the grid-like NWPs, the size of a price region, and the number of turbines in an area. We argue that these methods could be specifically used to learn more complicated patterns from the NWP data and turbine maps.

In recent years, research on probabilistic forecasting for day-ahead wind power forecasting started to appear that tries to provide a measure of uncertainty, e.g., an interval where, with a certain probability, the real value could lie within. These uncertainty quantifications help optimize the decision-making process of actors in the electric power industry [4,19]. The literature divides probabilistic forecasting into parametric and non-parametric approaches [3,4,7]. Parametric methods derive forecasting intervals from parametric distributions like Gaussian distribution, which differs from non-parametric approaches, that do not assume the underlying distribution. The non-parametric approach is the most researched probabilistic approach for (regional) wind power forecasting [12,13,15,16,24,25,29–33]. An advantage of this approach is that no assumptions on the underlying error distribution need to be made. Nevertheless, the fact that these approaches tend to be more complex is a disadvantage, but their superior performance generally outweighs this. Nonetheless, parametric approaches are still being proposed with promising results, however, using more complex parametric distributions such as beta mixture distributions [34]. In Bessa et al. [4], another distinction is made between statistical methods based on deterministic NWP forecasts and NWP ensemble forecasts. Methods based on deterministic NWP forecasts derive uncertainty using statistical methods and deterministic NWP as input. Contrarily, approaches based on ensemble forecasts determine uncertainty by applying power forecasts on each ensemble member, or by applying a reduction method on the ensemble members. The downside of ensemble forecasts is that they drastically increase the dimension of an already high-dimensional problem. This work will implement the latter approach as it is the one that tries to capture the most uncertainty. We will deal with the dimensionality increase by using the ensemble forecasts' summary statistics, like the mean and variance, as features for deep learning and tree-based models to forecast the power output and related uncertainty.

Tree-based ensembles and physics-inspired input features seem to be the most popular and successful method to generate a regional forecast with an uncertainty measure provided next to the predicted output [25,27,29]. However, more complex deep neural network models, like CNN and Long Short-Term Memory (LSTM), which proved to be successful on the single wind farm forecasting task [7] and in the upscaling approach [12,14,17,35], are also not thoroughly researched for the regional forecasting task. Basu et al. [24] propose an approach that uses a CNN network. However, the convolutional layers are mainly used for dimensionality reduction of the input data, a grid of NWP variables spanning the entire Swedish region. As mentioned before, we propose to use an extra grid-like feature representing the locations and capacities of the turbines in the region, i.e., a turbine map, on top of the NWP variables. A deeper and denser CNN model can be trained by linking the characteristics of the turbines and NWP variables, which will benefit from learning more complex, non-linear relations.

While the conformal prediction framework can provide non-asymptotic coverage guarantees with limited assumptions, see Section 4, it has not been often applied to wind power prediction [33,36,37]. The proposed approaches [33,36,37] specifically use Conformalized Quantile Regression (CQR) and present promising results. However, the methods are limited to interval forecasting and do not consider density forecasting, which would be of valuable use in decision-making frameworks. Therefore, this work will propose using Conformal Predictive Systems (CPS), which allows for generating calibrated predictive distributions.

3. Data description

This paper uses the Swedish regional wind power production dataset provided by the EEM20 competition to train and evaluate the regional wind power forecasting approaches. The competition's goal was to generate a day-ahead quantile forecast of aggregated wind power production for the four Swedish bidding areas, also known as price regions. The price regions are named SE1, SE2, SE3, and SE4 in this work. The lower the number in the name of the region, the higher in the north of Sweden it is situated; see Fig. 1. Also, note that in this setting, a day-ahead forecast means that on the current day at noon, the power output is forecasted for every hour of the next day. The competition consisted of six tasks:

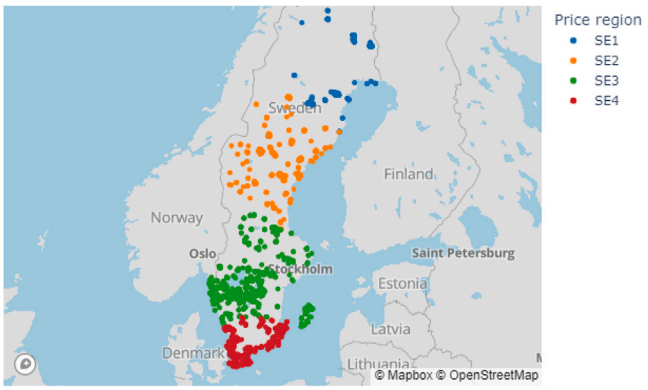


Fig. 1. Map of Swedish region with turbines from the EEM20 dataset classified per bidding area.

- Task 1: January until February 2001
- Task 2: March until April 2001
- Task 3: May until June 2001
- Task 4: July until August 2001
- Task 5: September until October 2001
- Task 6: November until December 2001

For each task, the participants are required to predict each price region's day-ahead hourly power production for the next two months based on NWP data and turbine records. The participants can input a grid of weather forecasts from that period. The dataset consists of three parts:

1. Records of all Swedish turbines containing the capacity, nacelle height, rotor diameter, installation date, and location of the turbines.
2. A 169 by 71 grid of numerical weather predictions covering the Swedish area (zonal wind, meridional wind, wind gust speed, temperature, pressure, relative humidity). For each grid of NWP variable, ten predictions by different ensemble members, i.e., forecasts, are available, which should reflect the uncertainty of the weather model estimate.
3. Hourly power time series aggregated over all turbines per bidding area.

3.1. Turbine data

The dataset consists of specifications of 4004 turbines located in the Swedish area. The dataset contains the capacity, nacelle height, rotor diameter, installation date, and location of the turbines. The dataset brings three major challenges for the forecasting task, related to the turbines. The first challenge is the heterogeneity between price regions. First of all, there is a difference in the number of tribunes in each region (see Fig. 1) and the aggregated capacity of each region (see Fig. 2). Additionally, there is considerable variability in the type of terrain, turbines, and climate in the different regions. The turbines in price regions SE1 and SE2 are installed in higher terrains, while those in regions SE3 and SE4 are closer to sea level. Notably, there is considerable variability among turbines within the same bidding area, with capacities ranging from 10 kW to 4.2 MW. This diversity is particularly pronounced in regions SE3 and SE4, where turbines

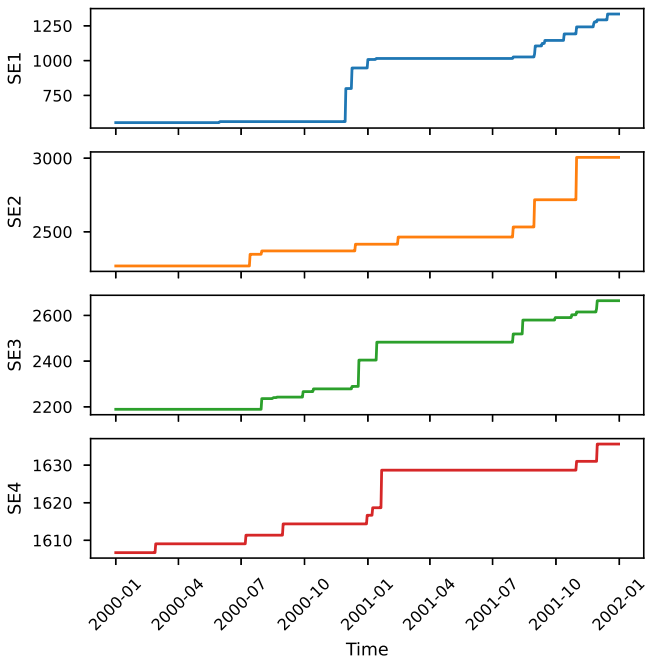


Fig. 2. Evolution of wind power capacity in MW for each bidding area.

exhibit a wide range of capacities. In contrast, regions SE1 and SE2 predominantly feature higher-capacity turbines.

The second challenge this dataset brings is the increase in power capacity over time; see Fig. 2. The capacity and the number of turbines in region SE1 more than doubled during 2001. New turbines are sometimes placed at existing farms but occasionally at entirely different places, making the forecasting problem more challenging. Consequently, these change the characteristics of the power time series and cause a distribution shift, which is difficult to handle for machine learning methods. To partially account for this distribution shift during forecasting and training, we introduce a "turbine map", defined in Section 4.1, which accounts for the evolution in the turbine mix. We also normalize the power time series by its region's installed capacity, to support changes over time.

A third challenge is the data quality of the records. In the normalized power time series of price region SE1, the power production regularly surpasses the theoretical boundary between October and December of 2000, see Fig. 3(a). This is impossible, indicating erroneous data, in the dataset. This phenomenon is also noticed in previous literature that uses this dataset [25,29]. There are three potential reasons for this error: (1) incomplete turbine records, as partly indicated by the competition organizers; (2) errors in some of the turbine records' installation dates; and (3) errors in the power time series of SE1. Since the forecasting method's target variable is the normalized power output, this can severely affect training and performance. Therefore, the normalized power output of SE1 is adjusted by moving some of the turbine's installation dates backward. More specifically, 66 turbine installation dates have been changed from 30/11/2000 to 30/09/2000. The 66 turbines are of the same type, each with a theoretical capacity of 3.2 MW, installed roughly at the exact location. By making this adjustment, the time series of SE1 does not output more power than the capacity, supporting this correction.

3.2. Numerical weather prediction data

For the EEM20 dataset, the Norwegian Meteorological Institute (MET) Norway provides a 10-member ensemble forecast for the Scandinavian region. According to Basu et al. [24], MET Norway uses the HARMONIE ensemble model [38] to forecast weather variables and quantify the uncertainty and predictability of NWP output. The NWP data consists of forecasts computed at 06:00 UTC the day before, thus in time for trading in the European day-ahead electricity markets [29]. The data consists of a spatial grid of 169 by 71 NWPs in hourly resolution for each day for two years (except for three missing dates). Each ensemble member provides predictions for seven meteorological variables. Hence, for every hour, 839,930 values are available, meaning that there are in the entire dataset 14.7 billion values to use considering the NWP alone. As input for the deterministic models and further analysis in this section, the mean of the ensemble data is used. The reason for this is to reduce the dimensionality of the data. In addition, some literature indicates the absence of performance gains using ensemble data compared to the mean of the ensemble [29,39].

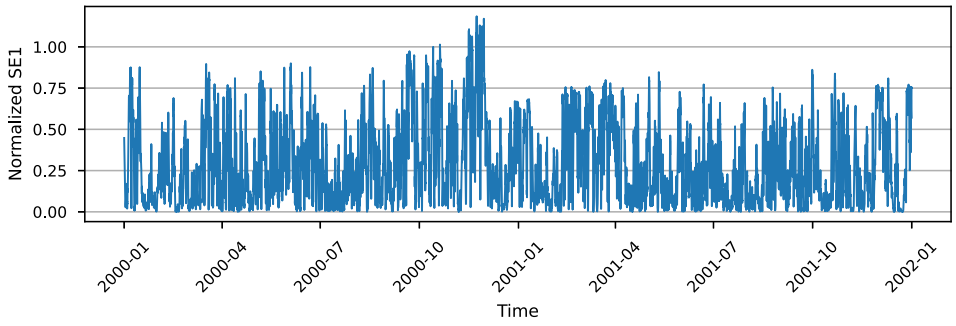
4. Methodology

This section will present our proposed forecasting framework that consists of four components, see Fig. 4: (1) a map construction component; (2) a deterministic forecasting component; (3) an uncertainty forecasting component; and (4) a calibration component. In this framework, we propose using a split conformal predictive system (SCPS) [40], a framework based on conformal prediction that outputs probability distributions instead of prediction sets. These probability distributions can then be used to generate quantile forecasting. We will compare this method to conformalized quantile regression (CQR) [41], which combines conformal prediction and quantile regression. SCPS and CQR require us to split the data into proper training and calibration sets to achieve well-calibrated probability distributions or quantile predictions. The different sets are used as follows: First, a deterministic model is trained on the proper training set, which uses feature maps from the map construction component as input. We use the calibration set for an early-stopping procedure to achieve greater predictive efficiency to prevent the deterministic model from overfitting. Note that this slightly violates the assumptions of CQR and SCPS; however, evaluations have shown that this has a minor effect on reliability and results in a significant increase in precision. Afterward, different regression methods are trained to quantify the uncertainty of each prediction or, in the case of CQR, to generate quantile predictions. These methods use the point forecast of the deterministic model as input, together with information about the price region, hour, and date of the requested forecast, as well as the current and previous mean region NWP predictions, and the variability in the ensemble NWPs. Finally, the calibration set is used to calculate nonconformity scores that quantify the error made by our uncalibrated quantile predictions, and these scores are then used to form probability distribution in the case of SCPS. In the case of CQR, the scores are used to recalibrate the quantile forecasts to adjust for possible under- or over-coverage.

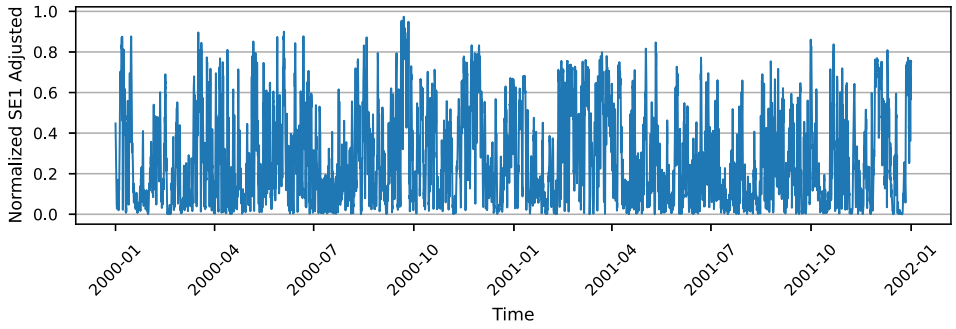
4.1. Map construction component

Our first component, the map construction component, performs data preprocessing and feature engineering to construct a tensor of dimensions $C \times H \times W$, where each channel C is a feature map that represents a particular variable, such as an NWP, with values at different coordinates. The 2D grid with dimension $H \times W$, where each index represents a specific coordinate, is a "zoomed-in" grid created such that each price region has values of NWP variables closely related to the turbines of the region. Each price region 2D grid has a dimension of 64×64 and is determined based on the location of the turbines in that region.

To reduce the dimensionality of the input data, not all available NWP variables are used as a feature map. After initial data analysis and applying a forward selection method with a hold-out validation set, two features were empirically found to be sufficient for the deterministic forecast: the absolute wind speed and wind gust speeds. Eq. (1) presents



(a) Initial



(b) Adjusted

Fig. 3. Time series of hourly produced normalized power for SE1.

the formula of the absolute wind speed, which uses the zonal and meridional wind speeds that indicate the direction and speed of the wind:

$$Wind = \sqrt{Wind_U^2 + Wind_V^2} \tag{1}$$

This new absolute wind speed variable is used instead of the zonal and meridional as, on the one hand, it reduces the dimension of the input data, and, on the other hand, it has a high correlation with power time series. Due to this transformation, we lose information about the direction of the wind. However, after several experiments, this information is of negligible value. We also evaluated the use of wind power density, frequently used in wind power forecasting for single turbine prediction [42], as a feature map; however, it did not increase performance on the hold-out set.

Crucially, a turbine map is also concatenated with the NWP feature maps to add more information to the input feature. Each turbine capacity of the price region is aggregated to a grid point on the turbine map. The turbine map has identical dimensions to the cropped NWP grid and represents equivalent locations. The map aggregates the capacity of a turbine in a weighted way to the closest four grid points, where a higher weight is assigned to a more nearly point, see Algorithm 1. An advantage of this feature map is that it can handle an increasing production capacity. It can also account for specific outages due to maintenance or grid-balancing decisions. In essence, the turbine map gives the model context. The grid also enables a single model to be used for modeling all the price regions. Therefore, general phenomena occurring in specific areas can be learned by the model, thus increasing the generalizability of the model. Additionally, we also notice that it speeds up training time. To the best of our knowledge, we are the first to use and create these turbine maps combined with the NWP grid. This concept could be extended in future work to use more information/context about the turbines and geospatial data, and train on more regions, by creating a foundational model for forecasting regional wind power production.

Algorithm 1 Turbine map construction

Input: $\mathcal{T}$, a set of operational turbines where ith turbine t_i has geographic coordinate (lat_i, lon_i) and maximal power capacity C_i

Input: G_{lat} and G_{lon}, both 64x64 coordinate matrix which represents respectively the latitude and longitude values of the geographic coordinates of the NWPs,

Output: M, a 64x64 matrix which represents the turbine map

$M \leftarrow 0$ ▷ *Initiate a 64x64 zero-matrix*

for all $i \in \{\mathcal{T}_1, ..., \mathcal{T}_k\}$ **do**

  $D_i = distance\{(G_{lat}, G_{lon}), (lat_i, lon_i)\}$ ▷ *Calculate a 64x64 distance matrix that represents the distance between G and* $\mathcal{T}_i$ (lat_i, lon_i)

  $\{(x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), (x_{i,3}, y_{i,3}), (x_{i,4}, y_{i,4})\} = \text{argsort}(D_i)[:4]$ ▷ *Find the four indexes with the smallest values (distance to turbine* $\mathcal{T}_i$*)*

  $scale_i = \sum_j^4 D_i[x_{i,j}, y_{i,j}]$

  $M[x_{i,1}, y_{i,1}] = M[x_{i,1}, y_{i,1}] + \frac{1}{3} C_i (1 - \frac{D_i[x_{i,1}, y_{i,1}]}{scale_i})$

  $M[x_{i,2}, y_{i,2}] = M[x_{i,2}, y_{i,2}] + \frac{1}{3} C_i (1 - \frac{D_i[x_{i,2}, y_{i,2}]}{scale_i})$

  $M[x_{i,3}, y_{i,3}] = M[x_{i,3}, y_{i,3}] + \frac{1}{3} C_i (1 - \frac{D_i[x_{i,3}, y_{i,3}]}{scale_i})$

  $M[x_{i,4}, y_{i,4}] = M[x_{i,4}, y_{i,4}] + \frac{1}{3} C_i (1 - \frac{D_i[x_{i,4}, y_{i,4}]}{scale_i})$

end for

4.2. Deterministic forecasting model

A CNN architecture, traditionally used in computer vision tasks, is proposed as the final deterministic model that uses the grid of two NWP variables, wind and wind gust speed, extended with the turbine map. A single model is used for modeling all the price regions.

CNNs have several interesting characteristics for regional wind power forecasting. First, CNNs usually have sparse interactions using kernels smaller than the input [43]. Consequently, CNN models need

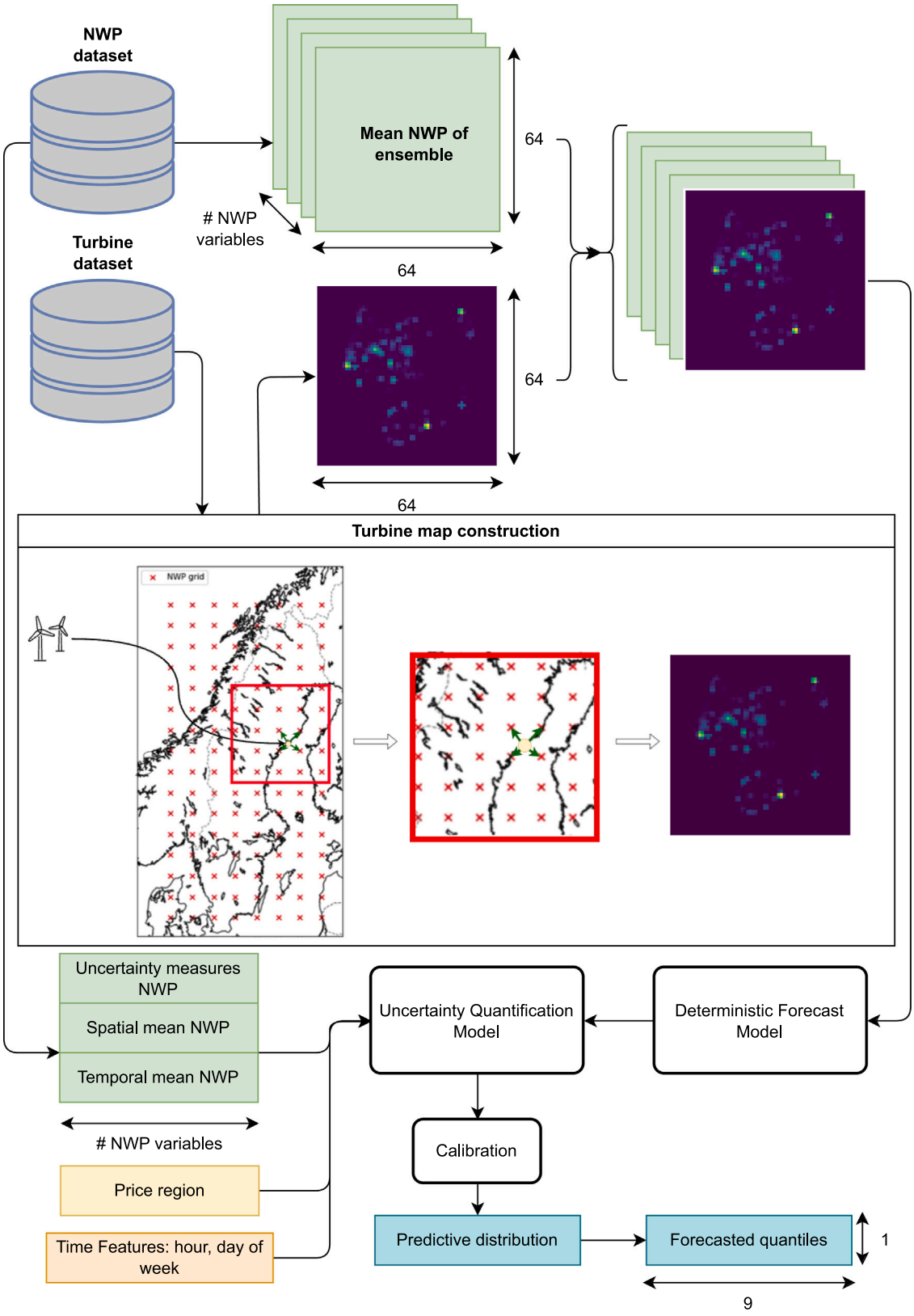


Fig. 4. Proposed regional day-ahead probabilistic forecasting framework.

to store fewer parameters. It also means it requires fewer operations to compute the output [43]. Second, another valuable aspect of CNN models is that they leverage parameter sharing and thus reduce the model's parameter size, which is helpful for this problem setting because of the high-dimensional input data available for forecasting. Due to parameter sharing, CNN models learn one set of parameters for

every location instead of separate sets for each location [43]. The latter aspect is not a problem because the turbine map together with a grid of NWP variables, are added as an input channel. Without the turbine map, the model would need a separate set of parameters for each location because it needs to learn which regions of the NWP grid are important. Given the highly correlated dataset, this would be a difficult task, as 8736 sometimes highly correlated samples per price region would be needed. The changing turbine mix for each price region over time makes this even more difficult. This is probably the reason, intentionally or not, why Basu et al. [24] only use a few CNN layers in the beginning and switch to dense layers further in the network to make use of a separate set of parameters for each location. Our choice for the turbine map, together with the normalization of the power time series, enables using a single model to model all the price regions.

In light of the recent rise of transformer architecture, claims are often made related to image classification that the transformer architectures are more robust, efficient, and provide better uncertainty estimates. Consequentially, one might think that since our problem setting works with 2D-dimensional data, this claim would translate to our setting. However, note that a recent thorough empirical analysis [44] states that state-of-the-art CNNs (such as ConvNext [45]) can be as reliable and robust, or even more, than state-of-the-art transformers. Nonetheless, CNNs and transformers have very different characteristics, each with its own inductive biases. Due to the self-attention mechanism and the use of patches of ViTs [46] and derivatives, they tend to focus more on shapes and curvature, thus essentially acting as low-pass filters [47]. In contrast, the convolutional layers prioritize texture and can be considered high-pass filters [48]. However, note that depending on the learned features in different layers, they can function as both high- and low-pass filters [48]. Most importantly, CNNs have an inductive bias for spatial invariance, which is beneficial in our problem setting. The ViT architecture does not exhibit this bias and requires more data than CNNs such as ResNet [49]. Therefore, Wu et al. [49] proposed an architecture combining convolution layers and vision transformers to get the best of both worlds. The MaxViT architecture [50] goes even further, introducing a new attention module named the dubbed blocked multi-axis self-attention (Max-SA), reducing the quadratic complexity of vanilla attention to linear without any loss of non-locality. The MaxViT architecture is then built by stacking alternative layers of Max-SA with MBConv in a hierarchical architecture [50]. Advantageously, MaxViT benefits from global and local receptive fields throughout the entire depth of the network. Therefore, and since the performance of the different types of architectures can be problem-dependent, we evaluated both CNN and hybrid transformer architectures. Note that the amount of practical ViT architectures and newly introduced CNN architectures for this forecasting task is limited, as most of these architectures have an enormous number of parameters and need more training data than we have to our availability.

Based on the performance results, see Section 5.2 and Table 1, the architecture of the proposed approach is an adapted version of the DenseNet architecture proposed by Huang et al. [51]. DenseNet builds on the observations by He et al. [52] that shorter connections between layers closer to the input and output enable deeper, more efficient, and consequently more accurate networks. DenseNet's approach is simple: connect all the layers directly with the same feature-map size. DenseNets combine features by concatenating them instead of adding them like in ResNets [52]; in consequence, the lth layer has l connections [51]. Hence in a L-layer network, there are $\frac{L(L+1)}{2}$ connections in contrast with L connection in a traditional network [51]. Fig. 5 shows an abstract overview of the proposed adaptation, DenseNet100-k12-BC, of the DenseNet architecture. The proposed DenseNet model uses a growth rate k equal to 12, consists of 100 layers, and uses bottleneck layers and compression [51]. The DenseNet architecture is compared to a ResNet architecture [52] that consists of three building blocks, with each a depth of 9, using bottleneck layers, and with feature maps sizes of respectively 16×16, 32×32, 64×64. Finally, the model is compared with the MaxViT architecture, which consists of three stages, each stage 2, 5, and 2 MaxViT blocks, respectively. Other hyperparameters of the MaxViT model are the following: the dimension of the convolutional stem is 16, the dimension of the first layer is 12, the dimension of the attention head is 6, the window size for the blocks and grids is 4, the expansion rate of MBConv layer is 4, the shrinkage rate of the squeeze-excitation in MBConv is 0.25, and the dropout rate is 0.2.

4.3. Uncertainty quantification and calibration component

Most machine learning regression models are primarily designed to provide accurate point predictions. However, as emphasized earlier, the reliability of these predictions is crucial for instilling trust and confidence in them from external stakeholders who rely on these forecasts. When dealing with probability forecasts, they must be well-calibrated, meaning that they provide reliable predictions. Without proper calibration, the utility of these predictions diminishes.

In this work, we propose disentangling the deterministic forecasts and uncertainty quantification. We do this by predicting the residual values of the deterministic model and the actual power output in our third component, i.e. the uncertainty quantification component. To achieve this, we fit quantile or probability density prediction models on the proper training dataset, such as Linear Quantile Regression (LQR), Quantile Gradient Boosting Tree (QGBT), and Quantile Regression Forests (QRF). As input features, we utilize a combination of both quantitative and qualitative input features. Among these features, the price region indicator is the sole static feature, while the remaining features exhibit a temporal dimension. Specifically, the temporal input features encompass the following elements:

- Normalized prediction of the deterministic model: This feature involves the prediction values obtained from the deterministic model, which have been normalized.
- Total region capacity: This dynamic feature varies over time and across distinct price regions. The capacity within the EEM20 dataset predominantly increases due to the introduction of new wind farms or the expansion of existing ones. However, it is important to note that capacity could also change by factors such as maintenance activities for specific wind turbines or strategic allocation decisions in this context.
- Hour and day-of-the-week indicator for the target prediction: This feature provides information about the specific hour and day of the week associated with the target prediction, thus capturing temporal patterns.
- Spatial mean of NWP: For each NWP variable and price region, this feature resembles the spatial mean by aggregating values across grid points. It provides insights into the overall behavior of NWP variables within a region.
- Lagged and potentially leading spatial mean of NWP: Similar to the previous feature, this attribute computes the spatial mean of NWP variables for a particular region. However, it incorporates data from previous or possibly future hours within the same NWP forecast run. This captures temporal dependencies in the NWP data. In addition, we also believe that this feature can capture icing phenomena, which can accumulate to significant power loss.
- Spatial standard deviations of NWP ensemble: This feature assesses the uncertainty associated with NWP data. Specifically, it calculates the spatial standard deviations of the predictions generated by a 10-member ensemble, offering valuable insights into the variability and reliability of NWPs.

We also evaluated the use of embeddings of the deterministic model as features for the uncertainty quantification models and found no increase in performance or reliability by including them.

In summary, our predictive modeling framework incorporates diverse input features for UQ, including static and temporal features and

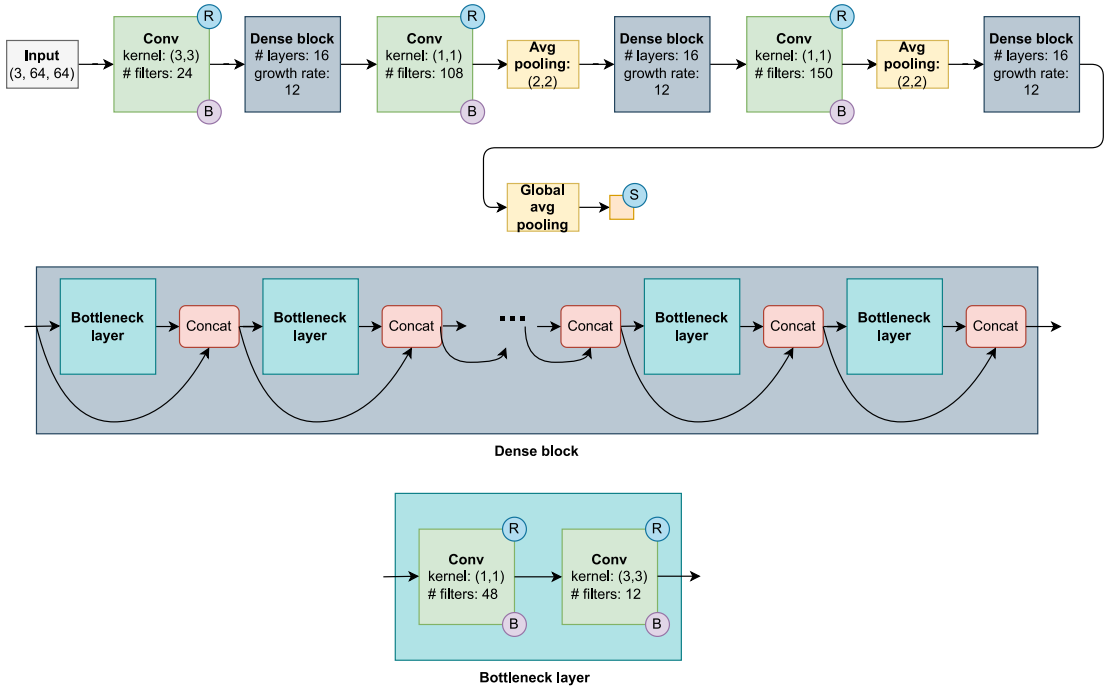


Fig. 5. DenseNet-100-BC architecture with growth rate $k = 12$, bottleneck layers, and compression in the transition layers. The green square represents a convolutional layer, the purple circle with the letter B represents a batch normalization operation, and the blue circle represents the activation function where R stands for ReLU and S for sigmoid. The batch normalization operation is always performed before the activation function and comes after the convolutional layer.

measures of spatial characteristics and uncertainty in NWP data. These features collectively enable us to enhance our UQ predictions' accuracy and robustness.

4.3.1. Conformal prediction

Conformal prediction is a model-agnostic methodology that can transform point predictions into reliable prediction intervals with non-asymptotic, distribution-free coverage guarantees under the exchangeability assumption. In practice, split or inductive conformal prediction is used, where the main idea is to split your data into a proper training set and a calibration set. The regression model is then trained on the training set, and the calibration set is used in our fourth component to calculate nonconformity scores. These scores, determined by a nonconformity measure, assess how unusual an example appears relative to the instances in the training set.

To transition from a point forecast to quantile predictions, we propose the use of signed error as a nonconformity score [53], as opposed to the more commonly used absolute error, which can only deliver asymmetric intervals, i.e., intervals where the probability mass below and above the lower and upper bound respectively are not guaranteed to be equal. Utilizing the signed errors $\alpha_i = y_i - \hat{y}_i$ of the calibration set, we sort them in ascending order. Subsequently, a specific quantile q_δ can be generated by taking the $\lfloor \delta(N_{cal} + 1) \rfloor + 1$ element of the sorted signed errors α, where N_{cal} represents the number of examples in the calibration set, and δ signifies the desired quantile. Calibrated predictions are then derived by adding q_δ to $\hat{y}_i$.

While this approach results in calibrated quantile predictions, it provides only marginal coverage, often failing to meet expectations. For instance, a 25% quantile prediction ($y < q_{0.25}$) may have 100% coverage in one price region (e.g., SE1) and 0% coverage in others, resulting in an overall marginal coverage of 25%. However, we desire conditional coverage to provide coverage guarantees for different strata.

An intuitive, simple, and effective approach is the Mondrian conformal predictor [54], which divides the example space Z with a measurable function κ that assigns each $z \in Z$ to its category k. The same procedure as described earlier can then be applied to each category. An example of a possible category is the price region in our setting. However, one could also categorize (bin) the point prediction of the deterministic model; this approach is referred to as Mondrian conformal regression [55].

Another approach for generating adaptive prediction intervals is Conformalized Quantile Regression (CQR) proposed by Romano et al. [41], which fuses quantile regression methods with a conformal way of thinking to get valid distribution-free prediction intervals, inheriting advantages of both approaches resulting in variable-width conformal prediction intervals. Simulation studies have demonstrated that CQR outperforms standard classical approaches regarding efficiency (a.k.a. sharpness) [41].

The CQR approach closely resembles the standard induction conformal prediction approach. First, the data is divided into proper training and calibration sets. The proper training set is used to fit a quantile regression model, while the calibration set is used to compute nonconformity scores α_δ for each desired quantile with level δ. These scores are used to derive an adjustment term to calibrate the quantile regression model, with the nonconformity score being the signed error between the target value and quantile prediction ($\alpha_{\delta,i} = y_i - \hat{y}_{\delta,i}$). The adjustment term is then obtained by calculating the δth quantile of the errors of the fitted model on the calibration set, $Q_\delta(\alpha_\delta)$. Calibrated quantile predictions can then be generated by adding $Q_\delta(\alpha_\delta)$ to $\hat{y}_{\delta,i}$.

One notable advantage of CQR is its ability to calibrate various quantile regression machine learning methods, including those known for delivering poorly calibrated predictions, such as neural networks, Gradient Boosting Machines (GBM), QRF, and LQR.

One of the problems with both split conformal prediction and CQR is that when the signed error is used as a nonconformity measure, it is assumed that the response variable y has an unbounded domain, i.e., $y \in \mathbb{R}$. To resolve this issue, we propose in this work the logit-nonconformity measure: $\alpha_i = logit(y_i) - logit(\hat{y}_i)$. A calibrated quantile prediction in the case of CQR is then generated by

$$\hat{q}_{\delta,i} = expit(logit(\hat{y}_{\delta,i}) + Q_\delta(\alpha_\delta)) \tag{2}$$

Simply put, we transform the domain of the response variable from a bounded to an unbounded domain, apply conformal prediction, and transform the predicted intervals or quantiles back to the original domain.

4.3.2. Conformal predictive systems

As previously discussed, calibrated probability density predictions are more desirable as they offer more insights and possibilities in decision-making processes. We propose using a Split Conformal Predictive System (SCPS) [40], modifying conformal predictors that output probability distributions instead of prediction sets as in conformal prediction. The SCPS has many parallels with inductive conformal prediction; instead of using the calibration scores to create these adjustments for a specific predefined prediction interval or quantile, we keep them to create a conformal predictive distribution. CPS uses the sorted (ascending order) calibration scores $C_1, \dots, C_{N_{cal}}$ with N_{cal} the number of examples in the calibration set, and set $C_0 = -\infty$ and $C_{N_{cal}+1} = \infty$. Given a test object x, we can generate a prediction $\hat{y}$ from the deterministic model and return a predictive distribution:

$$Q(y,\phi) = \begin{cases} \frac{i+\phi}{N_{cal}+1} & \text{if } y \in (\hat{y}+C_i, \hat{y}+C_{i+1}) \text{ for } i \in \{0,\dots,N_{cal}\} \\ \frac{i'-1+(i''-i'+2)\phi}{N_{cal}+1} & \text{if } y = \hat{y}+C_i \text{ for } i \in \{0,\dots,N_{cal}\} \end{cases}$$

where $i'' = \min\{m|C_m = C_i\}$, $i\varepsilon = \max\{m|C_m = C_i\}$, and ϕ is a random number which follows a Uniform distribution between 0 and 1. However, as recently pointed out in the literature [40,56–58], this approach has the same problem as the standard conformal prediction intervals, they are not adaptive. Therefore, extensions of the standard SCPS are proposed. Mondrian conformal predictive distribution [56] applies the same methodology of Mondrian conformal predictors [54,55] to SCPS. Another approach is the Conformal Predictive Distribution Tree (CPDT), which is, in essence, a single tree where each leaf contains a conformal predictive distribution, giving more adaptive prediction distribution while being interpretable due to the decision tree.

Although these more adaptive methods allow efficiency increases compared to the standard approach, they still do not provide the adaptiveness that CQR gives to conformal prediction intervals. Applying the framework of CPS to calibrate the quantile regression forest, as proposed by Wang et al. [57], is a step towards more adaptive conformal prediction distributions. They suggest calibrating QRF with CPS, where each tree's out-of-bag (OOB) samples are seen as a calibration set, see Algorithm 2. Note that this proposed algorithm is not proven to be valid in theory, and out of the evaluation on 20 public datasets, it is shown that there were some deviations (around 2.5%) compared to ideal coverage and to the standard SCPS method, which has proven non-asymptotic coverage guarantees. The proposed algorithm also has a lot of similarities with the jackknife method in conformal prediction, which has a fragile out-of-sample coverage and only has asymptotic coverage properties with non-trivial conditions on the base estimator [59,60].

Algorithm 2 Calibrating QRF with CPS based on OOB and weighted samples. (Wang et al. [57])

Input: Training set $Z^N = \{(X_1, y_1), ..., (X_N, y_N)\}$
Input: Test object x_0
Output: Predictive distribution Y_0

1: Fit QRF on the training set Z^N
2: Calculate $C_i = \hat{F}^{(i)}_{x_i}$ for $\{i \in \{1, ..., N\}$ ▷ *Let $\hat{F}^{(i)}_{x_i}$ be the probabilistic prediction based on all trees whose OOB samples includes ith training data.*
3: For $j \in \{1, ..., N\}$, calculate α_j quantile of $\hat{F}^{(0)}_{x_0}$, which we denote as q_j: $\hat{F}^{(0)^{-1}}_{x_0}(\alpha_j) = q_j$
4: Return the weighted empirical distribution of $\{q_1, ..., q_l\}$ whose weights are $\{w_1(x_0), ..., w_N(x_0)\}$ where w_i is the fraction of trees where x_i is in the rectangular subspace of a leaf containing x_0.

4.3.3. Split Conformal Distribution Regression Forest (SCDRF)

In this work, we additionally propose the Split Conformal Distribution Regression Forest (SCDRF), a novel algorithm that combines the philosophy of QRF [20] with CPS to generate adaptive conformal predictive distributions, see Algorithm 3. The idea is to create all trees with the proper training dataset and generate conformalized distributions in each of the rectangular subspaces of each tree. These conformalized distributions are constructed with examples in the calibration set in the same rectangular subspace. Given a specific example x_0, we get N_{tree} predictive distributions; these distributions are then averaged to get the calibrated predictive distribution. We hypothesize that this method could deliver adaptive calibrated predictive distributions. We want to point out that the resulting trees are called honest trees, as they do not re-use target values y_i for both determining split points and for making predictions. This property has been required to prove the consistency of random forests in specific settings [61,62], and thus could be of valuable use in establishing (non)-asymptotic coverage guarantees. Note that the discussion and proof of (non)-asymptotic coverage guarantees are left as future work. Nonetheless, this innovative approach marks a significant advancement in tackling the adaptiveness challenges encountered by the SCPS and the validity concerns of some of the alternative adaptive SCPS methods. Notably, a key advantage over the proposal by Wang et al. [57] is that SCDRF is model-agnostic, allowing its integration with any deterministic model.

Algorithm 3 Split Conformal Distribution Regression Forest (SCDRF)

Input: Proper training set $Z^{N_{train}} = \{(X_1, y_1), ..., (X_N, y_{N_{train}})\}$
Input: Calibration set $Z^{N_{cal}} = \{(X_1, y_1), ..., (X_N, y_{N_{cal}})\}$
Input: Test object x_0
Input: A generic regression algorithm $\mathcal{A}$
Output: Predictive distribution Y_0

1: Fit $Z^{N_{train}}$ on the generic regression algorithm $\mathcal{A}$
2: Infer the training residuals from $\mathcal{A}$: $\{y_1 - \hat{y}_1, ..., y_{N_{train}} - \hat{y}_{N_{train}}\}$
3: Fit Random Forest on the proper training set $Z^{N_{train}}$ targeting the training residuals.
4: For every leaf in each tree, compute the calibration scores (errors) of the calibration set examples that belong to that leaf, i.e. rectangular subspace. ▷ *In essence, a conformal predictive distribution is formed in every leaf for each tree.*
5: Return the distribution function of x_0 by averaging the conformal predictive distributions of every tree i rectangular subspace that x_0 belongs $Q_i(y|X = x_0)$:

$$\hat{F}(y|X = x_0) = \sum_{i=1}^{n_{trees}} Q_i(y|X = x_0)$$

5. Results and discussion

The benefit of our framework is that we can easily exchange different components and thus optimize every part separately. We will first discuss our point forecast, the deterministic model, and afterward, the quantile forecasting models with different calibration methods. Note that we also evaluate disentangling the deterministic and uncertainty quantification by optimizing a quantile CNN using the same architecture as the deterministic model but with the critical difference being the output of nine quantiles optimized by the pinball loss function, see Eq. (4).

5.1. Data splitting

To evaluate the different architectures, we train the models on the data from the year 2000 and assess them on the data from 2001. From this training dataset, we hold out a validation set that we use for an

Table 1
Overview of deterministic forecasting results. Best results are highlighted in bold.

	DenseNet		ResNet		MaxViT	
	MAE (MW)	NMAE	MAE (MW)	NMAE	MAE (MW)	NMAE
Task 1: Jan–Feb 2001	**121.98**	0.06823	124.73	0.06859	122.71	**0.06802**
Task 2: Mar–Apr 2001	**117.00**	**0.06381**	125.66	0.06771	127.04	0.06908
Task 3: May–Jun 2001	**101.47**	**0.05573**	104.19	0.05690	104.86	0.05721
Task 4: Jul–Aug 2001	**91.99**	**0.05056**	94.56	0.05162	94.63	0.05247
Task 5: Sep–Oct 2001	**110.41**	**0.05657**	114.48	0.05778	113.42	0.05754
Task 6: Nov–Dec 2001	**132.89**	**0.06327**	138.19	0.06607	130.80	0.06338
Tasks 1–6: Jan–Dec 2001	**112.62**	**0.05970**	116.97	0.06145	115.58	0.06128
Validation set	**82.80**	**0.05324**	83.69	0.05325	85.54	0.05444

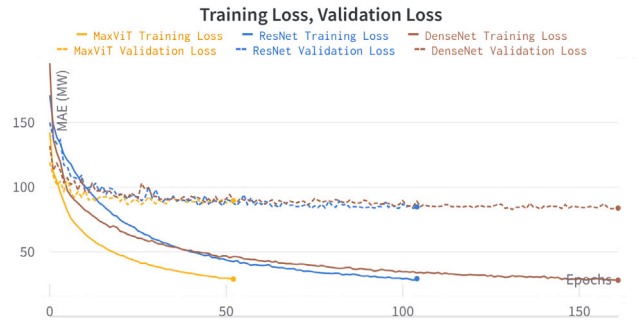


Fig. 6. Learning and validation curve of DenseNet, ResNet, and MaxVit model, trained with patience of 30 epochs for early stopping.

early-stopping procedure and calibration. This validation/calibration set is the same in all training runs for the different architectures and calibration approaches.

5.2. Deterministic models

For training both the CNN and ViT models, a batch size of 32, a learning rate of 1e−4, and an AdamW [63] optimizer are used. The models are optimized with the L1-loss, which optimizes the mean absolute error (MAE). This loss function is chosen as it is closely related to the pinball (quantile) loss function; 50% quantile loss is the same as the L1-loss function except for a constant factor since the L1 loss function optimizes to estimate the median. With a consistent size of the training datasets, we noted comparable training times for the three deterministic models, ranging from 6 to 12 h, with an increase corresponding to the size of the training set. On the other hand, performing inference for a 24-h period requires a few seconds (CPU only). Consequently, since the uncertainty quantification can almost be inferred instantly, this timeframe is more than sufficient for taking action based on the generated forecasts. In Fig. 6, we observe a large generalization gap, i.e., the gap between the training loss and validation loss, which we observe with all evaluated architectures. This is probably due to an unrepresentative training dataset and indicates that the model needs more training data. Table 1 shows the results of our experiments and indicates that the DenseNet architecture is a superior model for the point forecast compared to the ResNet and MaxViT architectures.

5.3. Uncertainty quantification and calibration

In this study, we evaluate the performance of CQR, Mondrian conformal prediction, and standard inductive conformal prediction for uncertainty quantification and calibration. For CQR, and as an uncalibrated reference, we assess a range of quantile forecasting methods to identify those best suited for this application. Specifically, we evaluate linear quantile regression (LQR), quantile gradient boosting trees (QGBT), and quantile random forests (QRF). More specifically, for the QGBT, we use the XGBoost boosting system (QXGB) [64]. Additionally, we explore a model that combines quantile and deterministic forecasting models, using the same architecture as the deterministic model but with the critical difference being the output of nine quantiles optimized by the pinball loss function, see Eq. (4). Additionally, we compare these conformal prediction methods with some of the discussed CPS approaches, such as standard SCPS, Mondrian SCPS, and SCDRF. Finally, we also compare the quality and reliability of the probability prediction distribution of the CPS approaches.

5.3.1. Evaluation metrics

The EEM20 competition evaluated different approaches using the average pinball loss function [65], a probability forecasting skill score incorporating both the sharpness and reliability of a quantile forecast. This work uses the same pinball loss function to evaluate quantile predictions. Eq. (3) presents the pinball (or quantile) loss function for a quantile δ:

$$\rho_i^{\delta}(q_i^{\delta}, y_i) = \begin{cases} \delta(y_i - q_i^{\delta}), & y_i \geq q_i^{\delta} \\ (1-\delta)(q_i^{\delta} - y_i), & y_i < q_i^{\delta} \end{cases} \tag{3}$$

where q_i represents the predicted value for quantile δ, and y_i represents the target value. The pinball loss function can be seen as an absolute error, with an additional penalty term when the prediction error exhibits a less probable sign, e.g., when the target quantile is 10%, and the quantile estimate is larger than the observed value, the absolute error gets a weight of 0.9 instead of 0.1. It thus accounts for both the sharpness and reliability of a quantile forecast. The quantile score is also a proper scoring rule for quantile predictions and thus, therefore, fit to evaluate different quantile regression models [66]. A scoring rule is considered "proper" if, on average, it rewards forecasters for providing accurate and well-calibrated probability estimates.

Eq. (4) represents the average pinball loss function, which is the average quantile losses on the predicted quantiles for this competition $\{0.1, 0.2, \ldots, 0.9\}$.

$$\rho_i = \frac{1}{N_{\delta}} \sum_{\delta \in \{0.1, 0.2, \ldots, 0.9\}} \rho_i^{\delta}(q_i^{\delta}, y_i) \tag{4}$$

However, the loss function slightly favors sharpness over reliability, which could result, to some extent, in sharp but less reliable forecasts when the pinball loss function is used for optimization or evaluation. This is also mentioned by Browell et al. [29]. They also noted that an alternative loss function that mitigates this characteristic has not been proposed yet. Since we slightly violated some of the assumptions of conformal prediction and introduced some new, unproven methods, we do not have (non)-asymptotic coverage guarantees. Therefore, and because reliability is a first priority, we must evaluate the reliability of forecasts separately. We do this visually by inspecting reliability plots, see Fig. 7, which visualize the reliability of the different methods. These plots show the deviation from perfect reliability against the required probability. Besides visually inspecting reliability, we also evaluate it numerically by measuring the mean absolute quantile coverage error:

$$\text{MQCE} = \sum_{\delta \in \{0.1, 0.2, \ldots, 0.9\}} \left| \left\{ \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathbb{1}_{y < q_{\delta}} \right\} - \delta \right| \tag{5}$$

Table 2
Evaluation of different quantile forecast approaches, trained on the first year of the EEM20 dataset and evaluated on the following year. The scale of the pinball loss function is in MW. Each category's best evaluation scores and approaches are highlighted in bold, and the overall best are underlined.

	Task 1 PL	Task 2 PL	Task 3 PL	Task 4 PL	Task 5 PL	Task 6 PL	Average PL
SCPS	48.5232	46.7655	40.2982	36.6175	**43.6214**	52.6067	44.7387
SCPS-logit	**48.2913**	**46.0982**	**40.0217**	36.4830	43.9583	**52.3617**	**44.5357**
MCPS-region	50.3513	47.7600	40.7393	**36.3723**	43.6352	53.2351	45.3489
MCPS-region-logit	52.0054	48.8611	41.7946	37.1885	44.8860	54.1332	46.4781
MCPS-bins	50.1302	46.5313	40.6215	36.9548	44.5432	52.8059	45.2645
MCPS-bins-logit	49.7768	46.3861	40.8495	37.2072	44.7020	52.8921	45.3023
LQR	57.9338	57.6905	48.4342	41.9310	52.6327	64.0425	53.7775
CQR-LQR	49.7479	49.5741	41.4726	36.3020	**44.6193**	54.3920	46.0180
CQR-logit-LQR	**49.0481**	**48.7776**	**41.0196**	**35.8194**	44.6329	**53.7915**	**45.5148**
QXGB	55.1557	55.9889	46.3781	40.9410	49.8649	62.4181	51.7911
CQR-QXGB	49.0458	49.2761	41.0522	36.5569	44.1579	55.2759	45.8941
CQR-logit-QXGB	48.6747	49.0930	40.7072	36.0756	44.0553	54.6871	45.5488
QXGB-error	55.6263	54.9954	46.4135	40.8464	49.4838	60.3533	51.2865
CQR-QXGB-error	48.4755	47.8969	40.4447	36.1064	43.2520	52.6716	44.8078
QXGB-logit-error	56.3808	55.9365	47.0402	41.0591	49.8320	60.9383	51.8645
CQR-QXGB-logit-error	**48.1212**	**47.6141**	**40.1808**	**35.5774**	**43.0909**	**52.0231**	**44.4346**
QRF	53.4628	53.1245	44.5703	39.6153	47.9132	58.6468	49.5555
CQR-QRF	48.0505	47.5879	40.1894	36.2964	43.2967	52.7650	44.6977
CQR-logit-QRF	**47.3824**	**47.0011**	**39.6635**	**35.6540**	**42.9732**	51.9436	**44.1030**
QRF-error	55.6766	55.3945	46.7267	40.9493	49.8511	60.7073	51.5509
CQR-QRF-error	48.4300	48.2043	40.6641	36.1780	43.4708	52.9077	44.9758
QRF-logit-error	55.0435	54.7372	46.0391	40.6291	49.3081	60.0475	50.9674
CQR-QRF-logit-error	47.5852	47.1441	39.9890	35.7857	43.2272	**51.8641**	44.2659
SCDRF	47.9301	**46.8926**	**40.1462**	**35.9818**	**43.3177**	**51.6380**	**44.3178**
SCDRF-logit	**47.7515**	47.1296	40.5283	36.2793	43.7620	51.7745	44.5375
DN-PBL	57.3887	55.7400	41.7017	38.0203	47.7551	56.5764	49.5304
CQR-DN-PBL	55.7219	**53.7137**	40.3180	36.9630	46.3374	55.0844	48.0231
CQR-logit-DN-PBL	**55.3741**	52.8741	**40.2268**	**36.9544**	**46.2589**	**54.6847**	**47.7289**

To evaluate the probability density predictions, we use the continuous ranked probability score (CRPS), used in several other studies related to CPS [40,56,57,67]. If we define F as the distribution function $F : \mathbb{R} \to [0, 1]$, and y_i as the actual value, then the CRPS is defined as follows:

$$\text{CRPS}(F, y_i) = \int_{-\infty}^{\infty} (F(y) - \mathbb{1}_{y>y_i})^2 dy \tag{6}$$

Performance improves as CRPS decreases, reaching its optimal at a minimum value of 0.

5.3.2. Evaluation of quantile prediction and calibration

The results of different quantile regression methods paired with varying calibration methods are presented in Table 2. Essentially, we can divide the evaluated methods by how the initial quantile forecast is performed: *SCPS/MCPS*, *LQR*, *QXGB*, *QRF*, and DenseNets optimized with pinball loss function *DN-PBL*. We also distinguish the quantile regression approaches into settings where the target is the actual power output, the residual error $y_i - \hat{y}_i$ (denoted by *-error*) or logit residual error $logit(y_i) - logit(\hat{y}_i)$ (denoted by *-logit-error*). Finally, for the conformal approaches, we also evaluated the use of two nonconformity measures, one with a regular signed error as a nonconformity measure and another one that uses our proposed logit nonconformity measure (denoted by *-logit*).

We generally observe that all uncalibrated regression algorithms are the worst-performing approaches due to overconfident predictions; see Fig. 7. There are two main reasons for this: first, these algorithms do not guarantee calibrated quantile predictions, and second, the large generalization gap of the deterministic model causes the quantile regression algorithms to overfit the problem, resulting in overconfident uncertainty predictions. This observation stresses the importance of calibration quantile regression approaches, certainly, if these approaches have a tendency to overfit, such as neural networks and gradient boosting.

Across all quantile regression algorithms with CQR (denoted by *CQR-*), the logit nonconformity measure increased efficiency compared to the classical residual error while observing roughly the same empirical coverage. For the SCPS, we observed the same phenomena; however, we observed the opposite for the MCPS and SCDRF. In future work, evaluating this logit-error nonconformity measure on different datasets with a target value with a bounded domain would be interesting.

The split (SCPS) and Mondrian (MCPS) conformal prediction systems approaches use the calibration set to generate prediction distribution from which quantile predictions can be deduced. For the MCPS, we consider two variants, one that considers the price region as a category *MCPS-region* and another where the binned categories of the deterministic prediction are considered as category *MCPS-bins*. The SCPS approach outperforms the MCPS approach across all tasks. The primary reason for this is that SCPS predictions show better calibration; see Table 3. We attribute this to a larger calibration set available for SCPS due to the grouping in MCPS. Additionally, we see that by focusing on just the price region with *MCPS-region*, we observe substantial coverage errors in price region SE1, see Fig. 8, where the capacity doubles during the testing period, indicating difficulties handling such distributional shifts in the data.

Using the CQR with QXGB and QRF improved the performance compared to the marginal coverage approaches, SCPS and MCSP. This gain in performance can be entirely attributed to the adaptability of uncertainty quantification since we observe an increase in sharpness while the coverage error only slightly increases. The increase in MQCE compared to *SCPS-logit* is only 0.6% and 1.8% for respectively *CQR-logit-QRF* and *CQR-QXGB-logit-error*.

The other quantile regression methods, LQR and DN-PB, underperformed compared to the SCPS approaches. For LQR, the primary reason for this underperformance can be attributed to uncalibrated predictions, even for the approaches that used CQR for calibration; see Table 3 and Fig. 7. For DenseNets optimized with a pinball loss function

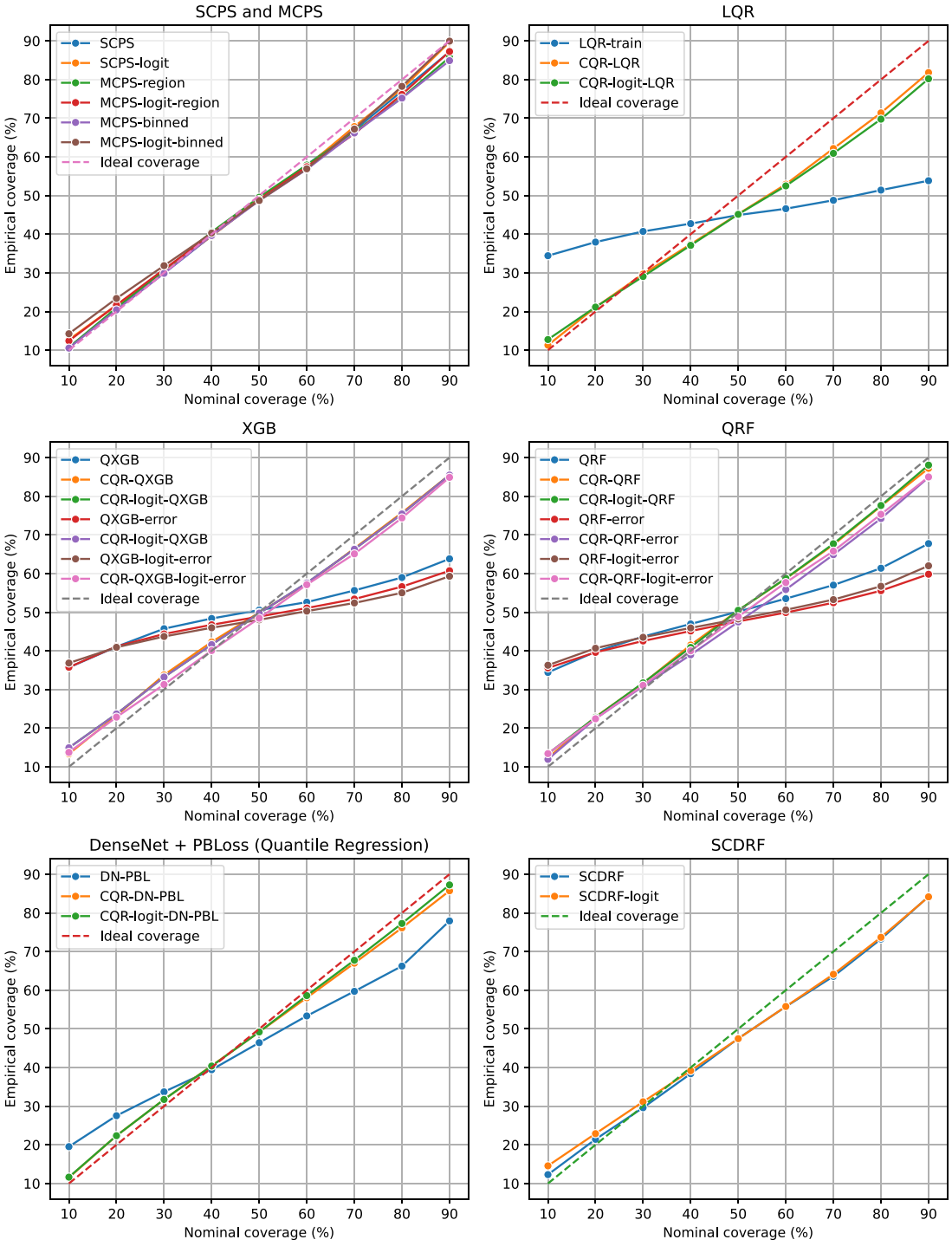


Fig. 7. Calibration plots, which plot the empirical coverage against the nominal/expected coverage of different quantile regression approaches.

that outputs quantile predictions, this was not the case since *CQR-DN-PBL* and *CQR-logit-DN-PBL* output highly reliable predictions; however, they are not as efficient since they are too wide compared to the other approaches. The primary reason for this is that during training, the model overfits way faster when we optimize for nine quantile values than if we optimize for the median value.

We also evaluated QRF and QXGB using different target values in combination with CQR for calibration. We observe that using the error and, more specifically, the logit error as the target was the performing approach for QXGB. However, for QRF, we found that targeting the actual power output was slightly better. A possible reason that QXGB benefits from this approach is that because of distracting the most influential feature, the prediction of the deterministic model, the QXGB will give more importance to other features, which helps generalizability.

Finally, we observe that our newly introduced CPS method *SCDRF* shows promising results; only QRF with CQR as calibration performed slightly better on the pinball loss. However, because the SCDRF method can produce an entire prediction distribution, see Fig. 9, instead of just quantile predictions, this approach is way more valuable as a tool in a decision-making system/framework. Note that the coverage error of the method is worse than the CQR or SCPS approaches; however, it is not on a scale that they become unreliable.

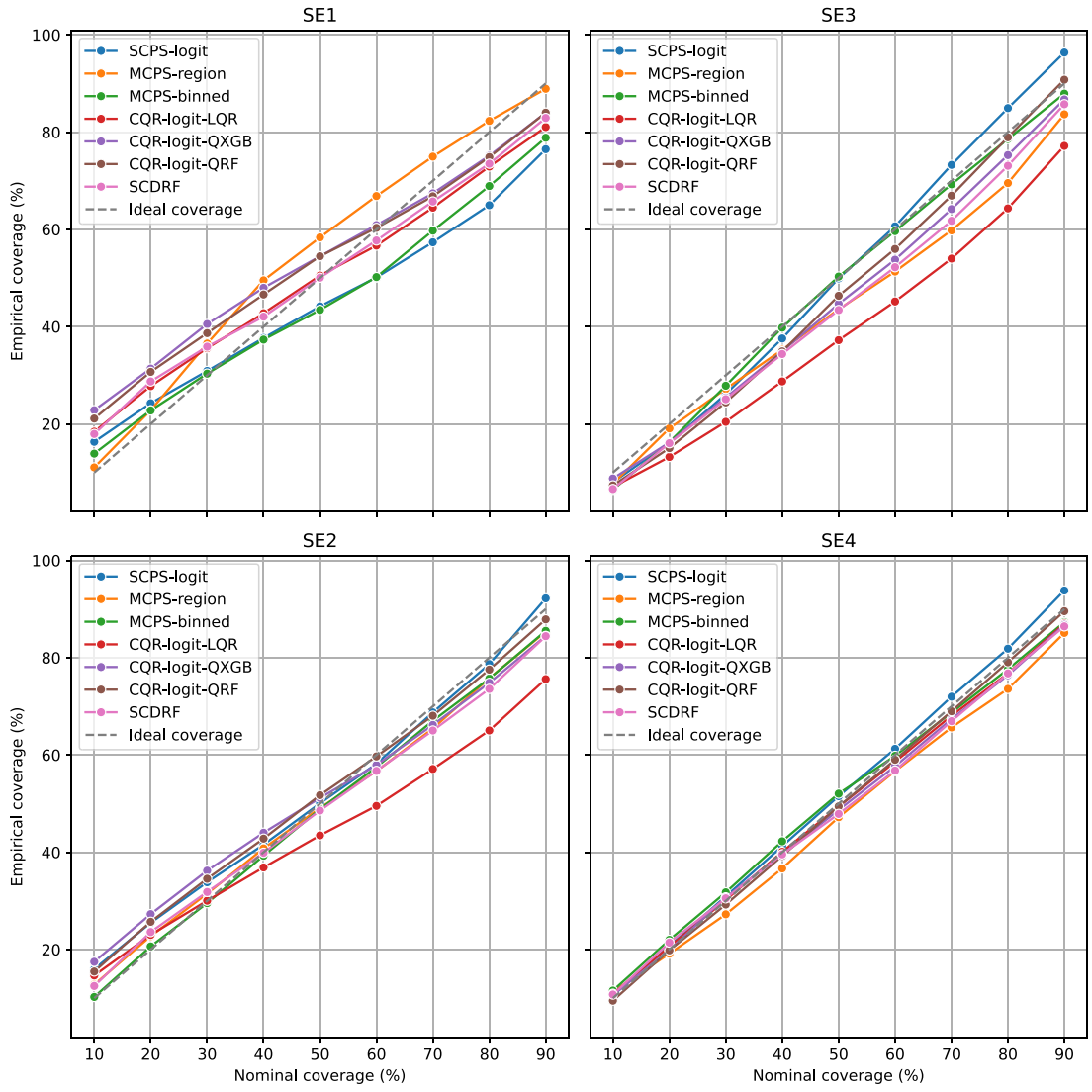


Fig. 8. Calibration plots for each price region, plots the empirical coverage in each price region against the nominal/expected coverage of different quantile regression approaches.

5.3.3. Evaluation of probability density forecasts

The quality of the probability density forecasts of the different CPS approaches is evaluated by the CRPSs shown in Table 4. Based on this score and the pinball loss, we can conclude that SCDRF approaches result in more efficient density forecasts while remaining reliable. This better efficiency, i.e., sharpness, is also clearly illustrated in Figs. 10(a) and 10(b), where the forecasted intervals of both approaches during a specific period are depicted. The SCRPs intervals are visually wider than the SCDRFs.

5.4. Comparison with state-of-the-art

Based on the results above, the best approaches for quantile and distributional forecasting, respectively DenseNet100-k12-BC with *CQR-logit-QRF* and *SCDRF* is trained six times with an iteration expanding dataset to replicate the EEM20 Wind forecasting competition setting. We compare both proposed models with the three top-performing models in the competition [24,25,29]. We refer to Table 5 for all results.

Our model, independent of the uncertainty quantification approach, outperforms all three state-of-the-art models on the average pinball score, with our proposed model, using *CQR-logit-QRF*, achieving an average pinball score of 44.10 MW. Expanding the training data reduced the pinball loss function by 1.41%. The proposed model outperforms the current state-of-the-art on the Swedish dataset, a QRF model with physics-inspired features, proposed by Bellinguer et al. [25], by 6.86%. Bellinguer et al. [25] achieved a pinball loss of 46.68 MW. Only on tasks 3 and 4, which consist of power time series from May till August 2001, the model proposed by Bellinguer et al. outperforms the proposed model. However, our proposed model performs better on all other tasks and has less variance between tasks and, thus, between seasons, which is a desirable property.

The DenseNet100-k12-BC with *CQR-logit-QRF* decreases the pinball loss by 19.45% compared to the model by Basu et al. [24], which also uses a CNN model. Besides a CNN, Basu et al. used physics-inspired input features and Monte Carlo simulations to get a probability forecast.

If we look at all the different approaches we evaluated in this work, we observe that all calibrated methods, except *DN-PBL*, outperform the state-of-the-art. This highlights the success of using deep and densely connected neural networks with turbine maps, on the one hand, making the networks context-aware and conformal prediction (systems), on the other hand, which results in reliable uncertainty quantification.

6. Conclusion

In this work, a day-ahead regional wind power forecasting framework is proposed, which provides a deterministic forecast and calibrated quantile and density forecasts that quantify the uncertainty of

Table 3
Evaluation of different quantile forecast approaches, trained on the first year of the EEM20 dataset and evaluated on the following year. Each category's best evaluation scores and approaches are highlighted in bold, and the overall best are underlined.

	Task 1 MQCE	Task 2 MQCE	Task 3 MQCE	Task 4 MQCE	Task 5 MQCE	Task 6 MQCE	Tasks 1–6 MQCE
SCPS	0.05481	0.03291	0.04159	**0.01651**	0.02055	0.04479	0.01745
SCPS-logit	0.05630	**0.02130**	**0.03520**	0.02229	0.02121	0.04935	**0.01320**
MCPS-region	0.06256	0.03907	0.04358	0.01928	**0.01838**	0.05224	0.02519
MCPS-region-logit	**0.04156**	0.02424	0.04349	0.02749	0.02286	**0.04440**	0.02106
MCPS-bins	0.06554	0.02731	0.05871	0.04214	0.02956	0.05631	0.02229
MCPS-bins-logit	0.07566	0.02322	0.03853	0.02223	0.01977	0.06525	0.02207
LQR	0.17908	0.18243	0.17834	0.16530	0.18607	0.18040	0.17808
CQR-LQR	0.05685	0.06638	**0.06354**	**0.03011**	**0.06850**	**0.05208**	**0.04643**
CQR-logit-LQR	**0.04247**	**0.06425**	0.07039	0.04190	0.07927	0.05794	**0.05468**
QXGB	0.16049	0.16540	0.15616	0.15173	0.15344	0.16526	0.15603
CQR-QXGB	0.04781	0.05064	0.02993	0.01787	**0.02591**	0.04891	**0.03086**
CQR-logit-QXGB	**0.03294**	0.04872	**0.03555**	0.03267	0.02846	0.04933	0.03215
QXGB-error	0.17308	0.17223	0.16748	0.15437	0.16130	0.17004	0.16371
CQR-QXGB-error	0.05102	0.04850	0.04742	**0.01322**	0.03121	**0.04432**	0.03215
QXGB-logit-error	0.17626	0.17712	0.17352	0.16063	0.16617	0.17509	0.16937
CQR-QXGB-logit-error	0.03777	**0.04811**	0.05197	0.02639	0.03745	0.04538	0.03115
QRF	0.15260	0.15025	0.14162	0.12561	0.13582	0.15124	0.13925
CQR-QRF	0.04763	0.03818	0.03575	**0.01096**	**0.01929**	0.04420	0.01999
CQR-logit-QRF	0.03610	**0.03568**	**0.03525**	0.01176	0.02032	0.04855	**0.01908**
QRF-error	0.17281	0.17230	0.16989	0.15321	0.16382	0.16397	0.16392
CQR-QRF-error	0.05116	0.05668	0.05436	0.02176	0.04038	**0.04144**	0.03214
QRF-logit-error	0.16981	0.17103	0.16456	0.15397	0.15691	0.16959	0.16175
CQR-QRF-logit-error	**0.03555**	0.04547	0.05076	0.02457	0.03367	0.04564	0.02701
SCDRF	0.04514	**0.05431**	0.06914	0.03983	0.04814	**0.03910**	**0.03492**
SCDRF-logit	**0.04265**	0.05691	**0.06183**	**0.03376**	**0.04153**	0.04865	0.03803
DN-PBL	0.07754	0.09018	0.07903	0.07153	0.07606	0.08405	0.02223
CQR-DN-PBL	0.03875	0.03906	0.03296	**0.01315**	0.02994	**0.04995**	0.02223
CQR-logit-DN-PBL	**0.02857**	**0.03132**	**0.02985**	0.01773	**0.02505**	0.05132	**0.01783**

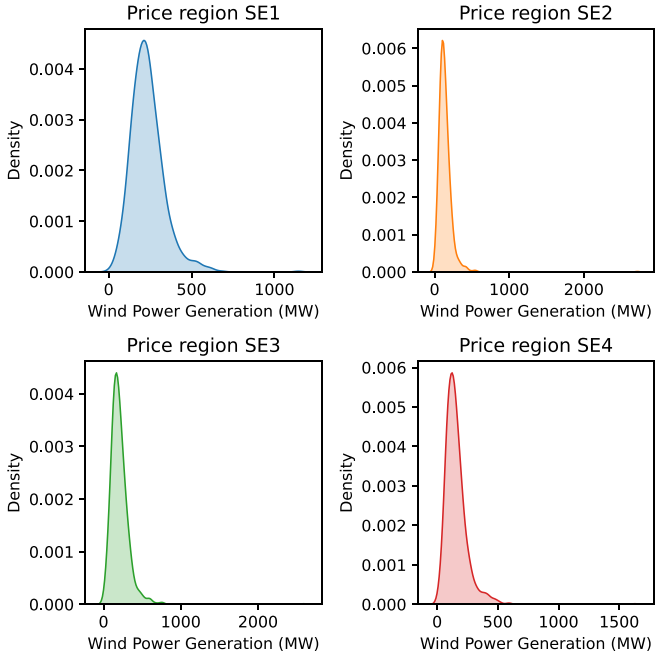


Fig. 9. Prediction density forecast of SCDRF on the 5th of October 2001.

the prediction. The approach consists of a deep and dense convolutional neural network that uses as input a grid of NWP variables that cover the entire region together with a specially constructed feature map that denotes the locations of all the turbines in a specific region. This feature map enables a deep CNN model and allows the model to be trained on multiple regions. In addition, the map allows the model to handle a

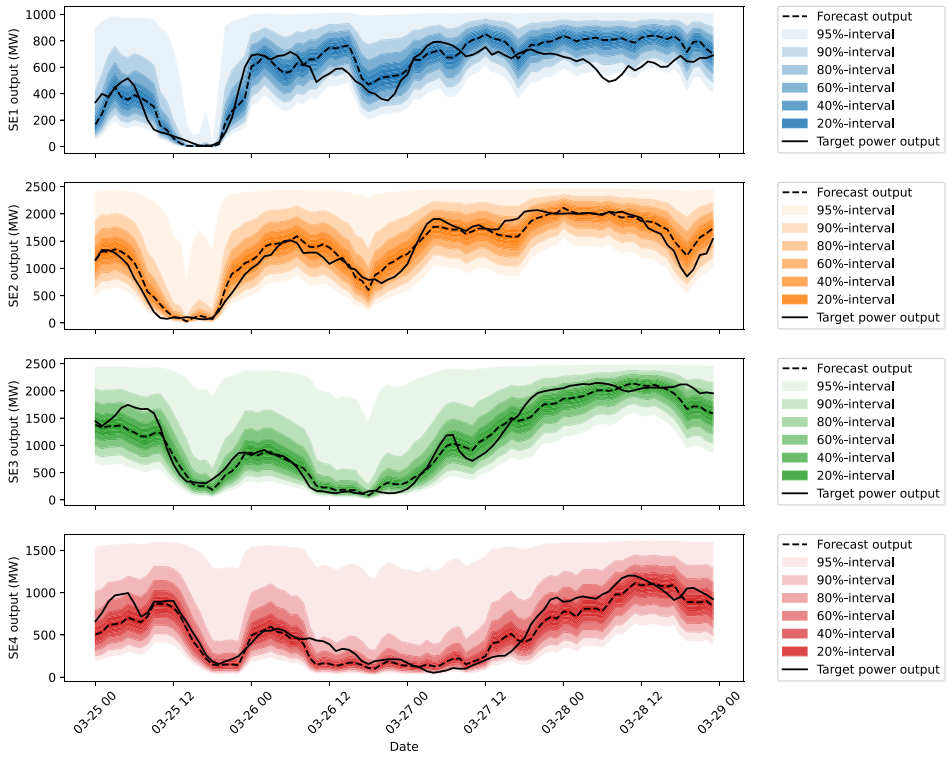
Table 4
Continuous Ranked Probability Score (CRPS) of different quantile forecast approaches, trained on the first year of the EEM20 dataset and evaluated on the following year. Each category's best evaluation scores and approaches are highlighted in bold, and the overall best are underlined.

	CRPS
SCPS	**0.04349**
SCPS-logit	0.04384
MCPS-region	0.04409
MCPS-region-logit	0.04653
MCPS-bins	0.04426
MCPS-bins-logit	0.04511
SCDRF	**0.04299**
SCDRF-logit	0.04339

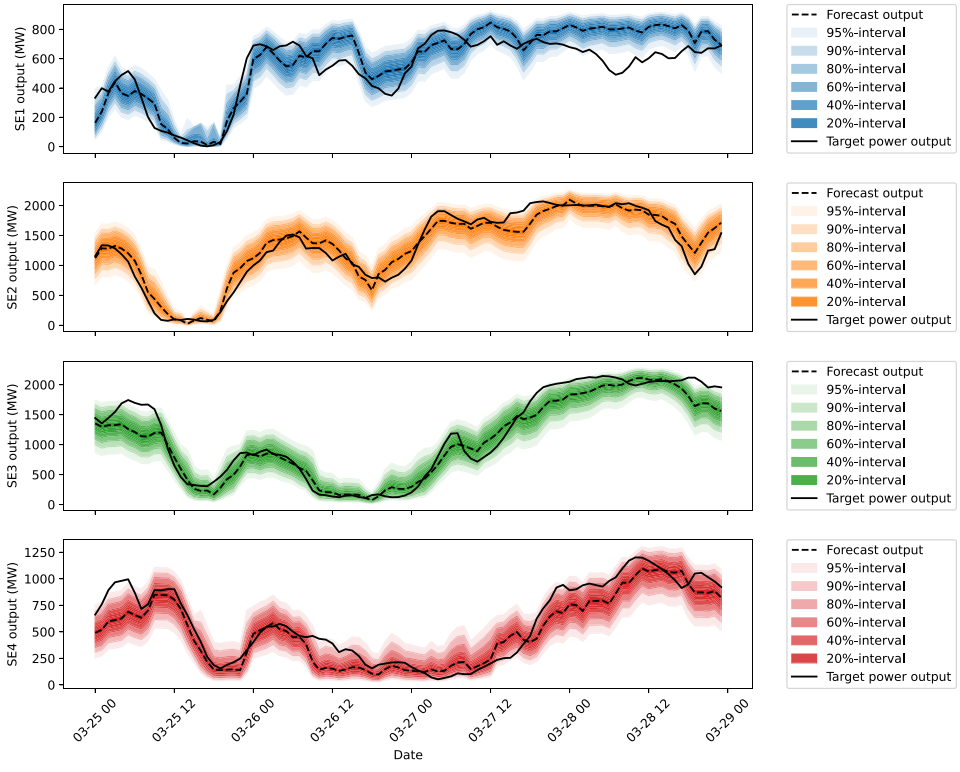
changing turbine mix due to capacity increases, maintenance outages, or grid-balancing decisions. One could say that the turbine map makes the network context-aware. To our knowledge, we are the first to use and create these turbine maps combined with the NWP grid.

We also show the importance of model calibration and that quantile conformal prediction and CPS can be valuable frameworks to achieve this. We successfully introduced a new adaptive CPS, Split Conformal Distribution Regression Forests (SCDRF), which allows for generating calibrated prediction distributions by combining the CPS with the philosophy behind QRF.

Our presented approach was evaluated on the EEM20 dataset, a Swedish dataset for regional forecasting, and compared with the state-of-the-art on that dataset. The proposed model decreases the pinball loss function by 6.86% compared to the best-performing model in the literature. This work shows that deep and dense CNN models with context-aware features are better than tree-based ensemble methods to cope with the large dimensionality of NWP and turbine data. Together with Conformalized Quantile Regression (CQR) and QRF for uncertainty quantification, it shows state-of-the-art performance for the regional probability forecasting task while ensuring reliability.



(a) SCPS-logit



(b) SCDRF

Fig. 10. Symmetric prediction intervals between 25/03/2001 and 28/03/2001. Symmetric prediction intervals mean that the same amount of probability mass lies above the upper bound as below the lower bound of the interval.

7. Future work

Although the proposed approach in this paper surpasses the current state-of-the-art, we believe there is still some untapped potential for this problem. We hypothesize that the deterministic forecast capabilities could be improved by using inputs (NWP grid, turbine map) of variable size so that for price regions SE1 and SE4, which cover a smaller region, the model can focus more on the grid points of the turbine locations

Table 5
Comparison of proposed CPS-QRF and CQP-QRF-logit for quantile forecasts against state-of-the-art on the Swedish EEM20 dataset. The scale of the pinball loss function is in MW and the best results are highlighted in bold.

	Task 1 PL	Task 2 PL	Task 3 PL	Task 4 PL	Task 5 PL	Task 6 PL	Average PL
CNN + MC sim. [24]	57.17	58.32	48.38	41.96	51.77	66.28	53.98
Quantile GBM [29]	66.71	53.84	42.53	34.31	46.18	62.81	51.06
QRF [25]	58.36	52.11	**37.56**	**33.07**	43.03	55.97	46.68
CQR-logit-QRF (proposed)	**47.38**	**47.02**	38.78	33.84	**41.63**	52.22	**43.48**
SCDRF (proposed)	47.93	48.23	39.39	33.86	42.07	**52.11**	43.93

by downsizing the grids. Also, it would be interesting to add more feature maps containing location-specific data like terrain information. It would also be interesting to compare input features as grid-like NWPs versus the NWPs per wind farm in future work. At the time of writing, all works related to regional forecasting, to our knowledge, do one or the other. A comparison of the same regional power times series of the two types of features would be an interesting and valuable addition to the literature. Finally, it would be interesting to research further the characteristics and performance of the newly introduced SCDRF approach and the use of the logit-nonconformity measure, for other datasets and use cases.

CRediT authorship contribution statement

Jef Jonkers: Writing – original draft, Validation, Software, Methodology, Conceptualization. **Diego Nieves Avendano:** Conceptualization, Writing – review & editing. **Glenn Van Wallendael:** Conceptualization, Supervision, Writing – review & editing. **Sofie Van Hoecke:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

The EEM20 dataset is openly available here.

Acknowledgments

Part of this work was funded by the Flanders AI Research Program, Belgium and the COOCK Smart Ports 2025 project. COOCK Smart Ports 2025 is a project financed by VLAIO, Belgium that aims to transfer knowledge via pilot projects, whereas the Flanders AI Research Program is financed by the Flemish Government, Belgium.

References

[1] Lew D, Milligan M, Jordan G, Piwko R. Value of wind power forecasting. Tech. rep. NREL/CP-5500-50814, Golden, CO (United States): National Renewable Energy Lab. (NREL); 2011, URL https://www.osti.gov/biblio/1011280.

[2] Jung J, Broadwater RP. Current status and future advances for wind speed and power forecasting. Renew Sustain Energy Rev 2014;31:762–77. http://dx.doi.org/10.1016/j.rser.2013.12.054, URL https://linkinghub.elsevier.com/retrieve/pii/S1364032114000094.

[3] Zhang Y, Wang J, Wang X. Review on probabilistic forecasting of wind power generation. Renew Sustain Energy Rev 2014;32:255–70. http://dx.doi.org/10.1016/j.rser.2014.01.033, URL https://www.sciencedirect.com/science/article/pii/S1364032114000446.

[4] Bessa RJ, Möhrlen C, Fundel V, Siefert M, Browell J, Haglund El Gaidi S, Hodge B-M, Cali U, Kariniotakis G. Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry. Energies 2017;10(9):1402. http://dx.doi.org/10.3390/en10091402, URL https://www.mdpi.com/1996-1073/10/9/1402.

[5] Giebel G, Kariniotakis G. 3 - Wind power forecasting—a review of the state of the art. In: Kariniotakis G, editor. Renewable energy forecasting. Woodhead publishing series in energy, Woodhead Publishing; 2017, p. 59–109, URL https://www.sciencedirect.com/science/article/pii/B9780081005040000032.

[6] Santhosh M, Venkaiah C, Vinod Kumar DM. Current advances and approaches in wind speed and wind power forecasting for improved renewable energy integration: A review. Eng Rep 2020;2(6). http://dx.doi.org/10.1002/eng2.12178, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/eng2.12178.

[7] Bazionis IK, Georgilakis PS. Review of deterministic and probabilistic wind power forecasting: Models, methods, and future research. Electricity 2021;2(1):13–47. http://dx.doi.org/10.3390/electricity2010002, URL https://www.mdpi.com/2673-4826/2/1/2.

[8] Toubeau J-F, Bottieau J, Vallée F, De Grève Z. Improved day-ahead predictions of load and renewable generation by optimally exploiting multi-scale dependencies. In: 2017 IEEE innovative smart grid technologies - Asia. ISGT-Asia, 2017, p. 1–5. http://dx.doi.org/10.1109/ISGT-Asia.2017.8378396.

[9] Zhu A, Li X, Mo Z, Wu R. Wind power prediction based on a convolutional neural network. In: 2017 international conference on circuits, devices and systems. ICCDS, 2017, p. 131–5. http://dx.doi.org/10.1109/ICCDS.2017.8120465.

[10] Shabbir N, AhmadiAhangar R, Kütt L, Iqbal MN, Rosin A. Forecasting short term wind energy generation using machine learning. In: 2019 IEEE 60th international scientific conference on power and electrical engineering of riga technical university (RTUCON). 2019, p. 1–4. http://dx.doi.org/10.1109/RTUCON48111.2019.8982365.

[11] Xue H, Jia Y, Wen P, Farkoush SG. Using of improved models of Gaussian processes in order to regional wind power forecasting. J Clean Prod 2020;262:121391. http://dx.doi.org/10.1016/j.jclepro.2020.121391, URL https://www.sciencedirect.com/science/article/pii/S0959652620314384.

[12] Yu Y, Han X, Yang M, Yang J. Probabilistic prediction of regional wind power based on spatiotemporal quantile regression. IEEE Trans Ind Appl 2020;56(6):6117–27. http://dx.doi.org/10.1109/TIA.2020.2992945.

[13] Zhang H, Liu Y, Yan J, Han S, Li L, Long Q. Improved deep mixture density network for regional wind power probabilistic forecasting. IEEE Trans Power Syst 2020;35(4):2549–60. http://dx.doi.org/10.1109/TPWRS.2020.2971607.

[14] Dong W, Sun H, Tan J, Li Z, Zhang J, Zhao YY. Short-term regional wind power forecasting for small datasets with input data correction, hybrid neural network, and error analysis. Energy Rep 2021;7:7675–92. http://dx.doi.org/10.1016/j.egyr.2021.11.021, URL https://www.sciencedirect.com/science/article/pii/S2352484721011665.

[15] Wang K, Zhang Y, Lin F, Xu Y. Regional wind power forecasting based on hierarchical clustering and upscaling method. In: 2021 3rd Asia energy and electrical engineering symposium (AEEES). 2021, p. 713–8. http://dx.doi.org/10.1109/AEEES51875.2021.9403004.

[16] Dong W, Sun H, Tan J, Li Z, Zhang J, Yang H. Regional wind power probabilistic forecasting based on an improved kernel density estimation, regular vine copulas, and ensemble learning. Energy 2022;238:122045. http://dx.doi.org/10.1016/j.energy.2021.122045, URL https://www.sciencedirect.com/science/article/pii/S0360544221022933.

[17] Pei M, Ye L, Li Y, Luo Y, Song X, Yu Y, Zhao Y. Short-term regional wind power forecasting based on spatial–temporal correlation and dynamic clustering model. Energy Rep 2022;8:10786–802. http://dx.doi.org/10.1016/j.egyr.2022.08.204, URL https://www.sciencedirect.com/science/article/pii/S2352484722016481.

[18] Wang Z, Wang W, Liu C, Wang B, Feng S. Short-term probabilistic forecasting for regional wind power using distance-weighted kernel density estimation. IET Renew Power Gener 2018;12(15):1725–32. http://dx.doi.org/10.1049/iet-rpg.2018.5282, URL https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2018.5282.

[19] Pinson P, Kariniotakis G, Nielsen HA, Nielsen TS, Madsen H. Properties of quantile and interval forecasts of wind generation and their evaluation. In: Proceedings of the European wind energy conference & exhibition. 2006, p. 11.

[20] Meinshausen N. Quantile regression forests. J Mach Learn Res 2006;7(35):983–99.

[21] Schmelas M, Feldmann T, da Costa Fernandes J, Bollin E. Photovoltaics energy prediction under complex conditions for a predictive energy management system. J Solar Energy Eng 2015;137(3). http://dx.doi.org/10.1115/1.4029378.

[22] Ogliari E, Dolara A, Manzolini G, Leva S. Physical and hybrid methods comparison for the day ahead PV output power forecast. Renew Energy 2017;113:11–21. http://dx.doi.org/10.1016/j.renene.2017.05.063, URL https://www.sciencedirect.com/science/article/pii/S096014811730455X.

[23] Mayer MJ, Gróf G. Extensive comparison of physical models for photovoltaic power forecasting. Appl Energy 2021;283. http://dx.doi.org/10.1016/j.apenergy.2020.116239, URL https://www.sciencedirect.com/science/article/pii/S0306261920316330.

[24] Basu S, Watson SJ, Lacoa Arends E, Cheneka B. Day-ahead wind power predictions at regional scales: Post-processing operational weather forecasts with a hybrid neural network. In: 2020 17th international conference on the European energy market (EEM). 2020, p. 1–6. http://dx.doi.org/10.1109/EEM49802.2020.9221979.

[25] Bellinguer K, Mahler V, Camal S, Kariniotakis G. Probabilistic forecasting of regional wind power generation for the EEM20 competition: a physics-oriented machine learning approach. In: 2020 17th international conference on the European energy market. EEM, 2020, p. 1–6. http://dx.doi.org/10.1109/EEM49802.2020.9221960.

[26] Browell J, Drew DR, Philippopoulos K. Improved very short-term spatio-temporal wind forecasting using atmospheric regimes: Improved very short-term spatio-temporal wind forecasting using atmospheric regimes. Wind Energy 2018;21(11):968–79. http://dx.doi.org/10.1002/we.2207, URL https://onlinelibrary.wiley.com/doi/10.1002/we.2207.

[27] Bochenek B, Jurasz J, Jaczewski A, Stachura G, Sekuła P, Strzyżewski T, Wdowikowski M, Figurski M. Day-ahead wind power forecasting in Poland based on numerical weather prediction. Energies 2021;14(8):2164. http://dx.doi.org/10.3390/en14082164, URL https://www.mdpi.com/1996-1073/14/8/2164.

[28] Lepetit M, Kurzrock F, Aillaud P, Sebastien N, Schmutz N. Regional-scale day-ahead wind power forecasting using deep learning. Tech. rep. EGU22-6872, Copernicus Meetings; 2022, http://dx.doi.org/10.5194/egusphere-egu22-6872, URL https://meetingorganizer.copernicus.org/EGU22/EGU22-6872.html.

[29] Browell J, Gilbert C, Tawn R, May L. Quantile combination for the EEM20 wind power forecasting competition. In: 2020 17th international conference on the European energy market. EEM, 2020, p. 1–6. http://dx.doi.org/10.1109/EEM49802.2020.9221942.

[30] Jalali SMJ, Khodayar M, Khosravi A, Osório GJ, Nahavandi S, Catalão JPS. An advanced generative deep learning framework for probabilistic spatio-temporal wind power forecasting. In: 2021 IEEE international conference on environment and electrical engineering and 2021 IEEE industrial and commercial power systems Europe. EEEIC / I&CPS Europe, 2021, p. 1–6. http://dx.doi.org/10.1109/EEEIC/ICPSEurope51590.2021.9584664, URL https://ieeexplore.ieee.org/abstract/document/9584664.

[31] Arora P, Jalali SMJ, Ahmadian S, Panigrahi BK, Suganthan PN, Khosravi A. Probabilistic wind power forecasting using optimized deep auto-regressive recurrent neural networks. IEEE Trans Ind Inf 2023;19(3):2814–25. http://dx.doi.org/10.1109/TII.2022.3160696, URL https://ieeexplore.ieee.org/abstract/document/9739990.

[32] Wen H, Ma J, Gu J, Yuan L, Jin Z. Sparse variational Gaussian process based day-ahead probabilistic wind power forecasting. IEEE Trans Sustain Energy 2022;13(2):957–70. http://dx.doi.org/10.1109/TSTE.2022.3141549, URL https://ieeexplore.ieee.org/document/9676440.

[33] Alcántara A, Galván IM, Aler R. Deep neural networks for the quantile estimation of regional renewable energy production. Appl Intell 2023;53(7):8318–53. http://dx.doi.org/10.1007/s10489-022-03958-7, URL https://doi.org/10.1007/s10489-022-03958-7.

[34] Dong X, Sun Y, Dong L, Li J, Li Y, Di L. Transferable wind power probabilistic forecasting based on multi-domain adversarial networks. Energy 2023;285:129496. http://dx.doi.org/10.1016/j.energy.2023.129496, URL https://www.sciencedirect.com/science/article/pii/S0360544223028906.

[35] Lu P, Ye L, Pei M, Zhao Y, Dai B, Li Z. Short-term wind power forecasting based on meteorological feature extraction and optimization strategy. Renew Energy 2022;184:642–61. http://dx.doi.org/10.1016/j.renene.2021.11.072, URL https://www.sciencedirect.com/science/article/pii/S0960148121016554.

[36] Hu J, Luo Q, Tang J, Heng J, Deng Y. Conformalized temporal convolutional quantile regression networks for wind power interval forecasting. Energy 2022;248:123497. http://dx.doi.org/10.1016/j.energy.2022.123497, URL https://www.sciencedirect.com/science/article/pii/S0360544222004005.

[37] Wang W, Feng B, Huang G, Guo C, Liao W, Chen Z. Conformal asymmetric multi-quantile generative transformer for day-ahead wind power interval prediction. Appl Energy 2023;333:120634. http://dx.doi.org/10.1016/j.apenergy.2022.120634, URL https://www.sciencedirect.com/science/article/pii/S0306261922018918.

[38] Bengtsson L, Andrae U, Aspelien T, Batrak Y, Calvo J, Rooy Wd, Gleeson E, Hansen-Sass B, Homleid M, Hortal M, Ivarsson K-I, Lenderink G, Niemelä S, Nielsen KP, Onvlee J, Rontu L, Samuelsson P, Muñoz DS, Subias A, Tijm S, Toll V, Yang X, Koltzow MO. The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. Mon Weather Rev 2017;145(5):1919–35. http://dx.doi.org/10.1175/MWR-D-16-0417.1, URL https://journals.ametsoc.org/view/journals/mwre/145/5/mwr-d-16-0417.1.xml.

[39] Juban J, Fugon L, Kariniotakis G. Uncertainty estimation of wind power forecasts: Comparison of probabilistic modelling approaches. In: European wind energy conference & exhibition EWEC 2008. 2008, p. 11.

[40] Vovk V, Petej I, Nouretdinov I, Manokhin V, Gammerman A. Computationally efficient versions of conformal predictive distributions. Neurocomputing 2020;397:292–308. http://dx.doi.org/10.1016/j.neucom.2019.10.110, URL https://www.sciencedirect.com/science/article/pii/S0925231219316042.

[41] Romano Y, Patterson E, Candes E. Conformalized quantile regression. In: Advances in neural information processing systems, vol. 32, Curran Associates, Inc.; 2019, URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

[42] Villanueva D, Feijóo A. Wind power distributions: A review of their applications. Renew Sustain Energy Rev 2010;14(5):1490–5. http://dx.doi.org/10.1016/j.rser.2010.01.005, URL https://www.sciencedirect.com/science/article/pii/S1364032110000134.

[43] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

[44] Pinto F, Torr PHS, K. Dokania P. An impartial take to the CNN vs transformer robustness contest. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer vision – ECCV 2022. Lecture notes in computer science, Cham: Springer Nature Switzerland; 2022, p. 466–80. http://dx.doi.org/10.1007/978-3-031-19778-9_27.

[45] Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 11976–86, URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

[46] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. 2021, http://dx.doi.org/10.48550/arXiv.2010.11929, URL http://arxiv.org/abs/2010.11929.

[47] Beyer L, Izmailov P, Kolesnikov A, Caron M, Kornblith S, Zhai X, Minderer M, Tschannen M, Alabdulmohsin I, Pavetic F. FlexiViT: One model for all patch sizes. 2023, http://dx.doi.org/10.48550/arXiv.2212.08013, URL http://arxiv.org/abs/2212.08013.

[48] Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. 2022, http://dx.doi.org/10.48550/arXiv.1811.12231, URL http://arxiv.org/abs/1811.12231.

[49] Wu H, Xiao B, Codella N, Liu M, Dai X, Yuan L, Zhang L. CvT: Introducing convolutions to vision transformers. 2021, http://dx.doi.org/10.48550/arXiv.2103.15808, URL http://arxiv.org/abs/2103.15808.

[50] Tu Z, Talebi H, Zhang H, Yang F, Milanfar P, Bovik A, Li Y. MaxViT: Multi-axis vision transformer. 2022, http://dx.doi.org/10.48550/arXiv.2204.01697, URL http://arxiv.org/abs/2204.01697.

[51] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 4700–8, URL https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html.

[52] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition. CVPR, Las Vegas, NV, USA: IEEE; 2016, p. 770–8. http://dx.doi.org/10.1109/CVPR.2016.90, URL http://ieeexplore.ieee.org/document/7780459/.

[53] Linusson H, Johansson U, Löfström T. Signed-error conformal regression. In: Tseng VS, Ho TB, Zhou Z-H, Chen ALP, Kao H-Y, editors. Advances in knowledge discovery and data mining. Lecture notes in computer science, Cham: Springer International Publishing; 2014, p. 224–36. http://dx.doi.org/10.1007/978-3-319-06608-0_19.

[54] Vovk V, Lindsay D, Nouretdinov I, Gammerman A. Mondrian confidence machine. 2003, URL https://pure.royalholloway.ac.uk/en/publications/mondrian-confidence-machine.

[55] Boström H, Johansson U. Mondrian conformal regressors. In: Proceedings of the ninth symposium on conformal and probabilistic prediction and applications. PMLR; 2020, p. 114–33, URL https://proceedings.mlr.press/v128/bostrom20a.html.

[56] Boström H, Johansson U, Löfström T. Mondrian conformal predictive distributions. In: Proceedings of the tenth symposium on conformal and probabilistic prediction and applications. PMLR; 2021, p. 24–38, URL https://proceedings.mlr.press/v152/bostrom21a.html.

[57] Wang D, Wang P, Wang C, Wang P. Calibrating probabilistic predictions of quantile regression forests with conformal predictive systems. Pattern Recognit Lett 2022;156:81–7. http://dx.doi.org/10.1016/j.patrec.2022.02.003, URL https://www.sciencedirect.com/science/article/pii/S016786552200037X.

[58] Johansson U, Löfström T, Boström H. Conformal predictive distribution trees. Ann Math Artif Intell 2023. http://dx.doi.org/10.1007/s10472-023-09847-0.

[59] Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. 2017, http://dx.doi.org/10.48550/arXiv.1604.04173, URL http://arxiv.org/abs/1604.04173.

[60] Fontana M, Zeni G, Vantini S. Conformal prediction: a unified review of theory and new challenges. 2022, http://dx.doi.org/10.48550/arXiv.2005.07972, URL http://arxiv.org/abs/2005.07972.

[61] Biau G, Devroye L, Lugosi G. Consistency of random forests and other averaging classifiers. J Mach Learn Res 2008;9:2015–33.
[62] Biau G. Analysis of a random forests model. J Mach Learn Res 2012;13(1):1063–95.
[63] Loshchilov I, Hutter F. Decoupled weight decay regularization. Tech. rep., 2019, arXiv. arXiv:1711.05101. URL http://arxiv.org/abs/1711.05101.
[64] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016, p. 785–94. http://dx.doi.org/10.1145/2939672.2939785, URL http://arxiv.org/abs/1603.02754.
[65] Steinwart I, Christmann A. Estimating conditional quantiles with the help of the pinball loss. Bernoulli 2011;17(1):211–25. http://dx.doi.org/10.3150/10-BEJ267, URL https://projecteuclid.org/journals/bernoulli/volume-17/issue-1/Estimating-conditional-quantiles-with-the-help-of-the-pinball-loss/10.3150/10-BEJ267.full.
[66] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Amer Statist Assoc 2007;102(477):359–78. http://dx.doi.org/10.1198/016214506000001437, URL https://doi.org/10.1198/016214506000001437.
[67] Werner H, Carlsson L, Ahlberg E, Boström H. Evaluating different approaches to calibrating conformal predictive systems. In: Proceedings of the ninth symposium on conformal and probabilistic prediction and applications. PMLR; 2020, p. 134–50, URL https://proceedings.mlr.press/v128/werner20a.html.