Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Inverted Transformers for Effective Post-Calibration in Multi-Site Wind Power Forecasting

Joseph Cohen[1] | **Tianyi Xu**[2] | **Youngchan Jang**[3] | **Pengwei Du**[4] | **Eunshin Byon**[2] | **Xun Huan**[5]

[1]Department of Mechanical and Aerospace Engineering, Rutgers University, New Jersey, United States

[2]Department of Industrial and Operations Engineering, University of Michigan, Michigan, United States

[3]Agency for Defense Development, Daejeon, South Korea

[4]Electric Reliability Council of Texas, Texas, United States

[5]Department of Mechanical Engineering, University of Michigan, Michigan, United States

Correspondence
Corresponding author: Joseph Cohen.
Email: joseph.cohen3@rutgers.edu

Funding Information
This research was supported by the National Science Foundation Grant CMMI-2226348. This research also received support from Schmidt Sciences, LLC.

Abstract

Accurate wind power forecasts are essential for energy management and resource allocation. However, because of complex weather dynamics and other nonlinearities, it is exceedingly difficult to forecast wind power on the multi-site level for dozens of wind farms at once. This paper proposes a hybridized approach that leverages deep learning to predict future forecast errors from physics-based numerical weather prediction (NWP) model estimates. Utilizing errors from NWP forecasts allows integration of critical atmospheric and meteorological dynamics into the forecasting model, and we demonstrate the importance of post-calibration based on the physics versus pure data-driven wind power prediction. This post-calibration approach is enabled by the inverted Transformer architecture, which efficiently and effectively learns meaningful wind farm variate representations resulting in accurate spatiotemporal corrections to the forecasts. We also investigate modifying the iTransformer with a new embedding approach, named SpaceEmbed, that explicitly encodes spatial distance information into the network. The proposed approach is validated with a case study using real-world data and forecasts from the Electric Reliability Council of Texas (ERCOT) in 2015 for 74 wind farms in Texas at different time scales. Using the high sustained limit as the metric for power generation, the iTransformer outperforms other state-of-the-art deep learning forecasting methods, succeeding at the post-calibration task by reducing NWP forecast error by up to 33% on average.

KEYWORDS
Deep learning, wind power prediction, time series forecasting

1 | INTRODUCTION

Electrification in modern society has accelerated demand for clean and renewable sources of energy. Over the last few decades, wind energy has emerged as a potent and reliable alternative, with China and the United States accounting for over 50% of the global cumulative wind capacity as of 2020 (Wang et al. 2021c). Accurate wind forecasting is essential for utilities' energy management and allocation, with massive economic implications. For example, a recent study estimated that National Oceanic and Atmospheric Administration (NOAA)'s High Resolution Rapid Refresh (HRRR) wind forecasts have saved consumers approximately $150 million USD on energy every year (Jeon et al. 2022). As stated in a technical report for the National Renewable Energy Laboratory (NREL), the economic benefits from improving short-term wind power forecasts such as those provided by the Electric Reliability Council of Texas (ERCOT) can lead to further savings for load payments and energy imbalance payments (Orwig et al. 2012).

Wind power generation relies on several dynamic meteorological factors such as wind speed, wind direction, humidity, temperature, and other nonlinearities; as a result, wind power forecasting remains a challenging and active area of research (Ding 2019, Byon et al. 2016, You et al. 2017). Classically, many wind power forecasting approaches have been on the turbine level, with more recent work attempting to forecast on the wind farm level (Wang et al. 2021c). Forecasting techniques can be categorized as statistical, physics-based, machine learning, or hybrid approaches (Wang et al. 2021a). Statistical methods such

as autoregressive integrated moving average (ARIMA) have been popular for time series forecasts, especially on short time scales (Eldali et al. 2016). On the other hand, physics-based numerical weather prediction (NWP) models focus on solving the governing partial differential equations, and have become standard tools for resolving complex weather dynamics and predicting wind speed and power, with potential for accurate longer term forecasts (Wang et al. 2011).

Machine learning, especially deep learning, has become exceedingly popular for time series forecasting problems in recent years. For example, deep convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and Transformer networks have been employed to autonomously learn nonlinear and predictive feature representations. Demolli et al. (2019) showed that such data-driven techniques could be useful for predicting wind power at unknown geographical locations. More recently, Wang and He (2022) illustrated ultra-short-term wind power predictions that capture highly localized spatiotemporal dynamics from historical data trends. The Transformer architecture, strengthened by multi-head attention mechanisms, has also become increasingly popular for sequential modeling, enhancing time series forecasting for wind speed and wind power applications (Sun et al. 2023, Qu et al. 2022, Wu et al. 2022a, Li et al. 2023, Pan et al. 2022, Wang et al. 2022). For example, Sun et al. (2023) demonstrated the effectiveness of Transformer networks at learning sequential relationships for short-term wind power forecasting from spatiotemporal correlations.

Hybrid methods that combine machine learning with physics-based NWP have also become prominent in recent years. Du (2018) investigated an ensemble learning approach that showed the advantages of integrating NWP information alongside meteorological covariates (e.g., wind speed, wind direction, temperature, relative humidity, barometric information). Elsewhere, Ye et al. (2023) demonstrated the effectiveness of utilizing NWP information in a probabilistic model for wind speed and wind power forecasts under ultra-short prediction horizons. These authors subsequently proposed correcting NWP biases with a machine learning approach guided by wind advection and diffusion physics (Ye et al. 2024). Despite these advancements, however, most existing short-term wind power forecasting approaches based on machine learning and hybrid methods have only considered predictions at the single-site level.

Training a single deep learning model to simultaneously forecast multiple wind farms has several distinct advantages. Firstly, the forecasts themselves may benefit from spatiotemporal connections between wind farms explicitly or implicitly encoded in the model. Secondly, training and updating a single model is more computationally efficient than needing to train and update a separate model for each site. Lastly, such a model can provide a "big picture" for decision making, providing practical insight when gauging energy management over a large geographical area (e.g., over an entire state). Qu et al. (2022) proposed using the Transformer architecture for wind power predictions at multiple wind farms, validating on a case study of 6 wind farms in northwest China. In a study for photovoltaic power forecasting, Simeunović et al. (2021) proposed a spatiotemporal graph attention network architecture for multiple sites that outperformed NWP-based models for 3- and 6-hour ahead predictions. However, as discussed by Liu et al. (2024), sequential models like the base Transformer may struggle with multivariate learning—such as in multi-site wind power forecasting—since the temporal tokens learned can muddle the information aggregated from different variates. To address this issue, Liu et al. (2024) proposed the *iTransformer* architecture that inverts the tokenizing dimension in the Transformer network, leading to "variate-centric" representations.

In this paper, we propose a new hybrid method that combines deep learning and NWP for *multi-site* time series forecasting. We argue that deep learning can improve short-term multi-site wind power forecasts via directly learning and compensating for the error from NWP projections. Specifically, we study the effectiveness of iTransformer for learning variate-centric NWP error representations, which allows for the user to receive recommendations of spatiotemporal corrections to the forecasts over different time scales. We benchmark and compare the iTransformer setup against other state-of-the-art models dedicated for time series forecasting—including models that explicitly model other static (e.g., spatial coordinates) and non-static covariates (e.g., wind speed and wind direction)—under a case study using ERCOT data from 2015 to simultaneously predict NWP error from 74 wind farm stations spread over a wide geographical area encompassing the state of Texas. Our main contributions are summarized as follows.

1. We present the iTransformer architecture and introduce the modified SpaceEmbed variant for learning spatiotemporal trends of NWP errors.
2. We demonstrate the effectiveness of the iTransformer-based approaches on a case study using ERCOT data from 74 wind farms with varying prediction horizons, with the approach resulting in an average improvement of up to 33% over the original NWP forecast.

We structure the paper as follows. Section 2 will present the proposed iTransformer deep learning methodology for multi-site wind power forecasting. Section 3 will detail the ERCOT case study, offering more comparisons between our approach and other state-of-the-art algorithms. We will then provide practical recommendations and discussions on the approach in Section 4, and conclude the paper in Section 5.

2 | METHODOLOGY

In this section, we will first formulate the time series forecasting problem. Then, we will expand on the base Transformer components to introduce the iTransformer, including augmentations with additional static covariates such as latitude and longitude coordinates.

2.1 | Problem Statement

Consider a multivariate time series for N wind farm stations, $\mathbf{Y}_{1:T+K} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T, \mathbf{y}_{T+1}, \ldots, \mathbf{y}_{T+K}\} \in \mathbb{R}^{N\times(T+K)}$. We are interested in the following problem: at time T given historical partial series $\mathbf{Y}_{1:T} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$, forecast the future series for $k = 1, 2, \ldots, K$ hourly steps ahead, $\mathbf{Y}_{T+1:T+K} = \{\mathbf{y}_{T+1}, \mathbf{y}_{T+2}, \ldots, \mathbf{y}_{T+K}\}$. These quantities could be any wind power metric of interest, such as wind power measured in MW or a proxy metric like wind speed. In our case study, we will use high-sustained limit (HSL)—the maximum rated capability for power generation from a resource at a given point in time (Taheri 2012)—updated hourly. We will consider both the single-step ahead case ($K = 1$), as well as multi-step ahead cases ($K > 1$) that produce multi-hourly forecasts for all wind farms.

Due to meteorological nonlinearities and complex physics, it is difficult to reliably forecast $\mathbf{Y}_{T+1:T+K}$ using data-driven autoregressive methods alone, particularly when N (the number of wind farms) and K (the maximum prediction horizon in hours) are large. Therefore, we propose to leverage available NWP short-term wind power forecasts (STWPF) as follows. Let $\widehat{\mathbf{Y}}_{1:T+K} = \{\widehat{\mathbf{y}}_T^{(1)}, \widehat{\mathbf{y}}_T^{(2)}, \ldots, \widehat{\mathbf{y}}_T^{(K)}\}$ denote the time series of predictions from a particular NWP model, where $\widehat{\mathbf{y}}_T^{(k)}$ is the k-step ahead forecast made at time T. We can then define the corresponding errors of the NWP model predictions:

$$\begin{aligned}\mathbf{B}_{T+1:T+K} &= \left\{\mathbf{b}_T^{(1)}, \mathbf{b}_T^{(2)}, \ldots, \mathbf{b}_T^{(K)}\right\} \\ &= \left\{\mathbf{y}_{T+1} - \widehat{\mathbf{y}}_T^{(1)}, \mathbf{y}_{T+2} - \widehat{\mathbf{y}}_T^{(2)}, \ldots, \mathbf{y}_{T+K} - \widehat{\mathbf{y}}_T^{(K)}\right\} \\ &= \mathbf{Y}_{T+1:T+K} - \widehat{\mathbf{Y}}_{T+1:T+K}. \end{aligned} \tag{1}$$

Here, $\mathbf{b}_T^{(k)}$ represents the prediction error of the NWP model when the forecast is made at time T, that is, $\mathbf{b}_T^{(k)} = \mathbf{y}_{T+k} - \widehat{\mathbf{y}}_T^{(k)}$ for $k = 1, 2, \ldots, K$. We seek to train a data-driven model f to forecast the future errors from its past history, $\mathbf{B}_{T+1:T+K} \approx \widehat{\mathbf{B}}_{T+1:T+K} = f(\mathbf{B}_{1:T})$ (see Fig. 1), and use it to correct upon the NWP baseline:

$$\begin{aligned}\mathbf{Y}_{T+1:T+K} &= \widehat{\mathbf{Y}}_{T+1:T+K} + \mathbf{B}_{T+1:T+K} \\ &\approx \widehat{\mathbf{Y}}_{T+1:T+K} + \widehat{\mathbf{B}}_{T+1:T+K}. \end{aligned} \tag{2}$$

The intuition behind such a "post-calibration" approach is that the error terms, i.e., differences between the measurements and the NWP baseline, may be easier to capture through f compared to $\mathbf{Y}_{T+1:T+K}$ directly since the "bulk" dynamics is expected to be already informed by the NWP's forecast $\widehat{\mathbf{Y}}_{T+1:T+K}$. The overall problem thus distills to building f, for example, via iTransformer or other deep learning model architectures.

Our approach is related to previous data-driven post-calibration efforts (Jang et al. 2022, Jeong et al. 2023, Jeong and Byon 2024, Jain et al. 2023), including those that use deep learning to accomplish spatiotemporal post-calibration (Grönquist et al. 2021, Laloyaux et al. 2022, Cho et al. 2020). Peng et al. (2016) proposed using machine learning techniques such as neural networks and support vector machines to predict error corrections to wind power forecasts. Furthermore, Costoya et al. (2020) and Liu et al. (2016) proposed methods for wind power forecasting based on NWP corrections for wind speed forecasts, using techniques such as linear regression and neural networks. However, our approach differs from the established literature in that we use a single model for correcting multi-site HSL forecasts. By learning the error corrections for dozens of wind farms simultaneously, the model is able to account for spatiotemporal correlations and variations in NWP error.

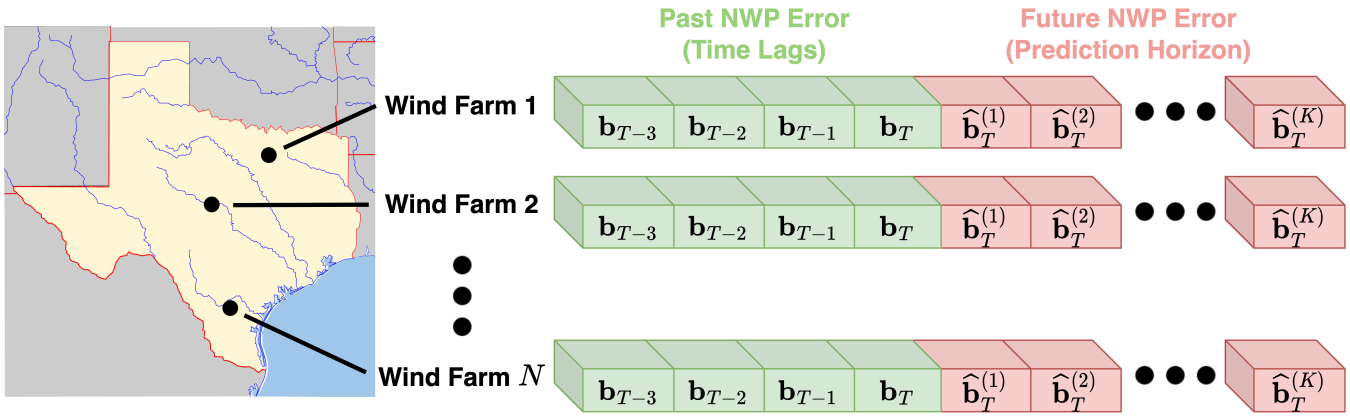

FIGURE 1 Summary of problem formulation for simultaneous multi-site forecast of NWP error. Note: $\mathbf{b}_T, \mathbf{b}_{T-1}, \mathbf{b}_{T-2}, \mathbf{b}_{T-3}$ represent the input series of NWP forecast errors, counting backwards from time T.

2.2 | iTransformer Architecture for Post-calibrating NWP Forecasts

The iTransformer architecture proposed by Liu et al. (2024) retains the core element of the popular Transformer model: attention. However, it differs in its encoder-only architecture that emphasizes learning representations of each variate using self-attention, in contrast to the original encoder-decoder architecture proposed for natural language processing applications (Vaswani et al. 2017). Most significantly, iTransformer's tokenizing dimension is "inverted" (i.e., transposed) from the standard Transformer, tokenizing along the variate dimension instead.

The iTransformer architecture, visualized in Fig. 2, can be summarized by three subcomponents: (input) embedding layer, attention-block, and (output) projection layer, as follows:

$$\mathbf{H}^0 = \{\mathbf{h}_1^0, \mathbf{h}_2^0, \ldots, \mathbf{h}_N^0\} = \text{Embed}(\mathbf{B}_{1:T}), \tag{3}$$

$$\mathbf{H}^\ell = \text{AttnBlock}(\mathbf{H}^{\ell-1}), \quad \ell = 1, 2, \ldots, L, \tag{4}$$

$$\widehat{\mathbf{B}}_{T+1:T+k} = \text{Proj}(\mathbf{H}^L). \tag{5}$$

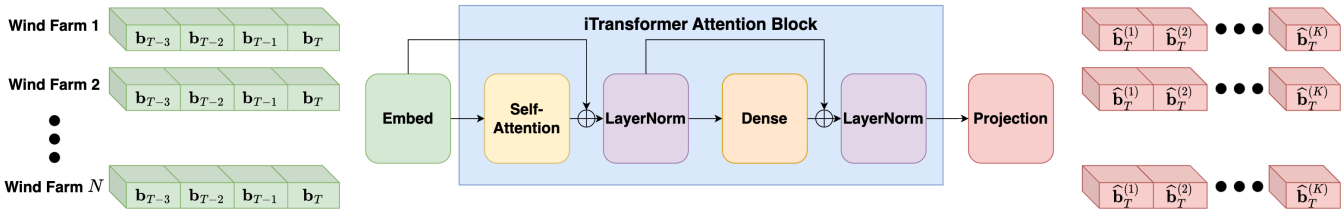

FIGURE 2 iTransformer architecture complete with subcomponents of embedding layer, attention-block, and output projection layer.

The embedding layer encodes the raw input $\mathbf{B}_{1:T}$ into latent variables $\mathbf{H}^0$ composed of tokens $\mathbf{h}_i^0 \in \mathbb{R}^{d_m}, i = 1, \ldots, N$, where N is the number of wind farms. The token dimension, d_m, is a tunable hyperparameter. We consider two embeddings in this work: Embed, the default iTransformer value embedding based on a single dense linear layer; and SpaceEmbed, an alternative embedding we propose to explicitly incorporate the spatial coordinates of wind farms:

$$\mathbf{H}^0 = \text{SpaceEmbed}(\mathbf{B}_{1:T}, \mathbf{S}) = \text{Embed}(\mathbf{B}_{1:T}) + \gamma \text{CoordEmbed}(\mathbf{S}), \tag{6}$$

where CoordEmbed are again simple dense linear layers, and γ is a trainable parameter weighing the spatial contributions to the embedding learned automatically via backpropagation. $\mathbf{S} \in \mathbb{R}^{N \times N}$ is the static symmetric distance matrix whose (i,j)-th

element is

$$\mathbf{S}_{ij} = 2r \arcsin \sqrt{\frac{1 - \cos(\lambda_i - \lambda_j) + \cos\lambda_i \cos\lambda_j (1 - \cos(\phi_i - \phi_j))}{2}}, \tag{7}$$

with λ_i and ϕ_i being the latitude and longitude of wind farm i in radians for $i,j = 1,2,\ldots,N$, and r is the radius of the Earth (6371 km). Specifically, $\mathbf{S}_{ij}$ is obtained using the Haversine formula—a spherical distance metric that accounts for the curvature of Earth (Maria et al. 2020). After computing the distance matrix, the elements of $\mathbf{S}_{ij}$ are normalized in the range [0,1] by dividing by the maximum distance.

Following the initial embedding, a sequence of L attention-blocks successively transform $\mathbf{H}^0$ to uncover its contextual representations. The cornerstone of each block is the multi-head attention mechanism proposed by Vaswani et al. (2017), which allows the network to "focus" on relevant information across the different wind farm station variates. In the self-attention layers, tokens are decomposed into queries, keys, and values by multiplying them with learnable weight matrices. This enables the computation of attention scores via scaled dot products, facilitating the identification of relationships between the variates. Such mechanisms also have the the potential to improve interpretability, as discussed by Liu et al. (2024). The self-attention layers are further connected by normalizing layers that perform z-score standardization on each individual variate. The normalization stabilizes training and convergence, and due to the inversion architecture, avoids mixing together information across variates at a given timestamp—an undesired drawback in the original Transformer design when applied for multi-site forecasting problems. Fully connected dense layers are added in between normalization layers, completing the block architecture.

Finally, a projection layer transforms the latest $\mathbf{H}^L$ to the output $\widehat{\mathbf{B}}_{T+1:T+K}$. For iTransformers, the projection layer is a simple linear dense layer rather than a more complex decoder, and thus often referred to be encoder-only.

3 | RESULTS

This section presents detailed implementation of the iTransformer model, complete with results and comparisons with other state-of-the-art deep learning architectures on a case study of wind farm data provided by ERCOT. We also provide comparisons with and without the SpaceEmbed variant for fusing spatial features, and other models that explicitly include other covariates such as wind speed and wind direction.

3.1 | ERCOT Case Study Description

The dataset provided by ERCOT includes both (normalized) HSL measurements and (normalized) HSL predictions obtained using the $K = 1,2,3,6$, and 12-hour ahead NWP models for 74 wind farms in Texas and for every hour during the year 2015 (see Fig. 3). Updated hourly, the HSL is a useful proxy for the physical capability of producing energy without the influence of external market factors (e.g., supplementary energy dispatches), making it preferable over conventional MW telemetry for forecasters (Taheri 2012).

For the forecasting problem, we set the length of the historical partial series (i.e., input sequence) to 48 regardless of K—that is, we always use the previous two days' data to make future forecasts. The choice for this parameter was motivated from short-term forecasting models benchmarked on the popular M4 dataset (Wang et al. 2024). Under this forecasting benchmark, the deep learning models were evaluated on sequence length values ranging from 6 to 48. As a result, 48 hours was chosen as a fixed setting for input length. We chose the highest benchmarked value because it would provide the most context for predicting NWP errors.

We note that proprietary software was used for the NWP simulations provided by ERCOT in this study. As a result, we are unable to share internal modeling parameters, and the focus of the case study is therefore to correct NWP forecasting errors *without needing to have direct access to the NWP simulation or potentially sensitive parameters*. This approach may be especially useful in situations where the privacy of sensitive parameters must be preserved.

The full dataset is divided into training and testing sets, and we ensure there is no overlap between any of the sequences in those sets (or any part of those sequences). The training period entails the first 10 months of 2015, and the testing period is the last 2 months. We further utilize the last 20% of the training set as a hold-out validation set for hyperparameter optimization and early stopping. Table 1 summarizes the training, validation, and testing periods and the number of sequences in each, which is

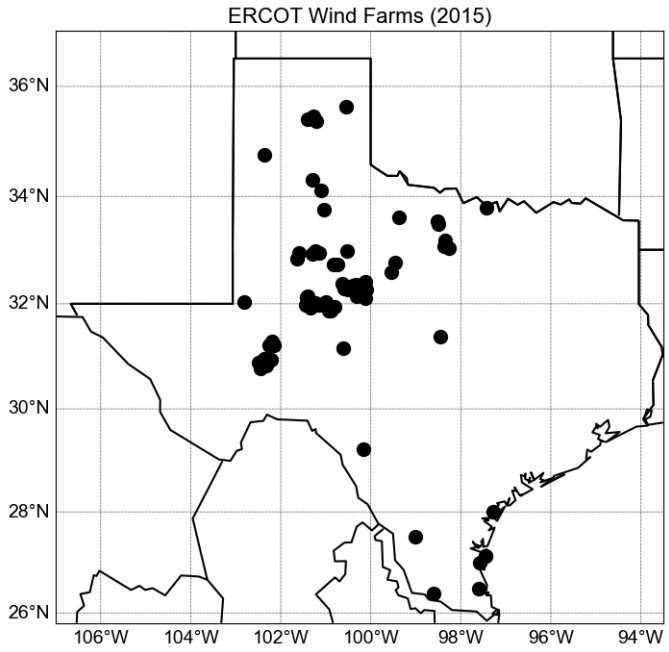

FIGURE 3 Location of 74 wind farms in Texas included in the 2015 ERCOT dataset.

dependent on k. We note that this case study features relatively small sample sizes in the context of deep learning, adding to the challenge of multi-site forecasting.

TABLE 1 Training, validation, and testing set information. The number of sequences is dependent on the forecast horizon, k. The dataset is split to ensure that there are no identical sequences between the subsets, and no overlap between the end of the training/validation period and the beginning of the testing set.

	Period	Number of Sequences
Training	2015-01-01 – 2015-08-30	$0.8(7248 - K)$
Validation	2015-08-30 – 2015-10-30	$0.2(7248 - K)$
Testing	2015-11-01 – 2015-12-30	$1416 - K$

3.2 | Deep Learning Results for NWP Post-Calibration

The training procedure and training hyperparameters are kept consistent across all deep learning models. While the hyperparameters are determined from a rough random search based on the validation set, results could be potentially improved further with a more extensive hyperparameter optimization. During training, mean-squared error (MSE) loss is minimized using ADAM. Mini-batching is employed with batch size set to 1024. From our hyperparameter tuning analysis, we find training to reasonably stabilize under just a few epochs, aligning with other deep time series models (Zhou et al. 2021, Wang et al. 2024); thus, we set the maximum training epoch to five with an early stopping patience parameter of three.

The initial learning rate is set to 0.001, exponentially decaying by a factor of 0.5 after the first two epochs. All models are trained on a PC laptop with an Intel i7-10750H CPU @ 2.60 GHz, 32 GB of RAM, and a NVIDIA GeForce RTX 2070 Super GPU. The Time Series Library (TSLib) (Wang et al. 2024) in Pytorch is utilized for implementing other deep learning models for comparison.

Specifically for the iTransformer and the proposed SpaceEmbed variant, we choose $d_m = 128$ for tokenization. γ was initialized to be 0.01, weighing the CoordEmbed portion of the embedding from two dense linear layers (with 64 and $d_m = 128$ units

in the hidden layers). The model architecture contains just two encoding layers ($L = 2$) and one output projection layer, keeping a parameter-efficient approach overall that appropriate for the dataset size. The embedding and projection components all use linear dense layers, as discussed in the previous section. For the attention-block, the self-attention layers adopt multi-head attention with eight heads, and the dense layers incorporate the Gaussian Error Linear Units (GELU) function for nonlinear activations (Hendrycks and Gimpel 2016) with a dropout probability of 0.1 to mitigate overfitting.

Table 2 reports the test performance of the final post-calibrated HSL forecasts, $\widehat{\mathbf{Y}}_{T+1:T+K} + \widehat{\mathbf{B}}_{T+1:T+K}$ from Equation (2), using the original iTransformer, the SpaceEmbed variant of iTransformer, and the other benchmark deep learning models for all values of K tested. We also included a model output statistics (MOS) approach as a baseline comparison via the linear regression modeling of $E[\mathbf{y}_T] = \beta_0 + \beta_1 \widehat{\mathbf{y}}_T$, where $\mathbf{y}_T$ and $\widehat{\mathbf{y}}_T$ are the observed and NWP-forecasted power at time T, respectively (Jahn et al. 2019). Here, β_0 and β_1 are the location and scale parameters to adjust the NWP forecast $\widehat{\mathbf{y}}_T$. Specifically, we report the mean test RMSE $\pm$ 1 standard deviation aggregated over the 74 wind farms. Results indicate that iTransformer models hold significant improvements over the NWP forecast and post-calibration using other deep learning models. The extent of improvement varies depending on K, with greater improvement observed under shorter time scales. Moreover, the standard deviations from the iTransformer models are much smaller than those from the original NWP baseline and other models, indicating a substantial decrease in forecast uncertainties.

TABLE 2 Test performance for forecasting HSL at different forecast horizons, K, using different deep learning models in post-calibrating NWP. Reported values are the mean station-wide test RMSE $\pm$ 1 standard deviation averaged over the 74 wind farms.

Model	$K = 1$	$K = 2$	$K = 3$	$K = 6$	$K = 12$
Autoformer (Wu et al. 2022b)	0.187 ± 0.018	0.202 ± 0.020	0.198 ± 0.021	0.201 ± 0.026	0.215 ± 0.030
DLinear (Zeng et al. 2022)	0.125 ± 0.020	0.142 ± 0.020	0.168 ± 0.018	0.182 ± 0.022	0.207 ± 0.026
FEDformer (Zhou et al. 2022)	0.182 ± 0.030	0.208 ± 0.038	0.202 ± 0.035	0.193 ± 0.028	0.218 ± 0.032
Informer (Zhou et al. 2021)	0.133 ± 0.022	0.144 ± 0.023	0.150 ± 0.024	0.163 ± 0.027	0.183 ± 0.033
iTransformer (Liu et al. 2024)	$\mathbf{0.095 \pm 0.014}$	$\mathbf{0.110 \pm 0.017}$	$\mathbf{0.122 \pm 0.019}$	$\mathbf{0.143 \pm 0.023}$	$\mathbf{0.169 \pm 0.030}$
SpaceEmbed	$\mathbf{0.095 \pm 0.014}$	$\mathbf{0.110 \pm 0.017}$	$\mathbf{0.122 \pm 0.019}$	$\mathbf{0.142 \pm 0.023}$	$\mathbf{0.168 \pm 0.030}$
Nonstationary Transformer (Liu et al. 2023)	0.126 ± 0.023	0.139 ± 0.024	0.148 ± 0.025	0.163 ± 0.028	0.184 ± 0.033
PatchTST (Nie et al. 2023)	0.100 ± 0.015	0.113 ± 0.017	0.124 ± 0.019	0.144 ± 0.023	0.169 ± 0.030
Pyraformer (Liu et al. 2021)	0.124 ± 0.022	0.133 ± 0.024	0.141 ± 0.025	0.154 ± 0.027	0.175 ± 0.033
TimesNet (Wu et al. 2023)	0.127 ± 0.024	0.139 ± 0.024	0.145 ± 0.025	0.158 ± 0.027	0.176 ± 0.033
Transformer (Vaswani et al. 2017)	0.139 ± 0.025	0.150 ± 0.025	0.157 ± 0.025	0.167 ± 0.027	0.184 ± 0.032
Linear Regression	0.162 ± 0.025	0.194 ± 0.029	0.214 ± 0.034	0.247 ± 0.044	0.291 ± 0.053
NWP Baseline	0.139 ± 0.023	0.146 ± 0.023	0.152 ± 0.024	0.162 ± 0.027	0.178 ± 0.031

Notably, iTransformer is significantly more effective for post-calibration at $K = 6$ and $K = 12$ compared to other models considered. It also reduces the errors from the baseline NWP forecast. This is particularly important given that NWP forecasts tend to be less accurate over longer time scales, making their inputs to iTransformer less predictive compared to shorter time scales. It is well-known that longer prediction horizons (higher values of K), specifically for $K = 6$ or 12, present considerable challenges. Nevertheless, iTransformer demonstrates its capability to correct the biases present in NWP forecasts for these extended forecasting horizons. While the performance of PatchTST is comparable for $K = 6$ and 12, iTransformer consistently exhibits superior performance across all forecasting horizons.

Taking a closer look at the results reported in Table 2, the two leading models for post-calibration are iTransformer and PatchTST, but the difference in performance between these two approaches is greatest at $K = 1$. Figure 4 illustrates the NWP forecast improvement when using the iTransformer approach compared to the next-best approach (PatchTST) in this 1-hour ahead setting. While the forecast improvement over NWP is not uniform across all wind farms, every wind farm experiences at least a 22% reduction in RMSE and upwards to 62% (with the average error reduction being approximately 33%) when adopting the iTransformer. Notably, the improvement brought by iTransformer is greater than that of PatchTST at *every* wind farm. Overall, the forecast accuracy is substantially increased in central Texas, where the wind farms are densely located. This provides evidence that the iTransformer architecture is better at learning and utilizing the spatial correlations in correcting the NWP forecasts.

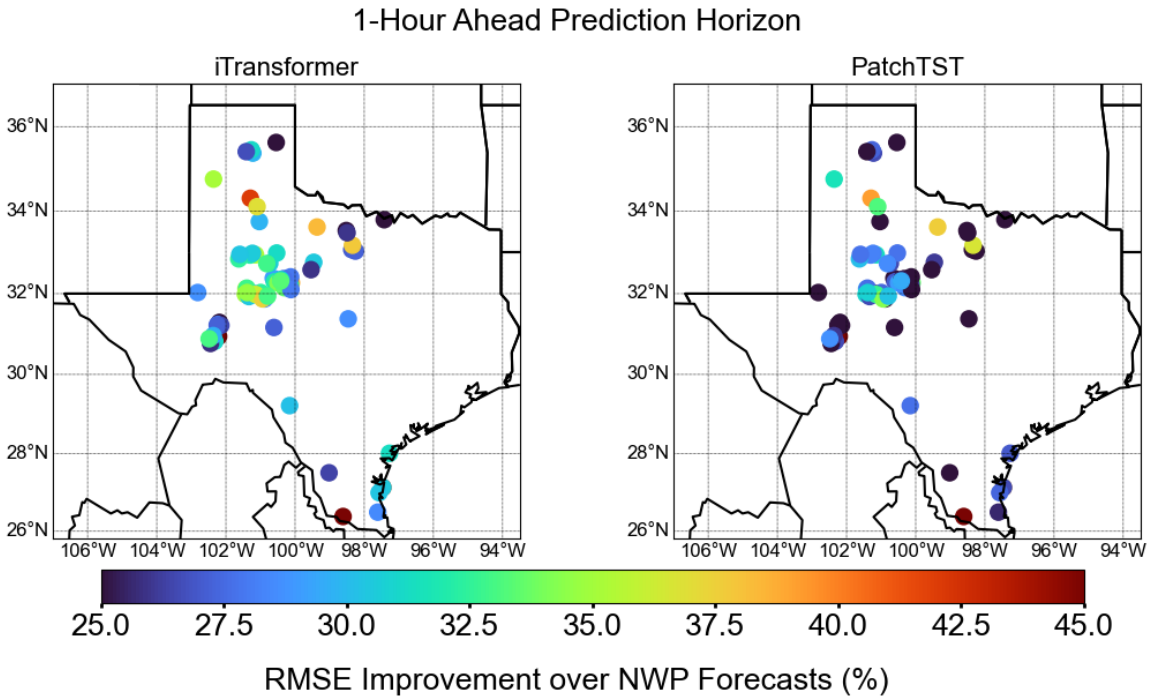

FIGURE 4 Percent improvement over NWP in RMSE across all 74 wind farms in Texas for 1-hour ahead predictions, comparing iTransformer-based post-calibration to the next-best PatchTST approach.

Figure 5 further presents the HSL error distributions for the iTransformer and the NWP predictions for the same 1-hour ahead setting over all test samples and all wind farms. It is noteworthy that the iTransformer post-calibration improves the bias over the physics-based NWP baseline model. The mean error in normalized HSL (actual - predicted) for the NWP baseline is -2.4×10^{-3} and the iTransformer model improves upon this overestimation with a mean error of just -3.6×10^{-4}. Correcting the NWP forecast is nontrivial, as evident from Table 2 where most of the state-of-the-art deep learning models fail to improve the NWP forecasts for longer prediction horizons (e.g., $K = 6$ and $K = 12$). The following subsection will provide additional comparisons using the same deep learning architectures for predicting the HSL metric directly without utilizing NWP forecast information in the inputs. Additionally, iTransformer post-calibration produces a sharper error distribution, indicating smaller prediction uncertainties and a reduced number of large prediction errors.

Furthermore, we note that the performance difference between the base iTransformer and its SpaceEmbed variant that explicitly incorporates spatial distance information into the network is subtle. While the mean RMSE and variance are nearly identical for all values of K, upon closer inspection, the SpaceEmbed variant enables slightly better improvement for $K = 6$ and $K = 12$ when post-calibrating NWP forecasts. The improvement can be shown more clearly when examining on the region-by-region level. Fig. 6 illustrates this for $K = 6$, which compares boxplots of the SpaceEmbed and iTransformer performance across wind farms in south, central, and north Texas. The SpaceEmbed variant slightly outperforms iTransformer for all 3 evaluated regions.

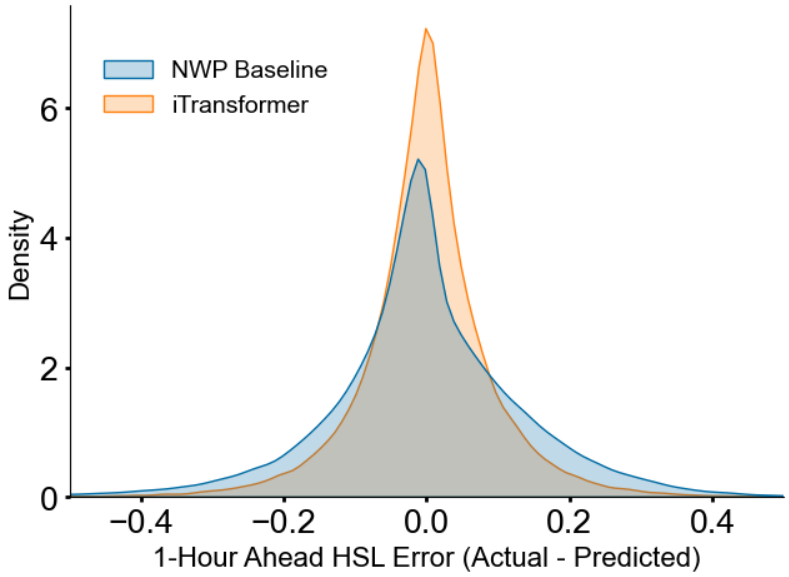

FIGURE 5 HSL errors with iTransformer corrections at 1-hour ahead prediction horizon, illustrating significant error compensation and reduction of variance.

These results indicate that the base iTransformer architecture is quite successful at implicitly learning these spatial correlations, but explicitly encoding coordinate information with approaches such as SpaceEmbed has the potential to offer additional advantages. Including the trainable parameter γ (see Eq. 6), which is learned automatically via backpropagation, enables the architecture to find the right balance between the implicit and explicit encoding mechanisms. These advantages may be further accented if the dataset only contains a few wind farms that are sparsely located, inhibiting the performance of iTransformer in its ability to implicitly learn spatial information. The next subsection will also compare four deep learning models that incorporate additional covariates.

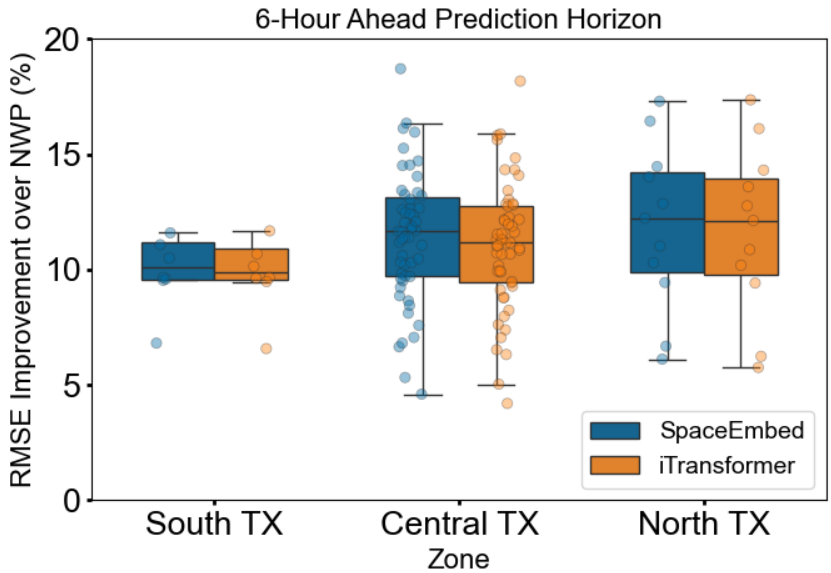

FIGURE 6 Boxplots comparing post-calibrated iTransformer and SpaceEmbed performance for $K = 6$, demonstrating regional performance improvements over NWP baseline.

3.3 | Additional Comparisons

Predicting HSL solely from sequential historical HSL estimates is difficult, as HSL is highly dependent on meteorological factors (Taheri 2012). The physics-based NWP forecasts simplify this problem by resolving some of the complex and nonlinear time-varying weather dynamics. However, it is helpful to gauge the impact of the NWP inputs when considering the viability of purely data-driven deep learning forecasts. Table 3 reports the test performance from all 74 wind farms when directly predicting HSL instead. We include the NWP results as reference in the last row in Table 3.

From $K = 3$ and upwards, all purely deep learning-based forecasts are significantly less accurate than NWP. These findings align with the general trend that purely data-driven models typically could outperform physics-based NWP models only with the shortest prediction horizons, while their performance degrades over longer time scales (Ding 2019). For prediction with $K = 1$ and $K = 2$, the iTransformer model outperforms the NWP forecast. While PatchTST also surpasses the NWP baseline model for $K = 1$ and $K = 2$, its performance is slightly inferior to that of iTransformer.

It is noteworthy that while iTransformer achieves a significant improvement over the NWP forecast at $K = 1$ even without post-calibration, all the implemented deep learning models show enhanced performance when used in conjunction with the NWP baseline (see Table 2). This observation suggests that the physical principles captured by the NWP model play a crucial role in forecast accuracy. As discussed in Section 3.2, combining NWP forecasts with deep learning post-calibration proves to be a more effective approach compared to relying solely on data-driven methods or purely on NWP forecasts.

TABLE 3 Test performance for forecasting HSL at different forecasting horizons, *K*, using different deep learning models directly without post-calibrating NWP. Reported values are the mean station-wide test RMSE $\pm$ 1 standard deviation averaged over the 74 wind farms.

Model	K = 1	K = 2	K = 3	K = 6	K = 12
Autoformer (Wu et al. 2022b)	0.297 ± 0.041	0.321 ± 0.040	0.314 ± 0.041	0.320 ± 0.043	0.332 ± 0.043
DLinear (Zeng et al. 2022)	0.235 ± 0.037	0.248 ± 0.039	0.258 ± 0.040	0.279 ± 0.042	0.311 ± 0.046
FEDformer (Zhou et al. 2022)	0.295 ± 0.043	0.315 ± 0.041	0.315 ± 0.043	0.314 ± 0.041	0.331 ± 0.044
Informer (Zhou et al. 2021)	0.209 ± 0.044	0.215 ± 0.044	0.221 ± 0.043	0.239 ± 0.041	0.270 ± 0.041
iTransformer (Liu et al. 2024)	$\mathbf{0.109 \pm 0.018}$	$\mathbf{0.133 \pm 0.023}$	$\mathbf{0.154 \pm 0.026}$	$\mathbf{0.198 \pm 0.033}$	$\mathbf{0.248 \pm 0.040}$
SpaceEmbed	$\mathbf{0.109 \pm 0.018}$	$\mathbf{0.135 \pm 0.023}$	$\mathbf{0.158 \pm 0.027}$	$\mathbf{0.200 \pm 0.033}$	$\mathbf{0.249 \pm 0.040}$
Nonstationary Transformer (Liu et al. 2023)	0.175 ± 0.037	0.211 ± 0.041	0.223 ± 0.042	0.251 ± 0.042	0.281 ± 0.044
PatchTST (Nie et al. 2023)	0.115 ± 0.017	0.139 ± 0.021	0.159 ± 0.024	0.205 ± 0.032	0.251 ± 0.039
Pyraformer (Liu et al. 2021)	0.165 ± 0.034	0.177 ± 0.035	0.188 ± 0.037	0.215 ± 0.039	0.254 ± 0.043
TimesNet (Wu et al. 2023)	0.180 ± 0.041	0.198 ± 0.042	0.204 ± 0.041	0.227 ± 0.042	0.248 ± 0.041
Transformer (Vaswani et al. 2017)	0.279 ± 0.057	0.295 ± 0.055	0.300 ± 0.054	0.308 ± 0.052	0.308 ± 0.054
Linear Regression	0.169 ± 0.028	0.207 ± 0.036	0.233 ± 0.043	0.301 ± 0.056	0.387 ± 0.075
NWP Baseline	0.139 ± 0.023	0.146 ± 0.023	0.152 ± 0.024	0.162 ± 0.027	0.178 ± 0.031

Additional covariate information such as wind speed and wind direction are also available in the ERCOT dataset, and four alternative deep learning models for post-calibration incorporating these covariates alongside spatial coordinates are compared. Unlike iTransformer, these models have existing capabilities for handling static covariates (such as spatial coordinates) and other past covariate series (wind speed and wind direction) that are explicitly separate from the target series (NWP error prediction). These models are trained using the Darts implementation (Herzen et al. 2022), a popular Python library for time series

forecasting. The same training settings listed previously are used once again, with the exception of batch size, reduced to 32 due to the additional memory requirement of processing the new covariate series. The results for these models are shown in Table 4.

TABLE 4 Test performance for forecasting HSL at different forecasting horizons, K, now incorporating wind direction, wind speed, and spatial coordinate covariates, using different deep learning models in post-calibrating NWP. Reported values are the mean station-wide test RMSE $\pm$ 1 standard deviation averaged over the 74 wind farms.

Model	K = 1	K = 2	K = 3	K = 6	K = 12
TFT (Lim et al. 2020)	0.140 ± 0.025	0.146 ± 0.024	0.152 ± 0.024	0.162 ± 0.028	0.179 ± 0.033
TSMixer (Chen et al. 2023)	0.142 ± 0.033	0.150 ± 0.029	0.155 ± 0.029	0.167 ± 0.037	0.184 ± 0.036
TiDE (Das et al. 2024)	0.111 ± 0.016	0.128 ± 0.019	0.134 ± 0.020	0.145 ± 0.022	0.167 ± 0.029
Transformer (Vaswani et al. 2017)	0.140 ± 0.027	0.141 ± 0.029	0.148 ± 0.035	0.157 ± 0.048	0.175 ± 0.036
NWP Baseline	0.139 ± 0.023	0.146 ± 0.023	0.152 ± 0.024	0.162 ± 0.027	0.178 ± 0.031

Some models, such as TSMixer (Chen et al. 2023) and TiDE (Das et al. 2024), are based entirely on multilayer perceptrons and are useful to contrast against the Transformer-based models. These models are flexible and powerful deep learning forecasting tools that can be fused with other past, future, and static covariates to improve prediction quality (Herzen et al. 2022), which is especially useful in the single-site context. However, they generally did not perform as well on post-calibration as the iTransformer models trained on the NWP errors, particularly for shorter prediction horizons. Interestingly, pairing the NWP forecasting error series with the wind speed and wind direction covariates improved post-calibration from the standard Transformer model for $K \geq 2$. Of these methods, TiDE offers the most significant improvement over the NWP baseline forecast, with relatively strong performance for $K = 12$. The extensive benchmarking performed with state-of-the-art deep learning techniques showcase the difficulty of the multi-site forecasting problem, especially given the small training dataset sample size.

4 | DISCUSSION

As seen in the results from the previous section, we find that even with the most recent advancements, state-of-the-art deep learning models developed for time series forecasting are not effective replacements for physics-based wind power forecasting approaches such as NWP, especially for longer prediction horizons (see Table 3). While these models may be effective for ultra short-term predictions, the importance of incorporating NWP atmospheric dynamics and physics is apparent when comparing the direct HSL prediction to the post-calibrated HSL prediction (see Table 2). Combining the physics-based insights provided from NWP with iTransformer is an effective way to learn spatiotemporal trends for wind power, as validated on ERCOT data across 74 wind farms in Texas.

One of our key findings is that the difference in generalization performance between the SpaceEmbed and regular iTransformer architectures was subtle, indicating that the iTransformer architecture may be able to efficiently learn spatial correlations via context. By embedding spatial coordinates via the SpaceEmbed variant and obtaining similar results, we believe we have uncovered a fascinating phenomenon: *that simply inverting the attention mechanism may provide spatial context between wind farm variates*. Explicitly encoding coordinate information may achieve additional regional improvements, as seen in Fig. 6. However, we acknowledge that our conclusion is based solely on the ERCOT dataset studied in this paper. In the future, we will conduct experiments with more datasets as they become available to us.

To the best of our knowledge, this is the first study investigating the effectiveness of a single deep learning model utilized for NWP post-calibration for dozens of wind farms, resulting in significant forecast improvement over differing time scales. The approach is general and future work may expand to multi-site forecasting of other quantities and meteorological attributes (e.g., predicting wind speed over various locations). The proposed approach for improving wind power forecasts is exceedingly practical and resource-efficient. As an alternative to single-site methods with potentially millions of trainable parameters per wind farm, a single model is designed to learn Nk forecast corrections—a difficult task for conventional statistical and shallow regression techniques. Naturally, this computationally efficient paradigm is much better suited for retraining, which may be

necessary in practice to account for unpredictable and rare weather events and dynamics. Another advantage is the capability of mixing a general multi-site model for efficient and "coarse" adjustments with more specialized single-site models for further fine-tuned corrections, mirroring the ideas of transfer learning and foundation models. As deep learning methods continue to improve at sequence prediction and generation, so does the potential for using them to improve forecasts, unlocking savings in energy spending.

Furthermore, the approach is promising in its performance evaluated under a small training dataset, as well as its potential for direct wind power prediction for short time scales (see Table 3). This has positive implications for implementation under conditions where physics-based forecasts may not be available or accurate for short-term prediction, lowering costs associated with high-fidelity numerical simulation methods. The iTransformer methods are also superior on the post-calibration tasks when compared to other methods paired with covariate series such as wind direction and wind speed. While the performance improvements are most pronounced on shorter prediction horizons, we also observe modest improvements at longer time scales (e.g. $K = 12$ hours ahead). Finally, we note the computational efficiency of this approach: under the parameter and local hardware settings described in the case study, it takes approximately 2 minutes to train the iTransformer architecture and about 15 seconds for evaluation to obtain all test set results.

5 | CONCLUSION

This study considers the problem of multi-site forecasting for wind farms spread over a wide geographical area. Due to complex dynamics and nonlinearities, multi-site wind power forecasting remains a difficult and relevant research challenge. Accurate forecasts are necessary for utility demand management and decision-making, with massive economic implications placed on consumers. Our results demonstrate that most state-of-the-art deep learning models do not perform well for direct (i.e., without NWP post-calibration) multi-site forecasting on this scale, particularly for longer prediction horizons such as 6 and 12 hours ahead. The case study examines the potential of a hybrid approach, where we pair deep learning methods such as the iTransformer architecture with physics-based NWP forecasts for learning farm-level trends in wind power error. The greatest forecasting accuracy is achieved in this setting, where iTransformer architectures correcting NWP forecasting errors results in an up to 33% average improvement over NWP in HSL prediction across dozens of wind farms. The effectiveness of this NWP and deep learning hybrid approach compared to pure deep learning for HSL estimation confirms the importance of utilizing NWP forecasts for learning atmospheric and meteorological dynamics for wind power forecasting. The obtained case study results may be further improved with more extensive hyperparameter optimization and expanded training data.

Despite these strides, there remain important directions for future work. Acquiring additional data from a few more years (e.g., 5 years) could be very helpful to better capture yearly trends, which could improve generalization and forecasting performance. For example, in the current study, the testing set evaluates performance in November and December—months that the model has not seen before. Additionally, higher resolution data (e.g., HSL estimates that update every half hour or every 15 minutes) could also improve the model by learning finer and granular trends on wind power production capabilities. In addition, the results of our case study could be quite sensitive to the length of the input sequence, and context length remains a field of critical interest for Transformer-based models. We leave the tuning of this parameter and further experimentation with context lengths and exploring potential implications on wind energy forecasting for future work.

Furthermore, we believe that rigorous uncertainty quantification and explainability are necessary to improve model confidence and trustworthiness, which is essential for utility companies to make decisions on energy management and resource allocation (Choe et al. 2018, Wang et al. 2021b). In particular, ensuring that data-dependent model explanations (such as those discussed by Cohen et al. (2023) and Cohen et al. (2024)) are robust to uncertainties could have important implications for research in digital twins (Cohen and Huan 2024). Future work should also continue to investigate and interpret the learned variate-based representations. These insights could prove useful for understanding the correlations between existing wind farms as well as for planning the construction of new wind farms (Jang and Byon 2020). We hope that the contributions of this paper encourage future research avenues for multi-site forecasting, especially for hybridized approaches that fuse physics-based and deep learning techniques.

ACKNOWLEDGMENTS

The authors would like to thank ERCOT for the operational data and assistance in this work. This research received support through Schmidt Sciences, LLC. This research is also supported by U.S. National Science Foundation Grant CMMI-2226348.

REFERENCES

Byon, E., Choe, Y. & Yampikulsakul, N. (2016) Adaptive learning in time-variant processes with application to wind power systems. *IEEE Transactions on Automation Science and Engineering*, 13(2), 997–1007.

Chen, S.A., Li, C.L., Yoder, N., Arik, S.O. & Pfister, T. (2023) *TSMixer: An all-MLP architecture for time series forecasting*.

Cho, D., Yoo, C., Im, J. & Cha, D.H. (2020) Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth and Space Science*, 7(4), e2019EA000740. doi:10.1029/2019EA000740.

Choe, Y., Lam, H. & Byon, E. (2018) Uncertainty quantification of stochastic simulation for black-box computer experiments. *Methodology and Computing in Applied Probability*, 20, 1155–1172.

Cohen, J., Byon, E. & Huan, X. To trust or not: Towards efficient uncertainty quantification for stochastic shapley explanations. In: *PHM Society Asia-Pacific Conference*. Vol. 4*, 2023*.

Cohen, J. & Huan, X. (2024) Uncertainty-aware explainable AI as a foundational paradigm for digital twins. *Frontiers in Mechanical Engineering*, 9, 1329146. doi:10.3389/fmech.2023.1329146.

Cohen, J., Huan, X. & Ni, J. (2024) Shapley-based explainable AI for clustering applications in fault diagnosis and prognosis. *Journal of Intelligent Manufacturing*, 35, 4071–4086. doi:10.1007/s10845-024-02468-2.

Costoya, X., Rocha, A. & Carvalho, D. (2020) Using bias-correction to improve future projections of offshore wind energy resource: A case study on the iberian peninsula. *Applied Energy*, 262, 114562.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R. & Yu, R. (2024) *Long-term forecasting with tide: Time-series dense encoder*.

Demolli, H., Dokuz, A.S., Ecemis, A. & Gokcek, M. (2019) Wind power forecasting based on daily wind speed data using machine learning algorithms. *Energy Conversion and Management*, 198, 111823. doi:10.1016/j.enconman.2019.111823.

Ding, Y. (2019) *Data science for wind energy*. : CRC Press.

Du, P. (2018) Ensemble machine learning-based wind forecasting to combine nwp output with data from weather station. *IEEE Transactions on Sustainable Energy*, 10(4), 2133–2141. doi:10.1109/TSTE.2018.2880615.

Eldali, F.A., Hansen, T.M., Suryanarayanan, S. & Chong, E.K. Employing arima models to improve wind power forecasts: A case study in ercot. In: *2016 North American Power Symposium (NAPS). IEEE, 2016*, pp. 1–6.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S. et al. (2021) Deep learning for post-processing ensemble weather forecasts. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200092. doi:10.1098/rsta.2020.0092.

Hendrycks, D. & Gimpel, K. (2016) *Gaussian error linear units (gelus)*.
URL https://arxiv.org/abs/1606.08415

Herzen, J., Lässig, F., Piazzetta, S.G., Neuer, T., Tafti, L., Raille, G. et al. (2022) Darts: User-friendly modern machine learning for time series. *Journal of Machine Learning Research*, 23(124), 1–6.

Jahn, D.E., Gallus Jr, W.A., Nguyen, P.T., Pan, Q., Cetin, K., Byon, E. et al. (2019) Projecting the most likely annual urban heat extremes in the central united states. *Atmosphere*, 10(12), 727.

Jain, P., Shashaani, S. & Byon, E. (2023) Wake effect parameter calibration with large-scale field operational data using stochastic optimization. *Applied Energy*, 347, 121426.

Jang, Y. & Byon, E. (2020) Probabilistic characterization of wind diurnal variability for wind resource assessment. *IEEE Transactions on Sustainable Energy*, 11(4), 2535–2544.

Jang, Y., Byon, E., Vanage, S., Cetin, K., Jahn, D.E., Gallus, W. et al. (2022) Spatiotemporal post-calibration in a numerical weather prediction model for quantifying building energy consumption. *IEEE Transactions on Automation Science and Engineering*, 20(4), 2732–2747. doi:10.1109/TASE.2022.3201475.

Jeon, H., Hartman, B., Cutler, H., Hill, R., Hu, Y., Lu, T. et al. (2022) Estimating the economic impacts of improved wind speed forecasts in the United States electricity sector. *Journal of Renewable and Sustainable Energy*, 14(3). doi:10.1063/5.0081905.

Jeong, C. & Byon, E. (2024) Calibration of building energy computer models via bias-corrected iteratively reweighted least squares method. *Applied Energy*, 360, 122753. doi:10.1016/j.apenergy.2024.122753.

Jeong, C., Xu, Z., Berahas, A.S., Byon, E. & Cetin, K. (2023) Multiblock parameter calibration in computer models. *INFORMS Journal on Data Science*, 2(2), 116–137. doi:10.1287/ijds.2023.0029.

Laloyaux, P., Kurth, T., Dueben, P.D. & Hall, D. (2022) Deep learning to estimate model biases in an operational NWP assimilation system. *Journal of Advances in Modeling Earth Systems*, 14(6), e2022MS003016. doi:10.1029/2022MS003016.

Li, N., Dong, J., Liu, L., Li, H. & Yan, J. (2023) A novel EMD and causal convolutional network integrated with transformer for ultra short-term wind power forecasting. *International Journal of Electrical Power & Energy Systems*, 154, 109470. doi:10.1016/j.ijepes.2023.109470.

Lim, B., Arik, S.O., Loeff, N. & Pfister, T. (2020) *Temporal fusion transformers for interpretable multi-horizon time series forecasting*.

Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A.X. et al. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In: *International Conference on Learning Representations, 2021*.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L. et al. (2024) *iTransformer: Inverted transformers are effective for time series forecasting*.

Liu, Y., Wang, Y., Li, L., Han, S. & Infield, D. (2016) Numerical weather prediction wind correction methods and its impact on computational fluid dynamics based wind power forecasting. *Journal of Renewable and Sustainable Energy*, 8(3).

Liu, Y., Wu, H., Wang, J. & Long, M. (2023) *Non-stationary transformers: Exploring the stationarity in time series forecasting*.

Maria, E., Budiman, E., Taruk, M. et al. Measure distance locating nearest public facilities using haversine and euclidean methods. In: *Journal of Physics: Conference Series*. Vol. 1450. *IOP Publishing, 2020*, p. 012080.

Nie, Y., Nguyen, N.H., Sinthong, P. & Kalagnanam, J. (2023) *A time series is worth 64 words: Long-term forecasting with transformers*.

Orwig, K., Hodge, B.M., Brinkman, G., Ela, E., Milligan, M., Banunarayanan, V. et al. (2012) *Economic evaluation of short-term wind power forecasts in ERCOT: Preliminary results*. National Renewable Energy Lab (NREL), Golden, CO.

Pan, X., Wang, L., Wang, Z. & Huang, C. (2022) Short-term wind speed forecasting based on spatial-temporal graph transformer networks. *Energy*, 253, 124095. doi:10.1016/j.energy.2022.124095.

Peng, X., Deng, D., Wen, J., Xiong, L., Feng, S. & Wang, B. A very short term wind power forecasting approach based on numerical weather prediction and error correction method. In: *2016 China International Conference on Electricity Distribution (CICED). IEEE, 2016*, pp. 1–4.

Qu, K., Si, G., Shan, Z., Kong, X. & Yang, X. (2022) Short-term forecasting for multiple wind farms based on transformer model. *Energy Reports*, 8, 483–490. doi:10.1016/j.egyr.2022.02.184.

Simeunović, J., Schubnel, B., Alet, P.J. & Carrillo, R.E. (2021) Spatio-temporal graph neural networks for multi-site PV power forecasting. *IEEE Transactions on Sustainable Energy*, 13(2), 1210–1220. doi:10.1109/TSTE.2021.3125200.

Sun, S., Liu, Y., Li, Q., Wang, T. & Chu, F. (2023) Short-term multi-step wind power forecasting based on spatio-temporal correlations and transformer neural networks. *Energy Conversion and Management*, 283, 116916. doi:10.1016/j.enconman.2023.116916.

Taheri, J. (2012) *White Paper: High Sustainable Limit (HSL)*. California Independent System Operator.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. et al. (2017) *Attention is all you need*.

Wang, J., AlShelahi, A., You, M., Byon, E. & Saigal, R. (2021) Integrative density forecast and uncertainty quantification of wind power generation. *IEEE Transactions on Sustainable Energy*, 12(4), 1864–1875.

Wang, J., Chung, S., AlShelahi, A., Kontar, R., Byon, E. & Saigal, R. (2021) Look-ahead decision making for renewable energy: A dynamic "predict and store" approach. *Applied Energy*, 296, 117068.

Wang, L. & He, Y. (2022) M2STAN: Multi-modal multi-task spatiotemporal attention network for multi-location ultra-short-term wind power multi-step predictions. *Applied Energy*, 324, 119672. doi:10.1016/j.apenergy.2022.119672.

Wang, L., He, Y., Liu, X., Li, L. & Shao, K. (2022) M2TNet: Multi-modal multi-task transformer network for ultra-short-term wind power multi-step forecasting. *Energy Reports*, 8, 7628–7642. doi:10.1016/j.egyr.2022.05.290.

Wang, X., Guo, P. & Huang, X. (2011) A review of wind power forecasting models. *Energy procedia*, 12, 770–778. doi:10.1016/j.egypro.2011.10.103.

Wang, Y., Wu, H., Dong, J., Liu, Y., Long, M. & Wang, J. (2024) *Deep time series models: A comprehensive survey and benchmark*.

Wang, Y., Zou, R., Liu, F., Zhang, L. & Liu, Q. (2021) A review of wind speed and wind power forecasting with deep neural networks. *Applied Energy*, 304, 117766. doi:10.1016/j.apenergy.2021.117766.

Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J. & Long, M. (2023) *Timesnet: Temporal 2D-variation modeling for general time series analysis*.

Wu, H., Meng, K., Fan, D., Zhang, Z. & Liu, Q. (2022) Multistep short-term wind speed forecasting using transformer. *Energy*, 261, 125231. doi:10.1016/j.energy.2022.125231.

Wu, H., Xu, J., Wang, J. & Long, M. (2022) *Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting*.

Ye, F., Brodie, J., Miles, T. & Ezzat, A.A. Ultra-short-term probabilistic wind forecasting: Can numerical weather predictions help? In: *2023 IEEE Power & Energy Society General Meeting (PESGM), 2023*, pp. 1–5.

Ye, F., Brodie, J., Miles, T. & Ezzat, A.A. (2024) AIRU-WRF: A physics-guided spatio-temporal wind forecasting model and its application to the US Mid Atlantic offshore wind energy areas. *Renewable Energy*, 223, 119934. doi:10.1016/j.renene.2023.119934.

You, M., Liu, B., Byon, E., Huang, S. & Jin, J. (2017) Direction-dependent power curve modeling for multiple interacting wind turbines. *IEEE Transactions on power systems*, 33(2), 1725–1733.

Zeng, A., Chen, M., Zhang, L. & Xu, Q. (2022) *Are transformers effective for time series forecasting?*

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H. et al. (2021) *Informer: Beyond efficient transformer for long sequence time-series forecasting*.

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L. & Jin, R. (2022) *Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting*.