Asset Bundling for Hierarchical Forecasting of Wind Power Generation

Hanyu Zhang, Mathieu Tanneau, Chaofan Huang, V. Roshan Joseph, Shangkun Wang, Pascal Van Hentenryck
Georgia Institute of Technology, Atlanta, USA
{hzhang747, chaofan.huang, roshan, sk_wang}@gatech.edu, {mathieu.tanneau, pascal.vanhentenryck}@isye.gatech.edu

***Abstract*—The growing penetration of intermittent, renewable generation in US power grids, especially wind and solar generation, results in increased operational uncertainty. In that context, accurate forecasts are critical, especially for wind generation, which exhibits large variability and is historically harder to predict. To overcome this challenge, this work proposes a novel Bundle-Predict-Reconcile (BPR) framework that integrates asset bundling, machine learning, and forecast reconciliation techniques. The BPR framework first learns an intermediate hierarchy level (the bundles), then predicts wind power at the asset, bundle, and fleet level, and finally reconciles all forecasts to ensure consistency. This approach effectively introduces an auxiliary learning task (predicting the bundle-level time series) to help the main learning tasks. The paper also introduces new asset-bundling criteria that capture the spatio-temporal dynamics of wind power time series. Extensive numerical experiments are conducted on an industry-size dataset of 283 wind farms in the MISO footprint. The experiments consider short-term and day-ahead forecasts, and evaluates a large variety of forecasting models that include weather predictions as covariates. The results demonstrate the benefits of BPR, which consistently and significantly improves forecast accuracy over baselines, especially at the fleet level.**

***Index Terms*—Asset bundling, wind power forecasting, machine learning, forecast reconciliation**

I. Introduction

The sustained growth of wind power generation in US power grids is causing operational and reliability challenges for Transmission System Operators (TSOs). In particular, the intrinsic variability and intermittency of wind resources creates significant operational uncertainty, which must be managed to ensure safe and reliable operations. This, in turn, requires accurate wind power forecasts at various temporal and geographical scales. Indeed, market participants use wind power forecasts at the asset level to inform their participation in day-ahead and real-time markets, while TSOs primarily use aggregated, fleet-wide forecasts to operate the grid. Forecasts must be accurate at both levels while remaining consistent, i.e., the sum of individual forecasts should be equal to the forecast for the overall fleet. As a result, this paper considers the task of producing high-dimensional, short-term and day-ahead, wind power forecasts that satisfy three requirements:

1) the forecasts predict the output of individual asset and the total output of the fleet;
2) the forecasts are *consistent*, i.e., the total forecast is the sum of the individual forecasts;
3) the forecasts are accurate both at the asset level and at the fleet level.

To carry this task, the paper proposes a novel hierarchical approach, *the Bundle-Predict-Reconcile* (BPR) framework, which (1) learns an additional forecast level (*the bundles*) between the assets and the fleet, (2) predicts outputs at the asset, bundle, and fleet levels, and (3) *reconciles* all levels for ensuring consistency. The rest of this section surveys the relevant literature and presents the paper's contributions.

A. Related Work

Because of its importance for power systems operations, wind power forecasting has received substantial attention. Readers are referred to [1], [2] for comprehensive reviews. For ease of presentation, the paper focuses on deterministic forecasts but BPR naturally extends to probabilistic settings. See [3] for an overview of probabilistic wind forecasting.

Wind power forecasting methods broadly fall under two categories: physical and statistical models. On the one hand, physical methods [4], [5] combine weather-based wind speed predictions with wind power curves [6] to produce wind power forecasts. On the other hand, statistical methods use historical data to train data-driven models. These include traditional time series forecasting models, such as Auto-Regressive Moving Average (ARMA) and its variants [1]. More recently, Machine Learning (ML) models have been applied to wind power forecasting, e.g., Artificial Neural Networks [1], recurrent neural networks [7], [8], and transformer-based architectures [9]–[11]. A comparison of several deep learning models for wind power forecasting is presented in [12]. Statistical models are typically better at short-term ($\leq$6 hours) predictions, but are outperformed by physical models for longer horizons [1]. Current state-of-the-art approaches thus combine physics-based weather predictions with data-driven models [2], [13].

Wind power forecasts are often produced in a hierarchical fashion, with different models predicting different levels of the hierarchy. For instance, one model predicts the output of individual wind farms, and another model predicts the fleet's total output. Forecast reconciliation [14], [15] can then be used to ensure consistency of hierarchical wind forecasts [16]–[20].

This research was partly supported by NSF award 2112533 and ARPA-E PERFORM award AR0001136.

Zhang et al. [16] compare the performance of several reconciliation approaches for very short-term ($\leq$1 hour) predictions. Bai et al [17] propose a distributed algorithm to solve the reconciliation problem, and an online reconciliation approach is presented in [18]. An end-to-end learning approach for hierarchical wind power forecast is introduced in [19] with a special focus on handling missing values. Hansen et al. [20] find that hierarchical reconciliation leads to improved day-ahead wind forecasts, especially for fleet-level predictions. In all these works [16]–[20], the hierarchical structure is fixed and given a priori. In contrast, the BPR framework proposed in this paper improves the forecast accuracy by *learning a hierarchical structure through asset bundling.*

Asset bundling is a popular technique in portfolio optimization, tracing back to the mean variance model of Markowitz [21]. In the context of power systems, a similar approach is known as the geographical smoothing effect, wherein the aggregated output of multiple wind farms exhibits lower variability [22]–[25]. The geographical smoothing effect has been employed in prior literature on optimal design, planning, and re-powering of wind farms [26]–[31]. References [32]–[36] use a similar approach, but consider a conditional value at risk (CVaR) objective. *Unlike the above portfolio selection-based approaches, the paper proposes to use asset bundling to improve the quality of wind power forecasts.*

B. Contributions

The BPR framework is a three step-approach that first learns a prediction hierarchy by bunding assets, before forecasting time series at the asset, bundle, and fleet levels, before reconciling the forecasts at all levels. A key innovation of the BPR framework is its ability *to learn the "best" bundles*. In other words, BPR is not given a forecasting hierarchy; rather it learns a hierarchy by choosing the bundles that will help improve the asset and fleet forecasts. As a result, BPR can be viewed as a multi-task learning approach: it introduces an auxiliary learning task, i.e., forecasting the bundles, to help the main learning tasks. A second innovation of BPR is the criteria used to select the bundles, leveraging the spatio-temporal correlations that exist in time series of wind power.

The paper presents a comprehensive evaluation of BPR that considers both short-term and day-ahead predictions, includes weather forecasts as covariates, and evaluates a large variety of forecasting models for the individual prediction tasks. Experimental results for industry-size test cases from NREL [37] clearly show the benefits of BPR. They also give significant insights on the impact of bundling and ML architecture on the quality of the forecasts. The learning architectures include recurrent neural networks, transformer models, temporal convolution transformers, and temporal fusion transformers. The bundling criteria include geographical distance, seasonal-adjusted variance, and the intermittency index.

In summary, the paper makes three main contributions:

1) It introduces a new approach, Bundle-Predict-Reconcile (BPR), that learn a forecasting hierarchy to improve the prediction accuracy at the asset and fleet level.
2) It proposes new bundling criteria that exploit the spatio-temporal structure of the time series for wind power.
3) It evaluates BPR on several forecasting tasks (short-term and day-ahead), numerous learning architectures, and covariate configurations, demonstrating its state-of-the-art performance for both short-term and day-ahead wind power forecasts. In particular, BPR improves fleet-level accuracy by about 25% over a strong baseline in short-term forecasting and about 10% over the best models in day-ahead forecasting.

The rest of the paper is organized as follows. Section II presents the problem formulation and summarizes the proposed approach. Section III presents the proposed asset-bundling methodology. Section IV describes the hierarchical forecasting and reconciliation framework, and the ML architectures used in the paper. Section V presents numerical experiments, and Section VI concludes the paper.

II. Problem Formulation

Let $\mathcal{N} = \{1, ..., N\}$ denote the set of wind farms in the system. The output of wind farm $i \in \{1, ..., N\}$ at time $t \in \mathbb{Z}$ is denoted by $\mathbf{x}_{i,t}$. The following notations are used in the paper for $i \in \mathcal{N}$, $t \in \mathbb{Z}$, and $\mathcal{T} \subseteq \mathbb{Z}$:

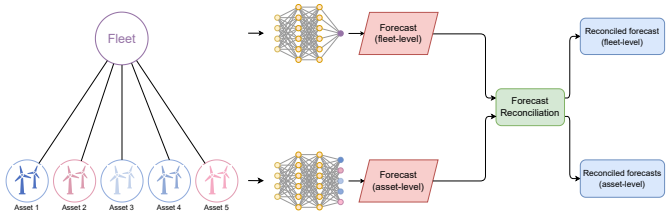
$$\mathbf{x}_t = (\mathbf{x}_{i,t})_{i \in \mathcal{N}} \in \mathbb{R}^N$$
$$\mathbf{x}_{\mathcal{T}} = (\mathbf{x}_{\tau})_{\tau \in \mathcal{T}} \in \mathbb{R}^{N \times |\mathcal{T}|}$$
$$\mathbf{x}_{i,\mathcal{T}} = (\mathbf{x}_{i,\tau})_{\tau \in \mathcal{T}} \in \mathbb{R}^{|\mathcal{T}|}$$

The paper takes the perspective of a system operator that has full visibility over the output of every wind farm in the system, thus eliminating data privacy issues. Note that BPR may be extended to settings where asset-level data must be kept private via, e.g., separate asset-level models, federated learning to train a joint asset-level forecasting model [38], [39], and distributed forecast reconciliation [17].

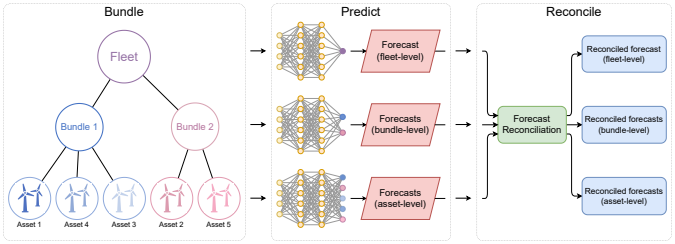
A. Wind Power Forecasting

Let $t \in \mathbb{Z}$ denote the current time period, and denote by $\mathcal{H} = \{t-H+1, ..., t\}$ and $\mathcal{T} = \{t+1, ..., t+T\}$ the *historical* and *forecast* windows, respectively. Given past observations $\mathbf{x}_{\mathcal{H}}$, the goal of wind power forecasting is to produce a forecast $\hat{\mathbf{x}}_{\mathcal{T}}$ that is as close as possible to the actual (future) realization $\mathbf{x}_{\mathcal{T}}$. While the paper considers point forecasts for ease of presentation, the BPR methodology naturally extend to probabilistic forecasting.

TSOs and market participants use wind power forecasts as input for market-clearing and to conduct reliability studies. In that context, it is important to note that spatial correlations (between wind farms) and temporal correlations (between time periods) have an impact on congestion management and flexibility (ramping) requirements. Therefore, forecasts should capture spatio-temporal dynamics, which is achieved by jointly predicting all $N \times T$ components of $\hat{\mathbf{x}}_{\mathcal{T}}$. The paper considers joint forecasts for hundreds wind farms and time horizons containing 24 to 48 time steps, giving prediction tasks with over 10,000 output dimensions. Such large output dimensions



(a) Baseline hierarchical wind power forecasting. Left: the initial two-level (asset, fleet) hierarchy; each asset is an individual wind farm. Middle: two models are trained for the asset and fleet levels. Right: consistent forecasts are obtained via reconciliation.



(b) The proposed Bundle-Predict-Reconcile (BPR) framework. Left: assets are grouped into bundles. Middle: a model is trained for each level (asset, bundle, fleet) of the hierarchy. Right: consistent forecasts are obtained via reconciliation.

Fig. 1: Illustration of the BPR Framework.

create scalability challenges and have an adverse impact on forecast accuracy.

B. Overview of BPR

Figure 1 sketches the BPR framework. A baseline approach to wind power forecasting is illustrated in Figure 1a: it trains a model to predict asset-level and fleet-level generations; a two-level reconciliation step ensures consistency. BPR is depicted in Figure 1b. First, assets, (i.e., wind farms), are grouped into K bundles based on spatial and temporal features (see Section III). Second, machine-learning models are trained to deliver forecasts at the asset, bundle, and fleet levels. Third, a hierarchical reconciliation step produces consistent forecasts.

While the presentation of BPR in this paper considers a 3-level hierarchy (asset, bundle, fleet), the proposed methodology directly extends to any hierarchy. The main difference between BPR and existing hierarchical forecasting approaches is the initial bundling step that *learns* a hierarchy based on spatio-temporal features, so as to improve forecast accuracy.

III. Asset Bundling

This section describes the asset-bundling methodology of BPR, from its problem formulation to its solution algorithm. The presentation assumes a set of N wind farms with given location, nameplate capacity, and historical output $\mathbf{x}_{\mathcal{T}}$ where $\mathcal{T}=\{1,...,T\}$. The distance between wind farms i and j is denoted by D_{ij}.

A. Problem Formulation

The asset bundling problem consists in grouping N wind farms into $K \in \{1,..,N\}$ groups, referred to as *bundles*, so as to optimize a certain criterion. Setting $K=1$ aggregates all wind farms into a single bundle, whereas $K=N$ considers all wind farms individually. The asset-bundling problem can be modeled as a combinatorial optimization problem.

Model 1 The Asset-Bundling Problem

Input: Time series $\mathbf{x}_{\mathcal{T}}$, distance matrix D, distance cutoff $\bar{D}$
Variables: $\lambda_{k,i}$ is 1 if wind farm i is in bundle k, 0 otherwise

$$\min_{\lambda} \quad f(\mathbf{z}_{\mathcal{T}}, \Lambda) \tag{1a}$$
$$\text{s.t.} \quad \mathbf{z}_{\mathcal{T}} = \Lambda \mathbf{x}_{\mathcal{T}} \tag{1b}$$
$$\sum_{i \in \mathcal{N}} \lambda_{k,i} \geq 1 \qquad \forall k \in \mathcal{K} \tag{1c}$$
$$\sum_{k=1}^{K} \lambda_{k,i} = 1 \qquad \forall i \in \mathcal{N} \tag{1d}$$
$$D_{ij}(\lambda_{k,i} + \lambda_{k,j}) \leq 2\bar{D} \qquad \forall (i,j) \in \mathcal{N}, k \in \mathcal{K} \tag{1e}$$
$$\Lambda \in \{0,1\}^{K \times N} \tag{1f}$$

Model 1 presents an optimization model for the asset-bundling problem and a criterion f. Binary variable $\lambda_{k,i}$ takes value 1 iff wind farm i belongs to bundle k. Constraints (1b) capture the bundled time series $\mathbf{z}_{\mathcal{T}} \in \mathbb{R}^{K \times T}$. Namely, the output of bundle k at time t is $\mathbf{z}_{k,t} = \sum_i \lambda_{k,i} \mathbf{x}_{i,t}$, i.e., it is the combined output of all wind farms assigned to bundle k. Constraints (1c) and (1d) ensure that each bundle is non-empty and that each wind farm is assigned to exactly one bundle. Constraints (1e) enforce that two wind farms cannot be assigned to the same bundle if the distance between them exceeds a maximum diameter $\bar{D}$. Finally, the objective (1a) captures properties of the bundled time series $\mathbf{z}_{\mathcal{T}}$ and of the bundling assignment Λ; possible choices for the objective function f are described next.

B. Bundling Criteria

This section describes various bundling criteria, which capture desirable properties of the bundled time series.

1) Variance: A natural bundling criterion is to minimize the total variance of the bundles, i.e.,

$$f(\mathbf{z}_{\mathcal{T}}, \Lambda) = \sum_{k=1}^{K} \mathrm{Var}(\mathbf{z}_{k,\mathcal{T}}) = \mathrm{tr}(\Lambda \mathbf{\Sigma} \Lambda^{\top}), \tag{2}$$

where $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$ is the empirical covariance matrix of $\mathbf{x}_{\mathcal{T}}$, tr stands for the trace of a matrix, and $\top$ represents the matrix (or vector) transpose. Variance-based bundling is illustrated in Figure 2. Note that bundling reduces the total variance only if negatively correlated assets are bundled together. Therefore, a key factor for its success is the proportion of negative correlations among the original assets.

2) Seasonal-Adjusted Variance: A large proportion of wind farms may exhibit positive correlations over a given study period due to, e.g., common daily patterns. This is detrimental to variance-based bundling, which relies on negative correlations. To mitigate this effect, the seasonal-adjusted variance is defined as

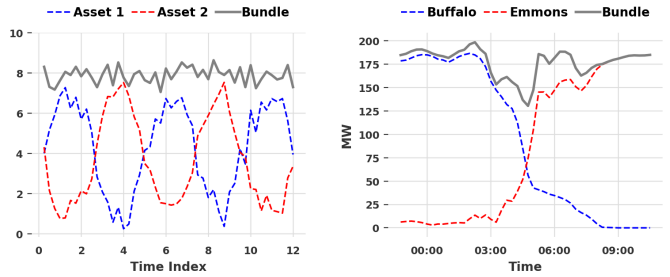


Fig. 2: Illustration of variance-based bundling. Left: synthetic example. Right: real example with two MISO wind farms from [37]. In each case, the two assets are negatively correlated, and bundling yields a reduction in variance.

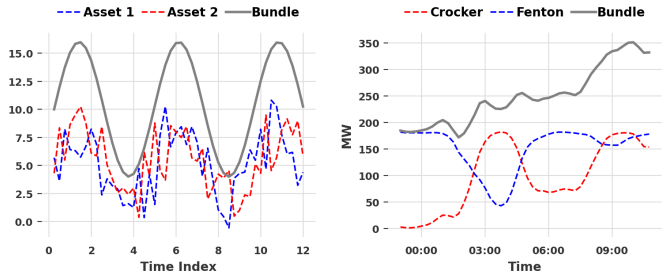


Fig. 3: Illustration of intermittency index-based bundling. Left: synthetic example. Right: real example with two MISO wind farms from [37]. In each case, bundling yields a reduction in intermittency: the bundled time series exhibits fewer high-frequency oscillations and clearer long-term patterns.

$$\mathrm{SAVar}(\mathbf{x}_{\mathcal{T}}) = \sum_{i\in\mathcal{N}} \mathrm{Var}(\mathbf{x}_{i,\mathcal{T}} - \boldsymbol{\mu}_{\mathcal{T}}), \tag{3}$$

where the average $\boldsymbol{\mu}_{\mathcal{T}} = \frac{1}{N}\sum_{i\in\mathcal{N}} \mathbf{x}_{i,\mathcal{T}}$ captures the seasonal trend. This yields the objective function

$$f(\mathbf{z}_{\mathcal{T}}, \Lambda) = \mathrm{SAVar}(\mathbf{z}_{\mathcal{T}}) = \mathrm{tr}(\Lambda\tilde{\boldsymbol{\Sigma}}\Lambda^\top), \tag{4}$$

where $\tilde{\boldsymbol{\Sigma}}$ is the covariance matrix of the seasonal-adjusted time series $\tilde{\mathbf{x}}_{\mathcal{T}} = \mathbf{x}_{\mathcal{T}} - \boldsymbol{\mu}_{\mathcal{T}}$. Seasonal-adjusted variance-based bundling promotes bundles that behave similarly to the mean of the assets, which is expected to improve the quality of the forecast when the mean exhibits simple patterns.

3) Intermittency Index: The *intermittency index* of $\mathbf{x}_{\mathcal{T}}$ is

$$\mathrm{Imcy}(\mathbf{x}_{\mathcal{T}}) = \sum_{i\in\mathcal{N}} \mathrm{Var}(\dot{\mathbf{x}}_{i,\mathcal{T}}), \tag{5}$$

where $\dot{\mathbf{x}}_{i,t} = \mathbf{x}_{i,t} - \mathbf{x}_{i,t-1}$; $\dot{\mathbf{z}}$ is defined similarly. Substituting $\dot{\mathbf{z}} = \Lambda\dot{\mathbf{x}}$, the objective function (1a) becomes

$$f(\mathbf{z}_{\mathcal{T}}, \Lambda) = \mathrm{Imcy}(\mathbf{z}_{\mathcal{T}}) = \mathrm{tr}(\Lambda\dot{\boldsymbol{\Sigma}}\Lambda^\top), \tag{6}$$

where $\dot{\boldsymbol{\Sigma}}$ is the covariance matrix of $\dot{\mathbf{x}}_{\mathcal{T}}$. Intermittency-based bundling is illustrated in Figure 3 where the bundled time series exhibit long-term patterns that are easier to learn. In the context of power systems operations, the ability of intermittency-based bundling to identify long-term trends allows to better anticipate and manage ramping needs.

C. The Bundling Algorithm

The objective functions in Eqs. (2), (4) and (6) share the same intrinsic structure, differring only in the choice of the covariance matrix $\boldsymbol{\Sigma}$, $\tilde{\boldsymbol{\Sigma}}$, or $\dot{\boldsymbol{\Sigma}}$. As a result, for each proposed bundling criterion, Problem (1) is a convex mixed-integer quadratic programming (MIQP) problem, which can be solved using MIQP solvers like Gurobi or CPLEX. However, for realistic test cases, the asset-bundling problem is intractable for MIQP solvers, mainly due to its intrinsic complexity and the symmetry of set-partitioning constraints (1d). The latter are notoriously detrimental to the performance of MIQP solvers.

To remedy these scalability issues, the BPR framework uses a fast greedy algorithm presented in Algorithm 1. For simplicity, Algorithm 1 is presented for the case of Variance-based bundling. The algorithm takes as input the time series $\mathbf{x}$ and the number of bundles K, and initializes a set of N bundles, each corresponding to an individual wind farm. Then, at each step, the algorithm aggregates the two most negatively correlated bundles, subject to the maximum diameter constraint. Algorithm 1 is computationally fast, and was found to produce close-to-optimal solutions in early experiments.

Algorithm 1: The Greedy Asset-Bundling Algorithm.

Input: Time series $\mathbf{x}_{\mathcal{T}}$, distance matrix D, maximum diameter $\bar{D}$, number of bundles K
Initialization: $\mathcal{K} \leftarrow \{\{1\}, ..., \{N\}\}$
while $|\mathcal{K}| > K$ **do**
  $\mathbf{z}_k \leftarrow \sum_{i\in\mathcal{K}_k} \mathbf{x}_{i,\mathcal{T}}, \quad k = 1, ..., |\mathcal{K}|$
  $\tilde{D}_{k,l} \leftarrow \max\{D_{ij} | (i,j)\in\mathcal{K}_k\times\mathcal{K}_l\}, \quad k,l = 1, ..., |\mathcal{K}|$
  $(k,l) \leftarrow \arg\min \left\{ \mathrm{Cov}(\mathbf{z}_k, \mathbf{z}_l) \,\middle|\, \tilde{D}_{kl} \le \bar{D} \right\}$
  $\mathcal{K} \leftarrow (\mathcal{K} \setminus \{\mathcal{K}_k, \mathcal{K}_l\}) \cup (\mathcal{K}_k \cup \mathcal{K}_l)$
end
Return: set of bundles $\mathcal{K}$

IV. Forecasting

This section presents the hierarchical forecasting and reconciliation component of BPR, the architecture of the ML models considered in the experiments, and the metrics for evaluating the quality of the forecasts. The section assumes a fixed bundling of N wind farms into K bundles, denoted by $\Lambda \in \{0,1\}^{K\times N}$. For $t\in\mathbb{Z}$, $\mathbf{z}_t = \Lambda\mathbf{x}_t \in \mathbb{R}^K$ and $\mathbf{X}_t = e^\top \mathbf{x}_t \in \mathbb{R}$ denote the bundle-level and total-level wind power output, respectively. The forecast horizon at time t is $\mathcal{T}_t = \{t+1, ..., t+T\}$.

A. Reconciliation

BPR works on a three-level hierarchical structure illustrated in Figure 1b: asset-level ($\mathbf{x}$), bundle-level ($\mathbf{z}$), and total-level ($\mathbf{X}$). The paper leverages this structure by training three models that predict each level of this hierarchy: at time t, three forecasts $\hat{\mathbf{x}}_{\mathcal{T}_t}, \hat{\mathbf{z}}_{\mathcal{T}_t}, \hat{\mathbf{X}}_{\mathcal{T}_t}$ are produced that predict the

future asset-level, bundle-level, and total-level time series $\mathbf{x}_{\mathcal{T}_t}, \mathbf{z}_{\mathcal{T}_t}, \mathbf{X}_{\mathcal{T}_t}$, respectively.

The reconciliation aims at making these predictions consistent. Given forecasts $\hat{\mathbf{x}}, \hat{\mathbf{z}}, \hat{\mathbf{X}}$ at the asset, bundle, and fleet levels, reconciliation produces new forecasts $\tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \tilde{\mathbf{X}}$ such that

$$\forall k, \tilde{\mathbf{z}}_{k,\mathcal{T}_t} = \sum_{i \in \mathcal{N}} \lambda_{k,i} \tilde{\mathbf{x}}_{i,\mathcal{T}_t}$$
$$\tilde{\mathbf{X}}_{\mathcal{T}} = \sum_{i \in \mathcal{N}} \tilde{\mathbf{x}}_{i,\mathcal{T}} = \sum_{k} \tilde{\mathbf{z}}_{k,\mathcal{T}}$$

More formally, forecast reconciliation [14] takes as input incoherent forecasts, and produces the most accurate consistent forecasts. Let $\hat{\mathbf{h}} = (\hat{\mathbf{X}}, \hat{\mathbf{z}}, \hat{\mathbf{x}})$ and S be the summing matrix

$$S = \begin{bmatrix} e & \Lambda^\top & I \end{bmatrix}^\top \in \mathbb{R}^{(N+K+1)\times N}. \tag{7}$$

Reconciliation searches for a matrix $G \in \mathbb{R}^{N \times (N+K+1)}$ such that $\tilde{\mathbf{h}}_\tau = SG\hat{\mathbf{h}}_\tau$. In addition, to improve accuracy further, BPR computes T matrices $G_1, ..., G_T$ such that

$$\forall \tau \in \{1, ..., T\}, \ \tilde{\mathbf{h}}_\tau = SG_\tau \hat{\mathbf{h}}_\tau$$

to capture the variances at different lead times τ. BPR uses the MinT optimal reconciliation approach from [15], with a weighted least squares (WLS) estimator using variance scaling. Thereby, each G_τ has the form

$$G_\tau = (S^\top W_\tau^{-1} S)^{-1} S^\top W_\tau^{-1} \tag{8}$$

where $W_\tau = \text{Diag}(\hat{W}_\tau)$ and

$$\hat{W}_\tau = \frac{1}{M} \sum_{t \in \mathcal{M}} \epsilon_{t,\tau} \epsilon_{t,\tau}^\top \tag{9}$$

and $\epsilon_{t,\tau} = (\hat{\mathbf{h}}_{t,\tau} - \mathbf{h}_{t,\tau})$. For the case at hand, WLS using variance scaling used in [40] and [15] achieved better performance than other MinT-based reconcilation approaches such as Ordinary least squres in [14], MinT(sample) [15], and WLS with structure scaling in [40]. Therefore, only WLS using variance scaling is reported in Section V.

B. Forecasting Models

BPR was instantiated with various learning architectures for the experimental evaluation. These are summarized in Table I, which reports, for each learning model, whether it supports past, future, and static covariates (see [41]). The co-variates may include static information about each wind farm (e.g., its location and model), temporal information such as hour of the day or month of the year, and past and forecasted weather-related information such as temperature, humidity, wind speed and direction. Covariates have been shown to improve forecast accuracy [42], [43].

1) Persistence Model: The Persistence model (`Per`) is a naive baseline whose prediction for the entire forecasting horizon is equal to the last observed value.

2) Linear Model: The linear regression model is trained using ridge regression. Without loss of generality, a separate model is trained for each bundle, which makes training more efficient.

TABLE I: Overview of ML Architectures.

Model	Covariates*	Description
`Per`	✗ ✗ ✗	Persistence model
`Lin`	✓ ✓ ✓	Linear model
`RNN`	✗ ✓ ✗	Recurrent Neural Network
`Tfm`	✓ ✗ ✗	Transformer
`TCN`	✓ ✗ ✗	Temporal Convolution Transformer
`TFT`	✓ ✓ ✓	Temporal Fusion Transformer

*Whether a model supports past/future/static covariates [41]

3) Recurrent Neural Network: The Recurrent Neural Network (RNN) architecture follows the state-of-the-art DeepAR model [8]. It is applied in an auto-regressive way to compute multi-step predictions. The RNN models are trained with mean squared error (MSE) loss.

4) Transformer models: The baseline transformer model follows the approach of [44]. Namely, it employs a multi-head attention mechanism inside an encoder-decoder, sequence-to-sequence architecture.

5) Temporal Convolution Transformer: The Temporal Convolution Transformer (TCN) architecture follows the model from [10]. Unlike the baseline transformer model, TCN uses stacked dilated causal convolutional neural networks to extract features from historical observations.

6) Temporal Fusion Transformer: The Temporal Fusion Transformer (TFT) architecture [11] is a deep learning architecture designed for multi-horizon forecasting tasks with a mixed-type inputs, e.g., past observations and past, future, and static covariates. This allows the integration of weather-based, wind speed forecasts into the model, which improves longer-term predictions. The TFT architecture uses a combination of recurrent layers and self-attention layers to learn temporal relationships at different scales and provide interpretable insights into the temporal dynamics. The model utilizes specialized components to select relevant features, and gating layers to suppress unnecessary components.

C. Evaluation Metrics

The experimental evaluation uses various metrics to evaluate the quality of wind power forecasts. Each metric is presented for asset-level times series $\mathbf{x}$ in what follows, but they are similar for the bundle and fleet levels. See [45] for a detailed discussion of wind power forecast evaluation.

1) Normalized Mean Absolute Error: The Normalized Mean Absolute Error (NMAE) is defined as

$$\text{NMAE} = \frac{1}{MNT} \sum_{t \in \mathcal{M}} \sum_{i \in \mathcal{N}} \frac{\|\mathbf{x}_{i,\mathcal{T}_t} - \hat{\mathbf{x}}_{i,\mathcal{T}_t}\|_1}{\bar{\mathbf{x}}_i} \tag{10}$$

where $\bar{\mathbf{x}}_i$ is the maximum capacity of wind farm i and $\|\cdot\|_1$ is the vector ℓ_1 norm.

2) Root-mean-square Error: The Root-mean-square Error (RMSE) is expressed as

$$\text{RMSE} = \sqrt{\frac{1}{MTN} \sum_{i \in \mathcal{N}} \sum_{t \in \mathcal{M}} \|\mathbf{x}_{\mathcal{T}_t} - \hat{\mathbf{x}}_{\mathcal{T}_t}\|_2^2} \tag{11}$$

where $\|\cdot\|_2$ is the vector ℓ_2 norm. RMSE better captures large prediction errors than NMAE.

3) Variogram Score: The Variogram Score (VS) of order $p > 0$ is given by

$$\mathrm{VS}_p = \frac{1}{M} \sum_{t \in \mathcal{M}} \sum_{i,j \in \mathcal{N}} \sum_{\tau,\tau' \in \mathcal{T}_t} \left(\mathbf{\Delta x}^p_{i,j,\tau,\tau'} - \mathbf{\Delta \hat{x}}^p_{i,j,\tau,\tau'} \right)^2 \tag{12}$$

where $\mathbf{\Delta x}^p_{i,j,\tau,\tau'}$ and $\mathbf{\Delta \hat{x}}^p_{i,j,\tau,\tau'}$ are defined as

$$\mathbf{\Delta x}^p_{i,j,\tau,\tau'} = |\mathbf{x}_{i,\tau} - \mathbf{x}_{j,\tau'}|^p, \quad \mathbf{\Delta \hat{x}}^p_{i,j,\tau,\tau'} = |\hat{\mathbf{x}}_{i,\tau} - \hat{\mathbf{x}}_{j,\tau'}|^p.$$

The paper uses $p = 1/2$, as recommended in [46]. Note that computing VS_p requires $\mathcal{O}(MN^2T^2)$ operations, which is computationally expensive for the case at hand. Therefore, the paper only reports VS for total wind predictions, which reduces the complexity to $\mathcal{O}(MT^2)$.

4) Energy Distance: The Energy Distance (ED) is

$$\mathrm{ED} = \frac{2}{M} \sum_{t \in \mathcal{M}} \|\mathbf{x}_{\mathcal{T}_t} - \hat{\mathbf{x}}_{\mathcal{T}_t}\|_F \tag{13}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

V. Experimental Evaluation

A. Data

The results are presented on the NREL dataset [37] that contains 283 time series of existing wind farms in MISO system for 2018 and 2019 at different time granularities (1 hour and 15 minutes). This dataset is complemented by a separate dataset of weather forecasts from the European Center for Medium-Range Weather Forecasts (ECMWF) [47]. These forecasts are collected on a 1-degree by 1-degree grid for the MISO region and used as covariates in the learning models. They are generated twice a day, at noon and midnight in UTC time, and have a 6-hour granularity. The weather forecasts time series are interpolated from 6-hour granularity down to the desired granularity of the forecasts.

B. Experimental Setting

The experiments consider both short-term and day-ahead forecasting tasks. Short-term forecasts predict wind power generation over the next 6 hours with a granularity of 15 minutes; it is the forecasting task needed for the Look-Ahead Commitment (LAC) in the MISO market clearing pipeline. Short-term forecasts are purely data-driven, because the original ECMWF weather forecasts have a granularity of 6 hours and are produced only twice a day. Day-ahead forecasts predict wind power generation over the next 48 hours with a granularity of 1 hour; it is the forecasting task needed for the day-ahead reliability commitment (FRAC) in the MISO market clearing pipeline. The day-ahead forecasting task is data-driven as well but it also uses co-variates given by the latest available ECMWF weather forecasts.

The evaluation uses a test set comprising five weeks in 2019, i.e., the weeks of January 8th, March 8th, May 8th, July 8th, and September 8th. For short-term predictions, the training dataset consists of 90 days of historical data immediately preceding each test week. The training for day-ahead forecasting uses a more extensive dataset spanning a year. This historical data is used to compute the bundles, train the forecasting models, and estimate the reconciliation matrices (see Section IV-A). Note that, in practice, BPR can be executed on a rolling basis, e.g., every few weeks, effectively re-computing bundles and re-training models using the most recent information. This allows to capture changes in the underlying dynamics, and incorporate new assets as they are added to the system.

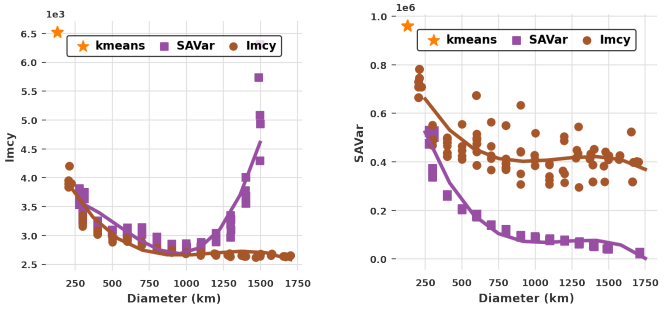


Fig. 4: Sensitivity of Imcy and SAVar with respect to diameter for three methods when considering 50 bundles. A cubic trend line is fitted for better visualization. Different bundles are obtained by varying the distance cutoff $\bar{D}$ in Algorithm 1. A single point is displayed for K-means ($\star$) because it does not support a distance cutoff.

All the experiments are conducted on a Linux cluster of dual Intel Xeon 6226 machines with 384GB of RAM equipped with RTX6000 GPUs [48]. The bundling and reconciliation steps are typically executed in a few seconds. Training times vary based on the architecture, prediction task (short-term vs day-ahead), and the number of bundles. For day-ahead predictions, training times are about 8min (resp. 40min) for an RNN model (resp. TFT) at the asset level; training times are comparable for short-term forecasting models. Note that these training times allow to re-train models on a weekly or daily schedule.

The evaluation uses two bundle sizes: 10 and 50 bundles. The diameter constraints are set to 500 kilometers for 50 bundles and 800 kilometers for 10 bundles. The evaluation trains three types of models for (i) fleet level ($K = 1$), (ii) bundle level ($K = 10$ and $K = 50$), and (iii) asset level ($K = 283$). These parameters were determined based on preliminary experiments, and allow to assess the impact of the number of bundles (K) on overall performance. K-means clustering is used as a baseline, corresponding to existing forecasting methodologies such as [19], [20]. Recall that the proposed BPR can support an arbitrary forecasting hierarchy, and an arbitrary number of bundles. Finally, forecast reconciliation is performed using MinT [49] with WLS scaling as presented in Section IV-A.

Figure 4 depicts the tradeoffs between the diameter constraints for the bundles and the quality of the bundling criteria (Imcy and SAVar). Recall that the baseline K-means clustering groups assets purely based on their pair-wise distances, i.e., it does not take any temporal information into account. Therefore, Figure 4 displays a single point for the K-means

bundling algorithm, which exhibits low diameter by design (since nearby assets are bundled together), but high variance and intermittency index due to the lack of temporal information in K-means. Figure 4 shows that it is possible to obtain low values for the intermittency and the SAVar criteria with a reasonable diameter constraint. This informed the selection of the diameter constraints specified earlier.

C. Prediction Performance

Preliminary experiments, whose results are not reported for lack of space, showed that, for the two-level hierarchy (K=1), forecast reconciliation improved overall performance. For instance, reconciliation improved the fleet-level accuracy of the RNN model by roughly 10% for short-term predictions. Accordingly, all results presented next are for reconciled forecasts. Tables II and III present the results obtained by BPR for the short-term and day-ahead forecasting tasks, respectively. All results are averaged over the five selected weeks. For each model architecture, each table reports: the number of bundles (K), the bundling criterion, and performance metrics of fleet-level forecasts (NMAE, RMSE, ED, VS) and asset-level forecasts (NMAE, RMSE, ED). Rows with $K = 1$ correspond to the baseline two-level hierarchy illustrated in Figure 1a, i.e., the corresponding forecasts are obtained by reconciling an asset-level and a fleet-level forecasts, without any bundling. A detailed analysis of each forecasting task is presented next.

a) Short-Term Predictions: Observe first that, for short-term forecasting, the persistence model emerges as a strong baseline, achieving a 7.6% NMAE at the fleet level, and a 13% NMAE and a 25.2MW RMSE at the asset level. Only the Ridge regression, transformer, and TFT models are capable of improving over the baseline. For the TFT model with 10 SAVar bundles, BPR achieves a NMAE of 5.7% at the fleet level and a 15.3% NMAE and a 23.5MW RMSE at the asset level. The best results come from reconciling asset level forecasts from the Ridge regression and bundle level forecasts from TFT (`TFT+Lin`, last row of the table). The latter was found to produce the most accurate asset-level forecast, and the former to produce the most accurate bundle-level and fleet-level forecast. In this case, BPR achieves a NMAE of 5.8% at the fleet level and a 18.8% NMAE and a 21.7MW RMSE at the asset level. When correlations are taken into account, as captured by ED and VS metrics, BPR achieves consistently better results, especially for VS at the fleet level. *In summary, BPR improves the accuracy by about 25% (NMAE, RMSE and ED) at the fleet level and by about 14% (RMSE and ED) at the asset level over the baseline, and reduces VS by 75% at the fleet level.*

When looking at a specific model (e.g., TFT), the main benefit of BPR for short-term forecasting is at the fleet level. For instance, BPR on TFT with SAVar and 10 bundles improves the fleet level accuracy by about 10% (NMAE, RMSE and ED) over the base setting ($K = 1$); VS is also improved by 6%. The improvements of BPR at the asset level through reconciliation is about 2% (RMSE), again for TFT with SAVar and 10 bundles. Note also that the bundling criteria achieve rather similar results: SAVar achieves consistently strong results overall, but there is no clear overall winner. Similarly, there is no significant difference between 10 and 50 bundles for short-term forecasting.

TABLE II: BPR for 6 Hours Forecasting: Accuracy Results.

			Fleet				Asset		
Model	K	criterion	NMAE	RMSE	ED	VS	NMAE	RMSE	ED
`Per`	–	–	7.6	2444	18797	631229	13.9	25.3	3911
`Lin`	**1**	–	6.5	2038	16370	171212	14.0	22.4	3524
	10	lmcy	6.2	1958	15637	165624	13.9	22.0	3469
		SAVar	6.1	1955	15631	167449	13.9	22.0	3471
		kmeans	6.2	1959	15662	166057	13.9	22.1	3471
	50	lmcy	6.0	1925	15200	163855	**13.7**	**21.3**	**3355**
		SAVar	6.1	1940	15407	163822	13.8	21.9	3437
		kmeans	6.0	1930	15348	165105	13.7	21.7	3406
`RNN`	**1**	–	8.1	2524	20275	199126	15.6	24.8	3928
	10	lmcy	6.7	2114	16927	175617	15.2	23.5	3752
		SAVar	7.4	2284	18374	191918	15.4	23.9	3807
		kmeans	8.1	2463	20057	196443	15.6	24.6	3904
	50	lmcy	8.6	2549	20919	199591	15.6	24.4	3863
		SAVar	7.1	2199	17534	175830	15.2	23.8	3781
		kmeans	7.7	2387	19316	202821	15.3	24.0	3807
`Tfm`	**1**	–	7.2	2054	17026	173000	15.3	22.7	3615
	10	lmcy	6.4	1878	15348	168573	15.2	22.3	3554
		SAVar	6.5	1893	15508	167283	15.2	22.3	3559
		kmeans	6.8	1956	16121	170853	15.3	22.5	3591
	50	lmcy	6.5	1885	15517	170166	15.2	22.3	3547
		SAVar	6.7	1931	15775	161601	15.2	22.4	3561
		kmeans	6.8	1950	16029	162314	15.2	22.5	3572
`TCN`	**1**	–	10.6	2811	24388	211609	19.3	27.7	4428
	10	lmcy	10.3	2725	23763	202882	19.3	27.4	4390
		SAVar	10.4	2738	23851	209404	19.3	27.5	4405
		kmeans	10.3	2730	23682	209485	19.3	27.4	4397
	50	lmcy	10.0	2672	23182	213046	19.1	27.0	4328
		SAVar	10.1	2691	23375	207217	19.1	27.3	4365
		kmeans	10.2	2735	23552	213180	19.2	27.3	4374
`TFT`	**1**	–	6.4	1899	15432	152863	15.3	23.7	3755
	10	lmcy	5.8	1753	14163	142315	15.3	23.5	3723
		SAVar	**5.7**	**1721**	**13891**	142824	15.3	23.4	3708
		kmeans	6.0	1780	14387	**140360**	15.3	23.5	3725
	50	lmcy	6.0	1778	14472	145387	15.3	23.5	3724
		SAVar	6.1	1797	14689	148588	15.3	23.5	3729
		kmeans	5.9	1746	14256	142929	15.2	23.3	3697
`TFT + Lin`	**10**	SAVar	5.8	1806	14574	152841	13.8	21.7	3423

K=1 denotes the baseline two-level hierarchy. All forecasts are reconciled.

b) Day-Ahead Predictions: The persistence model is not relevant for the day-ahead setting, so the discussion focuses on the benefits of BPR for specific models. Nonetheless, the short-term forecasting results provide a good reference point, indicating the high quality of the day-ahead forecasts.

For day-ahead forecasting, the `lmcy` criterion dominates the others (almost always) and BPR with 50 bundles is typically superior to BPR with 10 bundles. In addition, BPR always improves against the baseline (K=1), especially when considering ED and VS metrics, which capture correlations. BPR over the TFT model with 50 bundles and the `lmcy` criterion achieves the best results overall. *At the fleet level, BPR yields a 6–8% improvement in NMAE and RMSE, and a 20% improvement in ED and VS over the baseline.* At the asset level, BPR yields a 3% improvement in RMSE and ED.

D. Comparison with NREL Forecasts

The original NREL dataset [37] includes intra-day and day-ahead forecasts. To the best of the authors' knowledge, these are produced by NREL using a combination of weather forecasts, physical models [50], and unspecified machine learning

TABLE III: BPR for 48 hours Forecasting: Accuracy Results.

			Fleet				Asset		
Model	K	criterion	NMAE	RMSE	ED	VS	NMAE	RMSE	ED
RNN	**1**	–	10.5	2848	38389	1370109	19.5	28.8	6673
	10	lmcy	9.8	2655	35722	1228135	19.3	28.0	6478
		SAVar	9.8	2662	35818	1242135	19.3	28.1	6514
		kmeans	10.2	2743	36973	1283293	19.3	28.3	6562
	50	lmcy	9.2	2444	33102	1167797	18.9	26.7	6189
		SAVar	10.3	2749	37135	1296811	19.2	28.2	6525
		kmeans	10.1	2720	36611	1331087	19.1	28.1	6516
TFT	**1**	–	8.8	2480	33236	1214637	16.1	25.7	5918
	10	lmcy	8.4	2364	31586	1117189	16.0	25.3	5819
		SAVar	8.6	2391	31915	1139472	16.0	25.5	5867
		kmeans	8.6	2408	32227	1153448	16.0	25.4	5862
	50	lmcy	**8.1**	2324	30835	1099162	15.9	**24.9**	**5726**
		SAVar	8.3	2330	31178	1119034	15.9	25.2	5797
		kmeans	8.2	**2302**	**30615**	**1092823**	**15.8**	25.0	5753

$K=1$ denotes the baseline two-level hierarchy. All forecasts are reconciled.

models.[1] NREL's intra-day forecasts are produced every 6 hours, with a 6-hour lead time, 6-hour horizon, and hourly granularity. They are therefore not comparable to the short-term forecasts considered in the paper.

NREL's day-ahead forecasts are produced once a day (at midnight UTC), with an 11-hour lead time, 48-hour horizon, and hourly granularity. While not directly comparable to the day-ahead setting considered here, which has no lead time and is updated hourly, the accuracy of NREL's day-ahead forecasts nonetheless provides a sense of the quality of the BPR forecasts. Namely, NREL's day-ahead forecasts achieve, at the fleet level, a NMAE of 10.5%, an RMSE of 2954MW, an ED of 37589, and a VS of 854604 and, at the asset level, NMAE of 14.4%, an RMSE of 24.5MW, and an ED of 5469. Overall, the data-driven BPR forecasts compare favorably in terms of NMAE, RMSE and ED at the fleet level, while NREL's asset-level forecasts achieve a slightly lower NMAE and similar RMSE and ED. Recall that NREL's forecasts use advanced physical models [50], which are known to improve long-term forecasts. These results confirm the overall potential of BPR.

VI. Conclusion

The paper has proposed a novel *Bundle-Predict-Reconcile* framework for hierarchical wind power forecasting, that integrates asset bundling, machine learning and forecast reconciliation techniques. In contrast with existing hierarchical forecasting approaches, BPR leverages asset-bundling techniques to identify a hierarchy that yields improved forecast accuracy. The proposed asset-bundling methodology captures spatio-temporal characteristics of the underlying time series and is computationally fast, thus allowing to update bundles and re-train models periodically. The paper has conducted extensive numerical experiments on a large industry-size dataset, thereby evaluating the impact of BPR across different forecasting tasks (short-term and day-ahead), bundling criteria, and machine learning architectures. The results demonstrate that, compared to a baseline hierarchical setting, BPR consistently improves forecast accuracy at the fleet and asset levels, for both short-term and day-ahead forecasts.

Future work will investigate several relevant directions. First, forecast accuracy may be further improved by considering additional layers in the hierarchy, which can be achieved by recursively aggregating time series into larger and larger bundles. Furthermore, a principled framework for automatically identifying the best number of layers and bundles, would support practitioners in deploying BPR in real-life settings. Finally, as system operators face increased operational uncertainty, future work will extend BPR to probabilistic hierarchical wind power forecasting.

[1]The forecast methodology used by NREL in [37] has not been published.

References

[1] S. S. Soman, H. Zareipour, O. Malik, and P. Mandal, "A review of wind power and wind speed forecasting methods with different time horizons," in *North American Power Symposium 2010*, 2010, pp. 1–8.

[2] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy forecasting: A review and outlook," *IEEE Open Access Journal of Power and Energy*, vol. 7, pp. 376–388, 2020.

[3] Y. Zhang, J. Wang, and X. Wang, "Review on probabilistic forecasting of wind power generation," *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 255–270, 2014.

[4] Q. Xu, D. He, N. Zhang, C. Kang, Q. Xia, J. Bai, and J. Huang, "A short-term wind power forecasting approach with adjustment of numerical weather prediction input by data mining," *IEEE Transactions on Sustainable Energy*, vol. 6, no. 4, pp. 1283–1291, 2015.

[5] J. Yan and T. Ouyang, "Advanced wind power prediction based on data-driven error correction," *Energy Conversion and Management*, vol. 180, pp. 302–311, 2019.

[6] Y. Wang, Q. Hu, L. Li, A. M. Foley, and D. Srinivasan, "Approaches to wind power curve modeling: A review and discussion," *Renewable and Sustainable Energy Reviews*, vol. 116, p. 109422, 2019.

[7] Z. Niu, Z. Yu, W. Tang, Q. Wu, and M. Reformat, "Wind power forecasting using attention-based gated recurrent unit network," *Energy*, vol. 196, p. 117081, 2020.

[8] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "Deepar: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181–1191, 2020.

[9] R. Meka, A. Alaeddini, and K. Bhaganagar, "A robust deep learning framework for short-term wind power forecast of a full-scale wind farm using atmospheric variables," *Energy*, vol. 221, p. 119759, 2021.

[10] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018.

[11] B. Lim, S. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting. arxiv," *arXiv preprint arXiv:1912.09363*, 2019.

[12] A. Mashlakov, T. Kuronen, L. Lensu, A. Kaarna, and S. Honkapuro, "Assessing the performance of deep learning models for multivariate probabilistic energy forecasting," *Applied Energy*, vol. 285, p. 116405, 2021.

[13] C. Sweeney, R. J. Bessa, J. Browell, and P. Pinson, "The future of forecasting for renewable energy," *WIREs Energy and Environment*, vol. 9, no. 2, p. e365, 2020.

[14] R. J. Hyndman, R. A. Ahmed, G. Athanasopoulos, and H. L. Shang, "Optimal combination forecasts for hierarchical time series," *Computational Statistics & Data Analysis*, vol. 55, no. 9, pp. 2579–2589, 2011.

[15] S. L. Wickramasuriya, G. Athanasopoulos, and R. J. Hyndman, "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization," *Journal of the American Statistical Association*, vol. 114, no. 526, pp. 804–819, 2019.

[16] Y. Zhang and J. Dong, "Least squares-based optimal reconciliation method for hierarchical forecasts of wind power generation," *IEEE Transactions on Power Systems*, pp. 1–1, 2018.

[17] L. Bai and P. Pinson, "Distributed reconciliation in day-ahead wind power forecasting," *Energies*, vol. 12, no. 6, 2019.

[18] C. Di Modica, P. Pinson, and S. Ben Taieb, "Online forecast reconciliation in wind power prediction," *Electric Power Systems Research*, vol. 190, p. 106637, 2021.

[19] A. Stratigakos, D. van der Meer, S. Camal, and G. Kariniotakis, "End-to-end Learning for Hierarchical Forecasting of Renewable Energy Production with Missing Values," in *2022 17th International Conference on Probabilistic Methods Applied to Power Systems*, 2022, pp. 1–6.

[20] M. E. Hansen, N. Peter, J. K. Møller, and M. Henrik, "Reconciliation of wind power forecasts in spatial hierarchies," *Wind Energy*, vol. 26, no. 6, pp. 615–632, 2023.

[21] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[22] W. Katzenstein, E. Fertig, and J. Apt, "The variability of interconnected wind plants," *Energy policy*, vol. 38, no. 8, pp. 4400–4410, 2010.

[23] M. M. Bandi, "Spectrum of wind power fluctuations," *Physical review letters*, vol. 118, no. 2, p. 028301, 2017.

[24] M. Shahriari and S. Blumsack, "Scaling of wind energy variability over space and time," *Applied Energy*, vol. 195, pp. 572–585, 2017.

[25] C. Wu, X.-P. Zhang, and M. Sterling, "Wind power generation variations and aggregations," *CSEE Journal of Power and Energy Systems*, vol. 8, no. 1, pp. 17–38, Jan 2022.

[26] L. M. Hansen, "Can wind be a firm resource-a north carolina case study," *Duke Envtl. L. & Pol'y F.*, vol. 15, p. 341, 2004.

[27] B. Drake and K. Hubacek, "What to expect from a greater geographic dispersion of wind farms?—a risk portfolio approach," *Energy Policy*, vol. 35, no. 8, pp. 3999–4008, 2007.

[28] F. Roques, C. Hiroux, and M. Saguan, "Optimal wind power deployment in europe—a portfolio approach," *Energy policy*, vol. 38, no. 7, pp. 3245–3256, 2010.

[29] Y. Degeilh and C. Singh, "A quantitative approach to wind farm diversification and reliability," *International Journal of Electrical Power & Energy Systems*, vol. 33, no. 2, pp. 303–314, 2011.

[30] N. S. Thomaidis, F. J. Santos-Alamillos, D. Pozo-Vázquez, and J. Usaola-García, "Optimal management of wind and solar energy resources," *Computers & Operations Research*, vol. 66, pp. 284–291, 2016.

[31] F. Santos-Alamillos, N. Thomaidis, J. Usaola-García, J. Ruiz-Arias, and D. Pozo-Vázquez, "Exploring the mean-variance portfolio optimization approach for planning wind repowering actions in spain," *Renewable Energy*, vol. 106, pp. 335–342, 2017.

[32] O. Grothe and J. Schnieders, "Spatial dependence in wind and optimal wind power allocation: A copula-based analysis," *Energy policy*, vol. 39, no. 9, pp. 4742–4754, 2011.

[33] T. Ma and C. Li, "The electricity portfolio decision-making model based on the cvar under risk conditions," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 7, no. 3, pp. 570–575, 2014.

[34] G. Gersema and D. Wozabal, "Risk-optimized pooling of intermittent renewable energy sources," *Journal of banking & finance*, vol. 95, pp. 217–230, 2018.

[35] A. Vinel and E. Mortaz, "Optimal pooling of renewable energy sources with a risk-averse approach: Implications for us energy portfolio," *Energy Policy*, vol. 132, pp. 928–939, 2019.

[36] C. Han and A. Vinel, "Reducing forecasting error by optimally pooling wind energy generation sources through portfolio optimization," *Energy*, vol. 239, p. 122099, 2022.

[37] B. Sergi, C. Feng, F. Zhang, B.-M. Hodge, R. Ring-Jarvi, R. Bryce, K. Doubleday, M. Rose, G. Buster, and M. Rossol, "ARPA-E PERFORM datasets," 08 2022.

[38] A. Ahmadi, M. Talaei, M. Sadipour, A. M. Amani, and M. Jalili, "Deep federated learning-based privacy-preserving wind power forecasting," *IEEE Access*, vol. 11, pp. 39 521–39 530, 2023.

[39] Y. Li, R. Wang, Y. Li, M. Zhang, and C. Long, "Wind power forecasting considering data privacy protection: A federated deep reinforcement learning approach," *Applied Energy*, vol. 329, p. 120291, 2023.

[40] G. Athanasopoulos, R. J. Hyndman, N. Kourentzes, and F. Petropoulos, "Forecasting with temporal hierarchies," *European Journal of Operational Research*, vol. 262, no. 1, pp. 60–74, 2017.

[41] J. Herzen, F. Lässig, S. G. Piazzetta, T. Neuer, L. Tafti, G. Raille, T. Van Pottelbergh, M. Pasieka, A. Skrodzki, N. Huguenin *et al.*, "Darts: User-friendly modern machine learning for time series," *Journal of Machine Learning Research*, vol. 23, no. 124, pp. 1–6, 2022.

[42] O. F. Eikeland, F. D. Hovem, T. E. Olsen, M. Chiesa, and F. M. Bianchi, "Probabilistic forecasts of wind power generation in regions with complex topography using deep learning methods: An arctic case," *Energy Conversion and Management: X*, vol. 15, p. 100239, 2022.

[43] Q. Cao, B. T. Ewing, and M. A. Thompson, "Forecasting wind speed with recurrent neural networks," *European Journal of Operational Research*, vol. 221, no. 1, pp. 148–154, 2012.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[45] J. W. Messner, P. Pinson, J. Browell, M. B. Bjerregård, and I. Schicker, "Evaluation of wind power forecasts—an up-to-date view," *Wind Energy*, vol. 23, no. 6, pp. 1461–1481, 2020.

[46] M. Scheuerer and T. M. Hamill, "Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities*," *Monthly Weather Review*, vol. 143, no. 4, p. 1321–1334, 2015.

[47] European Centre for Medium-Range Weather Forecasts, "TIGGE Data Retrieval," https://apps.ecmwf.int/datasets/data/tigge/levtype=sfc/type=cf/, 2023, [Online; accessed March 2023].

[48] PACE, *Partnership for an Advanced Computing Environment (PACE)*, 2017. [Online]. Available: http://www.pace.gatech.edu

[49] G. A. Shanika L. Wickramasuriya and R. J. Hyndman, "Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization," *Journal of the American Statistical Association*, vol. 114, no. 526, pp. 804–819, 2019.

[50] J. M. Freeman, N. A. DiOrio, N. J. Blair, T. W. Neises, M. J. Wagner, P. Gilman, and S. Janzou, "System advisor model (sam) general description (version 2017.9. 5)," National Renewable Energy Lab.(NREL), Golden, CO (United States), Tech. Rep., 2018.